

Biostatistics: The Book

A. J. Smit

Biostatistics: The Book

BCB744 Biostatistics

A. J. Smit

Placeholder title page for the compiled BCB744 biostatistics book.

Half-Title Placeholder

Biostatistics: The Book

Copyright Placeholder

Rights, edition, ISBN, and publisher details to be inserted here.

Dedication Placeholder

Dedication text to be inserted here.

Table of contents

I	Foundations	1
1	The Statistical Landscape	5
1.1	Important Concepts	8
1.2	Core Principles	8
1.3	Core Skills	9
1.4	Section Structure	9
1.4.1	Part I. Foundations	9
1.4.2	Part II. Inference	10
1.4.3	Part III. Modelling	10
1.4.4	Part IV. Extensions	10
1.4.5	Part V. Reproducibility	10
1.5	Assessment and Practice	11
1.5.1	Biostatistics Self-Assessment	11
1.5.2	Biostatistics Example 1	11
2	Summarising Biological Data	13
2.1	Key Concepts	14
2.2	Foundational Definitions	14
2.2.1	Variables and Parameters	14
2.2.2	Samples and Populations	14
2.2.3	When Is Something Random?	16
2.3	Descriptive Statistics	17
2.3.1	The Core Equations	17
2.3.2	Measures of Central Tendency	17
2.3.3	Measures of Variance or Dispersion Around the Centre	22
2.4	The Palmer Penguin Dataset	25
2.5	Exploring the Data Structure	26
2.5.1	Inspecting Type and Layout	26
2.5.2	Inspecting Size and Names	27
2.5.3	Previewing the Data	28
2.6	Data Summaries	29
2.6.1	<code>summary()</code>	30

2.6.2	<code>psych::describe()</code>	30
2.6.3	<code>skimr::skim()</code>	31
2.6.4	<code>summarytools::dfSummary()</code>	32
2.7	Descriptive Statistics by Group	32
2.8	Reporting	33
2.9	Conclusion	35
3	Visualising Data	37
3.1	Key Concepts	38
3.2	Choosing the Right Plot	38
3.3	Visualising Distributions	39
3.3.1	Histograms	39
3.3.2	Density Plots	41
3.4	Comparing Groups	43
3.4.1	Box Plots	43
3.4.2	Violin Plots	45
3.4.3	Bar Graphs	47
3.5	Visualising Relationships	50
3.5.1	Scatter Plots	51
3.5.2	Line Graphs	52
3.6	Extensions	53
3.6.1	Heatmaps	53
3.6.2	Hovmöller Diagrams	55
3.6.3	Other Examples (To Develop)	57
3.7	Conclusion	57
4	Distributions, Sampling, and Uncertainty	59
4.1	What Are Data Distributions?	60
4.2	Discrete Distributions	60
4.2.1	Mathematical Foundations	61
4.2.2	The Bernoulli Distribution	61
4.2.3	The Binomial Distribution	62
4.2.4	The Poisson Distribution	65
4.2.5	The Negative Binomial Distribution	67
4.2.6	The Geometric Distribution	69
4.3	Continuous Distributions	71
4.3.1	The Normal Distribution	71
4.3.2	The Log-Normal Distribution	74
4.3.3	The Gamma Distribution	76
4.3.4	The Exponential Distribution	78
4.3.5	The Beta Distribution	80

4.4	Identifying Your Data's Distribution	82
4.5	Inference-Related Distributions	84
4.6	Sampling-Level Thinking	85
4.6.1	Samples, Populations, and Sampling Variation	85
4.6.2	The Central Limit Theorem	86
4.6.3	Standard Error	87
4.6.4	Confidence Intervals	88
4.7	Linking Distributions to Later Chapters	88
4.8	Summary	89
II	Inference	91
5	Statistical Inference	95
5.1	The Basic Workflow of Hypothesis Testing	96
5.2	Hypotheses	97
5.3	Test Statistics, p -Values, and Decision Rules	98
5.3.1	Test Statistics	98
5.3.2	The p -Value	98
5.3.3	The Significance Threshold, α	99
5.3.4	The Decision Rule	99
5.3.5	Interpretation	100
5.4	Assumptions	100
5.5	Parametric and Non-Parametric Methods	100
5.6	Reporting Inferential Results	101
5.7	Conclusion	102
6	Raw-Data Assumptions and Transformations	103
6.1	Assumption-Checking Workflow	104
6.2	Identifying the Data Structure	105
6.3	Group Comparisons: Normality Within Groups	105
6.3.1	Graphical Checks	105
6.3.2	Formal Test	107
6.3.3	Other Tests for Normality	108
6.4	Group Comparisons: Variance Across Groups	109
6.4.1	Other Tests for Homogeneity	110
6.5	Paired Designs: Check the Differences	111
6.6	Associations: Check the Joint Structure	112
6.7	Independence	113
6.8	Reporting Assumption Checks	113
6.9	Responding to Assumption Violations	114

6.10	Why Transform Data?	114
6.11	Transformation Decision Guide	115
6.12	Common Transformations	115
6.12.1	Log Transformation	116
6.12.2	Square-Root Transformation	116
6.12.3	Arcsine Transformation	116
6.12.4	Reciprocal Transformation	116
6.12.5	Square and Cube Transformations	116
6.12.6	Anscombe Transformation	116
6.13	Worked Example with Transformations	117
6.14	Check the Assumptions Again	119
6.15	When Transformation Is Not the Right Solution	119
6.16	Summary	119
7	*t*-Tests	121
7.1	Design Determines the Test	122
7.2	R Functions	122
7.3	A Simple Experiment	124
7.4	One-Sample <i>t</i> -Tests	124
7.4.1	When to Use a One-Sample <i>t</i> -Test	124
7.5	Example 1: Tree Height Against Two Reference Means	125
7.5.1	Do an Exploratory Data Analysis (EDA)	125
7.5.2	State the Hypotheses	126
7.5.3	Apply the Test	127
7.5.4	Interpret the Results	127
7.5.5	Reporting	127
7.6	Two-Sample <i>t</i> -Tests	128
7.6.1	When to Use a Two-Sample <i>t</i> -Test	128
7.6.2	Choosing Between Student and Welch	129
7.6.3	Other Tests for Homogeneity	129
7.7	Example 2: Penguin Flipper Length in Two Species	130
7.7.1	Do an Exploratory Data Analysis (EDA)	130
7.7.2	State the Hypotheses	131
7.7.3	Apply the Test	131
7.7.4	Interpret the Results	132
7.7.5	Reporting	132
7.8	Example 3: A Two-Sample Comparison After Log Transformation	133
7.8.1	Do an Exploratory Data Analysis (EDA)	134
7.8.2	State the Hypotheses	135
7.8.3	Apply the Test	136
7.8.4	Interpret the Results	136

7.8.5	Reporting	137
7.9	Paired t -Tests	137
7.9.1	When to Use a Paired t -Test	138
7.10	Example 4: Before-and-After Measurements on Piglets	138
7.10.1	Do an Exploratory Data Analysis (EDA)	138
7.10.2	State the Hypotheses	140
7.10.3	Apply the Test	140
7.10.4	Interpret the Results	140
7.10.5	Reporting	140
7.11	When Assumptions Are Difficult	141
7.12	Optional: One-Sided t -Tests	142
7.12.1	One-Sided One-Sample t -Test	142
7.12.2	One-Sided Two-Sample t -Test	143
7.13	Rank-Based Alternatives Within the t -Test Family	143
7.14	Decision Steps for Any t -Test	144
7.15	A t -Test Workflow	145
7.15.1	Loading Data	145
7.15.2	Visualising Data	145
7.15.3	Formulating a Hypothesis	146
7.15.4	Choosing a Test	147
7.15.5	Checking Assumptions	147
7.15.6	Running an Analysis	148
7.15.7	Interpreting the Results	148
7.15.8	Drawing Conclusions	148
7.15.9	Reporting	148
7.15.10	Going Further	149
7.16	Conclusion	150
8	ANOVA	151
8.1	Why Not Do Multiple t -Tests?	152
8.2	How ANOVA Partitions Variation	154
8.3	Data Structure and Decision Rule	155
8.4	R Functions	156
8.5	Example 1: Chick Masses Under Four Diets	156
8.5.1	The One-Way Design	157
8.5.2	Do an Exploratory Data Analysis (EDA)	157
8.5.3	Apply the Test	159
8.5.4	Brief Residual Check	159
8.5.5	Identify Which Groups Differ	160
8.5.6	Interpret the Results	162
8.5.7	Reporting	162

8.6	Factorial ANOVA	163
8.6.1	A Detour We Should Avoid	164
8.6.2	Two Factors in One Model	164
8.6.3	Interaction in Factorial ANOVA	165
8.7	If Assumptions Fail	166
8.7.1	Welch's One-Way ANOVA	166
8.7.2	Kruskal-Wallis Rank-Sum Test	166
8.8	What This Chapter Does Not Cover	167
8.9	Conclusion	168
9	Correlation and Association	169
9.1	Choosing the Appropriate Correlation	170
9.1.1	How to Read Magnitude	170
9.2	Data Structure and Diagnostics	171
9.2.1	What is the Data Structure?	171
9.2.2	What is the Relationship Form?	171
9.2.3	What's the Data Quality?	171
9.2.4	Secondary Considerations	171
9.3	R Functions	172
9.4	Pearson Correlation	172
9.5	Example 1: Pearson Correlation in <i>Ecklonia maxima</i>	173
9.5.1	Do an Exploratory Data Analysis (EDA)	173
9.5.2	State the Hypotheses	174
9.5.3	Apply the Test	174
9.5.4	Interpret the Results	175
9.5.5	Reporting	175
9.6	Spearman Rank Correlation	176
9.7	Example 2: Spearman's ρ When the Rank Pattern Is Clearer Than the Linear One	176
9.7.1	Do an Exploratory Data Analysis (EDA)	176
9.7.2	State the Hypotheses	177
9.7.3	Apply the Test	177
9.7.4	Interpret the Results	178
9.7.5	Reporting	178
9.8	Kendall Rank Correlation	179
9.9	Example 3: Kendall's τ for Concordance in Rank Order	179
9.9.1	Do an Exploratory Data Analysis (EDA)	179
9.9.2	State the Hypotheses	180
9.9.3	Apply the Test	180
9.9.4	Interpret the Results	181
9.9.5	Reporting	181

9.10	Correlation Matrices and Heatmaps	181
9.10.1	Create the Correlation Matrix	182
9.10.2	Visualise the Matrix	182
9.11	Correlation and Non-Independent Data	184
9.12	Example 4: A Correlation Can Be Modified by Site Structure	185
9.13	Correlation Versus Regression	186
9.14	If Assumptions Fail	187
9.15	Summary	191
10	Finding Your Way on the Statistical Landscape	193
10.1	Two Important Questions	194
10.2	The Core Tests	194
10.2.1	Comparing a Single Group to a Reference	194
10.2.2	Comparing Two Groups	194
10.2.3	Comparing Three or More Groups	195
10.2.4	Asking Whether Two Variables Move Together	195
10.2.5	Modelling How One Variable Responds to Another	196
10.3	Beyond the Core Tests	196
10.4	Assumptions: Check Before You Proceed	198
10.5	Conclusion	199
III	Modelling	201
11	Residuals and Model-Based Diagnostics	205
11.1	Models, Errors, and Fitted Values	206
11.2	Residuals	207
11.3	Why Assumptions Apply to Residuals	208
11.4	Main Regression Assumptions	210
11.4.1	Linearity	211
11.4.2	Constant Variance	211
11.4.3	Independence	211
11.4.4	Normality of Residuals	212
11.5	Diagnostic Plots	212
11.5.1	Residual versus Fitted Plot	213
11.5.2	Q-Q Plot of Residuals	213
11.5.3	Example 1: Idealised Diagnostic Plots	214
11.6	Example 2: Worked Example with Real Data	215
11.7	Linking Diagnostics to Decisions	218
11.8	Connection to Earlier Chapters	219
11.9	Forward Links	219

11.10 Summary	220
12 Simple Linear Regression	221
12.1 Main Concepts	222
12.2 Nature of the Data and Assumptions	223
12.2.1 Requirements Before Fitting	223
12.2.2 Assumptions to Check After Fitting	223
12.2.3 Assumption and Diagnostics	224
12.3 The Model	224
12.4 The Fitting Rule	226
12.5 R Function	226
12.6 Outliers and Their Impact on Simple Linear Regression	227
12.7 Example 1: Sparrow Wing Length and Age	227
12.7.1 Do an Exploratory Data Analysis (EDA)	228
12.7.2 State the Model Question and Hypothesis	229
12.7.3 Fit the Model	230
12.7.4 Test the Assumptions	231
12.7.5 Interpret the Results	233
12.7.6 Reporting	234
12.8 Example 2: Adelie Penguin Bill Length and Body Mass	234
12.8.1 Do an Exploratory Data Analysis (EDA)	235
12.8.2 Create a Plot	235
12.8.3 State the Hypothesis	236
12.8.4 Fit the Model	237
12.8.5 Test the Assumptions	237
12.8.6 Interpret the Results	242
12.8.7 Reporting	243
12.9 Confidence and Prediction Intervals	243
12.10 Prediction Versus Explanation	245
12.11 What to Do When Assumptions Fail	246
12.12 Common Mistakes	246
12.12.1 Marking Rubric	247
12.13 Summary	248
13 Polynomial Regression	249
13.1 Important Concepts	250
13.2 When This Method Is Appropriate	251
13.3 Nature of the Data and Assumptions	251
13.3.1 Requirements Before Fitting	251
13.3.2 Assumptions to Check After Fitting	252
13.3.3 Assumption–Diagnostic Mapping	252

13.4	The Core Equations	253
13.5	The Fitting Rule	253
13.6	Implementation	254
13.7	Example: CO ₂ Uptake and Curved Response	254
13.7.1	Example Dataset	254
13.7.2	Do an Exploratory Data Analysis (EDA)	256
13.7.3	State the Model Question	257
13.7.4	Fit the Model	258
13.7.5	Test Assumptions	260
13.7.6	Interpret the Results	262
13.7.7	Reporting	263
13.8	What to Do When Assumptions Fail	264
13.9	Summary	265
14	Multiple Regression and Model Specification	267
14.1	Important Concepts	268
14.2	Model Specification vs Model Selection	269
14.3	Nature of the Data	269
14.3.1	Requirements Before Fitting	269
14.3.2	Assumptions After Fitting	270
14.4	The Core Equation	271
14.5	Outliers	271
14.6	R Function	272
14.7	Example 1: A Climate Model for the East Coast Transition Zone	272
14.7.1	Exploratory Data Analysis	273
14.7.2	Hypotheses	276
14.7.3	Fit the Initial Model	276
14.7.4	Check for Multicollinearity	279
14.7.5	Reduce Multicollinearity	280
14.7.6	Fit the Selected Model	281
14.7.7	Reading the Coefficient and ANOVA Tables	283
14.7.8	Scaling and Centring	283
14.7.9	Check Diagnostics	284
14.7.10	Interpret the Results	287
14.7.11	Report	288
14.8	Example 2: Categorical Predictors — Distance and Bioregion	289
14.8.1	Dummy Variable Coding	289
14.8.2	Dummy Variables in Practice	292
14.8.3	Biological Question and Hypotheses	295
14.8.4	Visualise the Main Effects	295
14.8.5	Fit and Assess Nested Models	296

14.8.6	Interpret the Additive Model	298
14.8.7	Where ANCOVA Fits	299
14.8.8	Report	299
14.9	Example 3: The Full Additive Model	300
14.9.1	Fit the Model	300
14.9.2	Assess Assumptions	301
14.9.3	Interpret the Model	305
14.9.4	Report	306
14.10	Alternative Categorical Variable Coding Schemes (Contrasts)	306
14.11	Towards Interaction Models	308
14.12	Summary	308
15	Interaction Effects	309
15.1	Important Concepts	310
15.2	The Core Equation	311
15.3	When Is an Interaction Appropriate?	312
15.4	Interpreting Conditional Coefficients	313
15.4.1	A Worked Algebra Example	314
15.4.2	The Role of the Reference Group	314
15.5	Example 1: Two Continuous Predictors with and without Centring	315
15.5.1	Example Dataset	315
15.5.2	Do an Exploratory Data Analysis (EDA)	316
15.5.3	State the Model Question	318
15.5.4	Fit the Uncentred and Centred Models	318
15.5.5	Interpret the Results	319
15.5.6	Scale Dependence	320
15.6	Example 2: Coastal Distance and Bioregional Context	320
15.6.1	Example Dataset	320
15.6.2	Do an Exploratory Data Analysis (EDA)	321
15.6.3	State the Hypotheses	321
15.6.4	Fit Nested Models	322
15.6.5	Interpret the Interaction	323
15.6.6	Diagnostic Checks for Interaction Models	325
15.6.7	Reporting	326
15.7	Common Mistakes	327
15.8	Summary	327
16	Collinearity, Confounding, and Measurement Error	329
16.1	Important Ideas	330
16.2	Problems of Note	331
16.3	R Functions	331

16.4	Collinearity	331
16.4.1	What Is Collinearity?	332
16.4.2	The Mechanism: Why Shared Variance Destabilises Coefficients	332
16.4.3	Proxy Variables and Mechanistic Variables	333
16.5	Example 1: Predicting Plant Growth	333
16.6	Example 2: Nutrient Limitation in Coastal Systems	334
16.7	Diagnosing Collinearity	334
16.8	Example 3: Worked Collinearity Example	335
16.8.1	Inspect the Predictor Set	335
16.8.2	Diagnose the Overlap with VIF	337
16.8.3	Compare Simple and Multiple-Regression Slopes	337
16.8.4	Interpret the Collinearity Problem	338
16.8.5	Reporting	338
16.9	Addressing Collinearity	339
16.10	Confounding	339
16.10.1	What Is Confounding?	339
16.10.2	Confounding and Causal Interpretation	340
16.11	Example 4: Worked Confounding Example	340
16.11.1	Do an Exploratory Data Analysis (EDA)	340
16.11.2	State the Model Question and Hypotheses	342
16.11.3	Fit Competing Models	342
16.11.4	Interpret the Results	343
16.11.5	Practical Implications of Confounding	343
16.11.6	Reporting	343
16.12	Measurement Error	344
16.13	Example 5: Worked Measurement-Error Example	345
16.13.1	Simulate the True and Noisy Predictors	345
16.13.2	Compare the Fitted Relationships	345
16.13.3	Interpret the Measurement-Error Problem	346
16.13.4	Reporting	347
16.13.5	Model II Regression as a Solution to Measurement Error	347
16.14	Revisiting Proxy Variables	348
16.15	A Practical Workflow	349
16.16	Summary	349
17	Model Checking and Evaluation	351
17.1	Important Concepts	352
17.2	Nature of the Data and Assumptions	352
17.3	R Functions	353
17.4	A Worked Diagnostic Example	353

17.4.1	Residual Diagnostics	354
17.4.2	Assumption Checks and Their Limits	355
17.4.3	Influence and Leverage	355
17.4.4	Model Comparison	356
17.4.5	Overfitting and Generalisation	357
17.4.6	Cross-Validation	357
17.4.7	What the Diagnostics Tell Us	359
17.5	A Practical Workflow	359
17.6	Reporting	360
17.7	Summary	361
IV	Extensions	363
18	Pseudoreplication	367
18.1	Important Concepts	368
18.2	What Is Pseudoreplication?	369
18.3	The Experimental Unit	369
18.3.1	What It Is	369
18.3.2	How to Identify It	370
18.3.3	The Domain of Inference	371
18.4	Types of Pseudoreplication	372
18.5	Why Is It a Problem?	372
18.6	Worked Examples	373
18.6.1	Example 1: Comparing Two Ponds	373
18.6.2	Example 2: Interspersion	374
18.6.3	Example 3: Temporal Structure	375
18.6.4	Example 4: Seasonal Comparison	376
18.7	Interpreting the Problem	377
18.8	Summary	377
19	Dependence and Mixed Models	379
19.1	Key Concepts	380
19.2	When This Method Is Appropriate	380
19.3	Nature of the Data and Assumptions	380
19.4	The Core Equations	381
19.5	R Functions	381
19.6	Example 1: Shorebird Landing Distance with Site-Level Dependence	381
19.6.1	Example Dataset	381
19.6.2	Do an Exploratory Data Analysis (EDA)	382
19.6.3	State the Model Question and Hypotheses	384

19.6.4	Fit the Model	385
19.6.5	Test Assumptions / Check Diagnostics	386
19.6.6	Interpret the Results	387
19.6.7	Reporting	388
19.7	Example 2: Repeated Snake Habituation Within Individuals	388
19.7.1	Why This Is a Dependence Problem	389
19.7.2	Fit a Naive Model and a Mixed Model	389
19.7.3	Interpret the Results	390
19.7.4	Reporting	391
19.8	What to Do When Assumptions Fail / Alternatives	391
19.9	Summary	391
20	Generalised Linear Models	393
20.1	Key Concepts	394
20.2	When This Method Is Appropriate	394
20.3	Nature of the Data and Assumptions	394
20.4	The Core Equations	394
20.5	R Functions	395
20.6	Simple Proportion Tests as a Bridge to GLMs	395
20.6.1	One-Sample Proportion Tests	395
20.6.2	Two-Sample Proportion Tests	396
20.7	Example 1: Logistic Regression for Presence-Absence Data	397
20.7.1	Example Dataset	397
20.7.2	Do an Exploratory Data Analysis (EDA)	397
20.7.3	State the Model Question and Hypotheses	399
20.7.4	Fit the Model	399
20.7.5	Test Assumptions / Check Diagnostics	399
20.7.6	Interpret the Results	400
20.7.7	Reporting	401
20.8	Example 2: Count Data and Overdispersion	401
20.8.1	Example Dataset	401
20.8.2	Do an Exploratory Data Analysis (EDA)	401
20.8.3	State the Model Question and Hypotheses	403
20.8.4	Fit the Model	403
20.8.5	Test Assumptions / Check Diagnostics	404
20.8.6	Interpret the Results	405
20.8.7	Reporting	406
20.9	What to Do When Assumptions Fail / Alternatives	406
20.10	Summary	406
21	Generalised Additive Models	409

21.1	Key Concepts	410
21.2	When This Method Is Appropriate	410
21.3	Nature of the Data and Assumptions	410
21.4	The Core Equations	410
21.5	R Functions	411
21.6	Example 1: Sea-Temperature Structure Through Time	411
21.6.1	Example Dataset	411
21.6.2	Do an Exploratory Data Analysis (EDA)	411
21.6.3	State the Model Question and Hypotheses	413
21.6.4	Fit the Model	414
21.6.5	Test Assumptions / Check Diagnostics	415
21.6.6	Interpret the Results	416
21.6.7	Reporting	418
21.7	What to Do When Assumptions Fail / Alternatives	418
21.8	Summary	418
22	Nonlinear Regression	421
22.1	Key Concepts	422
22.2	When This Method Is Appropriate	422
22.3	Nature of the Data and Assumptions	422
22.4	The Core Equation	422
22.5	R Functions	423
22.6	Example 1: Why Use a Mechanistic Nonlinear Model?	423
22.7	Example 2: Algal Nutrient Uptake Kinetics	423
22.7.1	Do an Exploratory Data Analysis (EDA)	425
22.7.2	State the Model Question	427
22.7.3	Fit the Model	427
22.7.4	Check Diagnostics	427
22.7.5	Interpret the Results	428
22.7.6	Reporting	430
22.7.7	What If We Add Treatments?	430
22.7.8	Repeated-Measures Extension	431
22.8	Example 3: Repeated-Measures Michaelis-Menten Kinetics	432
22.8.1	Example Dataset	432
22.8.2	Do an Exploratory Data Analysis (EDA)	434
22.8.3	State the Model Question	434
22.8.4	Fit the Model	434
22.8.5	Check Diagnostics	436
22.8.6	Interpret the Results	436
22.8.7	Reporting	439
22.9	Choosing Among Flexible Models	439

22.10 Practical Caution	440
22.11 Summary	440
23 Quantile Regression	441
23.1 Key Concepts	442
23.2 When This Method Is Appropriate	442
23.3 Nature of the Data and Assumptions	443
23.4 The Core Equations	443
23.5 R Functions	444
23.6 Example 1: Blade-Length Envelopes in <i>Laminaria pallida</i>	444
23.6.1 Example Dataset	444
23.6.2 Do an Exploratory Data Analysis (EDA)	444
23.6.3 State the Model Question and Hypotheses	446
23.6.4 Fit the Models	447
23.6.5 Test Assumptions / Check Diagnostics	448
23.6.6 Interpret the Results	450
23.6.7 Reporting	451
23.7 What to Do When Assumptions Fail / Alternatives	452
23.8 Summary	452
24 Prediction and Explanation	453
24.1 Key Concepts	454
24.2 Explanation Versus Prediction	454
24.3 Cross-Validation and Out-of-Sample Thinking	454
24.4 A Short Paired Example	455
24.5 What This Means in Practice	456
24.6 Practical Guidance	456
24.7 Why This Distinction Matters for the Rest of the Course	457
24.8 Summary	457
25 Regularisation	459
25.1 Key Concepts	460
25.2 When This Method Is Appropriate	460
25.3 Why Regularisation Matters	460
25.4 The Core Equations	461
25.4.1 Ridge Regression	461
25.4.2 Lasso Regression	461
25.4.3 Elastic Net	462
25.5 Cross-Validation	462
25.6 R Functions	462
25.7 Example 1: Ridge Regression with the Seaweed Data	462

25.7.1	Prepare the Data	463
25.7.2	Fit the Cross-Validated Ridge Model	463
25.7.3	Inspect the Cross-Validation Curve	463
25.7.4	Coefficient Paths and the Meaning of Shrinkage	463
25.7.5	Extract the Fitted Model	465
25.7.6	Interpret the Ridge Results	466
25.7.7	$\lambda_{\min}$ Versus λ_{1se}	466
25.7.8	Reporting	467
25.8	Example 2: Lasso Regression	468
25.8.1	Reporting	469
25.9	Example 3: Elastic Net Regression	470
25.9.1	Reporting	471
25.10	Theory-Driven and Data-Driven Variable Selection	472
25.11	Summary	472
 V Reproducibility		473
26	Reproducible Workflow	477
26.1	Key Concepts	478
26.2	When This Method Is Appropriate	478
26.3	Nature of the Data and Assumptions	478
26.4	Tools and Practice	479
26.5	Example 1: A Reproducible Quarto Workflow Built from Project Files	479
26.5.1	Example Dataset	479
26.5.2	Do an Exploratory Data Analysis (EDA)	480
26.5.3	State the Workflow Question	481
26.5.4	Generate the Outputs	482
26.5.5	Check the Workflow	482
26.5.6	What This Means for Us	483
26.6	Common Failures	483
26.7	Summary	483

List of Figures

2.1	Drawing increasingly larger sample sizes from a population with a true mean of 13 and an SD of 1.	16
2.2	Generated normal, right-skewed, left-skewed, and bimodal distributions, shown in panels A-D, with the mean and median indicated in each panel.	19
2.3	The idealised Normal distribution showing the proportion of data within 1, 2, and 3SD from the mean.	23
2.4	The figures represent A) a scatterplot of the mean and raw chicken mass values; B) a bar graph of the chicken mass values, showing whiskers indicating $1 \pm \text{SD}$; C) a box and whisker plot of the chicken mass data; and D) chicken mass as a function of both Diet and Time (10 and 21 days).	34
3.1	Example histograms for the Old Faithful data. A) A default frequency histogram. B) A relative frequency histogram with 1-minute bins. C) A relative frequency histogram with 0.5-minute bins. D) An ECDF.	40
3.2	Grouped histograms for the four <i>Iris</i> variables.	41
3.3	Density-plot variations based on the Old Faithful data. A) A basic density plot. B) A density plot with a rug. C) A histogram overlaid with a density curve. D) A count histogram with a rescaled density curve.	42
3.4	Box plot summarising the amount of sleep required by different ‘vores’.	43
3.5	A panelled collection of box plots, one for each of the four variables, with a scatter plot to indicate the spread of the raw data points.	45
3.6	Variations of violin plots applied to the <i>Iris</i> data.	46
3.7	Bar graphs for categorical summaries in the built-in <i>Iris</i> data. A) A stacked count bar graph. B) A stacked proportion bar graph. C) A pie chart using the same proportions. D) A side-by-side count bar graph.	49
3.8	Bar graphs indicating the mean size ($\pm \text{SD}$) for various flower features of three species of <i>Iris</i>	50
3.9	Examples of scatter plots made for the <i>Iris</i> data. A) A default scatter plot showing the relationship between petal length and width. B) The same data with fitted lines added.	51

3.10	A time-series plot showing the monthly climatology for several sites around South Africa. The specific line drawn here forms a stepped graph.	53
3.11	A calendar heatmap showing a time series of SST for Western Australia. The marine heatwave of summer 2011 is visible in the warm colours.	54
3.12	A Hovmöller diagram of zonally averaged SST for a region off North-west Africa in the Canary upwelling system.	55
3.13	Zonally averaged time series of SST in the Canary Current system displayed as a horizon plot.	57
4.1	Binomial distribution with 40 observations per simulation, 10 simulations, 100 trials, and $p = 0.75$. This could represent 10 separate germination experiments, each planting 100 seeds with a 75% germination probability. The Normal Q-Q plot reveals the characteristic staircase pattern of a discrete distribution and the mild right-skew: points curve away from the reference line at both ends.	63
4.2	Poisson distribution with 50 observations per simulation, 10 simulations, and $\lambda = 2$. This could represent the expected number of a particular invertebrate species per 1 m ² quadrat in a habitat where the mean density is 2 individuals per quadrat.	67
4.3	Negative Binomial distribution with 50 observations per simulation, 10 simulations, size parameter $r = 5$, and $p = 0.7$. The long right tail reflects the typical pattern of overdispersed ecological count data.	69
4.4	Geometric distribution with 50 observations per simulation, 10 simulations, and $p = 0.7$. Most observations cluster near zero (few failures before success), but the long tail captures the occasional run of bad luck.	71
4.5	The idealised Normal distribution showing the proportion of data within 1, 2, and 3 SDs from the mean.	72
4.6	Normal distribution with 100 observations per simulation and 5 simulations ($\mu = 0$, $\sigma = 1$, i.e. the standard normal). In practice, biological variables will have different means and standard deviations, but the symmetric bell shape will be the same. The Q-Q plot points follow the reference line closely, confirming normality.	73
4.7	Log-Normal distribution with 100 observations per simulation and 5 simulations ($\mu = 0$, $\sigma = 0.5$). The strong right skew in the density plot is echoed by the upward curve of Q-Q points away from the reference line in the upper tail. On a log scale, this distribution becomes Normal.	76

4.8	Gamma distribution with 100 observations per simulation, 5 simulations, shape $\alpha = 1$, and scale $\beta = 0.4$. With $\alpha = 1$, this reduces to an Exponential distribution. Increasing α shifts the mass rightward and reduces skewness. The Q–Q plot shows the pronounced right-skew departure from normality.	78
4.9	Exponential distribution with 100 observations per simulation, 5 simulations, and rate $\lambda = 1$. The steep decline from zero reflects the high probability of short waiting times and the low probability of long ones. The Q–Q plot curves strongly upward, indicating heavy right skewness.	80
4.10	Beta distribution with 100 observations per simulation, 5 simulations, $\alpha = 2$, and $\beta = 5$. This right-skewed shape on the $[0, 1]$ interval could represent, for example, occupancy probabilities for a moderately rare species across a set of habitat patches. The Q–Q plot shows moderate departure from normality in the upper tail.	82
4.11	Four-panel diagnostics for 200 normally distributed observations: histogram, kernel density estimate, empirical CDF, and Normal QQ plot. The close adherence to the 45° line in the QQ plot confirms approximate normality.	83
4.12	A simulation demonstrating the central limit theorem. The underlying population is strongly right-skewed (Gamma distribution), yet the sampling distribution of the sample mean becomes increasingly normal as sample size grows from $n = 2$ to $n = 99$	87
5.1	A reminder that “almost significant” is not significant.	99
6.1	Normality checks for grouped data. A) Histograms of bill length within each penguin species. B) A pooled Q–Q plot that ignores species. C) Q–Q plots drawn separately within species.	106
6.2	Variance checks for grouped data. A) Box plots of body mass by penguin species. B) Sample variances by species.	109
6.3	Checks for paired data. A) Histogram of paired differences. B) Q–Q plot of paired differences.	111
6.4	Scatter plot of tree height and volume from the base R <code>trees</code> dataset.	112
6.5	A real-data example showing how a log transformation changes the distribution of ozone concentrations. A) Raw ozone values by season. B) Log-transformed ozone values by season.	118
7.1	One-sample tree-height data in metres. The blue line marks a reference mean of 23.0 m and the red line marks 24.5 m.	126

7.2	Flipper length in Adelie and Chinstrap penguins. The group centres differ, while the group spreads are similar.	131
7.3	Ozone concentrations before and after log transformation. A) Raw ozone values (ppb) by season. B) Log-transformed ozone values by season.	134
7.4	Distribution of paired differences in piglet body mass after minus before the feeding period.	139
7.5	Boxplots showing differences in morphometric properties of the kelp <i>Ecklonia maxima</i> at two sites in False Bay.	146
7.6	Boxplots showing the difference in stipe mass (kg) of the kelp <i>Ecklonia maxima</i> at two sites in False Bay.	147
8.1	Visual intuition for ANOVA. The dashed blue line is the grand mean. The red points are the diet means. Distance from the red points to the blue line contributes to between-group variation. Distance from observations to their group mean contributes to within-group variation.	154
8.2	Boxplot and jittered observations for chick mass at Day 21 under four diets. Black points show the diet means and the vertical bars show one standard deviation above and below each mean.	158
8.3	Brief residual diagnostics for the chick-mass ANOVA. The left panel shows residuals versus fitted values and the right panel shows a normal Q-Q plot of the residuals.	160
8.4	Tukey HSD comparisons among the four chick diets at Day 21.	161
9.1	Scatterplot showing the relationship between <i>Ecklonia maxima</i> stipe length and frond length. The fitted line is included only to show the overall linear tendency.	174
9.2	Relationship between stipe diameter and primary blade length in <i>Ecklonia maxima</i> . The fitted line is shown only as a visual guide to the overall trend.	177
9.3	Relationship between primary blade length and primary blade width in <i>Ecklonia maxima</i> . The question here is whether longer blades also tend to rank as wider blades.	180
9.4	Pairwise correlations showing the strength of all Pearson correlations between variables as a scale from red (negative) to blue (positive).	183
9.5	Pairwise correlations of the <i>Ecklonia</i> dataset visualised as a heatmap in ggplot2	184
9.6	Relationship between stipe length and epiphyte length in <i>Ecklonia maxima</i> , coloured by site. The black dashed line is the naive overall fit, while the coloured lines show the within-site trends.	185

10.1	Links from common biological question types to the main method and the next extension when the problem becomes more complex. .	195
11.1	The systematic part of a relationship is removed when residuals are extracted. Panel A shows sloped scattered data with a fitted line. Panel B shows the residuals from that same fitted model against the predictor. The fitted line is now horizontal because the linear trend has been removed.	209
11.2	Idealised residual-versus-fitted plot for a well-behaved straight-line model.	214
11.3	Idealised Q-Q plot of residuals for a well-behaved straight-line model.	215
11.4	Residuals versus fitted values for a pooled penguin body-mass model.	216
11.5	Q-Q plot of residuals for a pooled penguin body-mass model.	217
12.1	Wing length of sparrows at different ages.	229
12.2	Standard diagnostic plots for the sparrow simple linear regression. Clockwise from top left: residuals versus fitted values (linearity and constant variance), normal Q-Q plot (residual normality), residuals versus leverage (influential observations), and scale-location plot (homoscedasticity).	232
12.3	Wing length as a function of age in the sparrow example. The straight line is the fitted simple linear regression and the blue shading is the 95% confidence interval.	233
12.4	Scatter plot of the Palmer Station Adelie penguin data with a best fit line.	236
12.5	Diagnostic plots for the simple linear regression of Adelie penguin bill length on body mass. A) Normal Q-Q plot. B) Histogram of residuals. C) Residuals versus fitted values. D) Scale-location plot. . . .	238
12.6	Diagnostic plots for visual inspection of outliers in the Adelie penguin regression. A) DFFITS. B) Cook's distance. C) Residuals versus leverage. D) Cook's distance versus leverage. Observations beyond the Cook's distance threshold are shown in black and labelled by row number.	240
12.7	Scatter plot of Adelie penguin bill length against body mass, with observations exceeding the Cook's distance threshold highlighted. .	241
12.8	Adelie penguin bill length model with confidence interval (blue) and prediction interval (pink) around the fitted values.	245
13.1	CO ₂ uptake in chilled Quebec plants across concentration levels. The pattern is positive but clearly decelerating.	256

13.2	Residual-versus-fitted plot for the straight-line model. The curved smooth line confirms that the linear mean structure is insufficient. .	257
13.3	Standard diagnostics for the quadratic model.	261
13.4	Linear, quadratic, and cubic polynomial fits to the CO ₂ uptake data. .	262
14.1	Simple linear regressions fitted separately for five candidate climate predictors against mean Sørensen dissimilarity (Y) in the East Coast Transition Zone.	274
14.2	The null model fitted to Sørensen dissimilarity (Y) in the East Coast Transition Zone. Each point is one observation; the magenta line is the null model's single estimate — the grand mean of Y — and the pink band is its 95% confidence interval. Every residual is a vertical deviation from this horizontal line.	278
14.3	Added-variable (partial regression) plots for the selected model <code>mod1</code> . Each panel shows the relationship between the response and one predictor after removing the effects of the other two predictors. . . .	284
14.4	Standard diagnostic plots for <code>mod1</code> . Top left: residuals vs fitted values; top right: normal Q-Q; bottom left: scale-location; bottom right: residuals vs leverage.	285
14.5	Component-plus-residual plots for <code>mod1</code> . Each panel shows whether the relationship between the response and that predictor is approximately linear in the fitted model.	286
14.6	(A) A one-way ANOVA model ($Y \sim \text{bio}$) produces one horizontal fitted line per bioregion, sitting at each group's mean dissimilarity. Distance plays no role. (B) The additive model ($Y \sim \text{dist} + \text{bio}$) tilts all four lines by the same slope but preserves their vertical separation. The lines are strictly parallel: same gradient, different heights. The vertical gap between any two lines equals the difference in their bioregion coefficients.	292
14.7	Visualisation of the dummy-variable coding shown in Table 14.2. AMP (reference level) has all zeros; the three other bioregions each have a single 1 in their corresponding indicator column.	293
14.8	Main effects of (A) geographic distance along the coast and (B) bioregional classification on mean Sørensen dissimilarity. Points in panel A are coloured by bioregion to show how the distance–dissimilarity relationship maps onto regional structure.	296
14.9	Standard diagnostic plots for <code>full_mod3</code> . Top left: residuals vs fitted values; top right: normal Q-Q plot of residuals; bottom left: scale-location; bottom right: residuals vs leverage with Cook's distance contours.	302

14.10	Component-plus-residual plots for the three continuous predictors in <code>full_mod3</code> . A straight fitted line confirms that a linear term is adequate for that predictor after accounting for the others.	303
14.11	Results of the additive climate-plus-bioregion model. (A) Observed mean Sørensen dissimilarity plotted against August mean temperature, coloured by bioregion. Parallel lines show the partial effect of <code>augMean</code> with <code>febSD</code> and <code>augSD</code> held at their means; the shared slope reflects the additive structure of <code>full_mod3</code> . (B) Bioregion coefficients ($\hat{\gamma}_i$) with 95% confidence intervals, expressed as differences in baseline dissimilarity relative to the AMP reference level. A point to the right of the dashed line indicates higher mean dissimilarity than AMP at any given set of climate values.	305
15.1	Algal growth as a function of temperature at low, medium, and high nutrient levels. The fitted lines suggest that the temperature-growth slope steepens as nutrient concentration increases.	317
15.2	Sørensen dissimilarity as a function of coastal distance in four bioregions.	321
16.1	August mean temperature and annual mean temperature are tightly linearly related in the ECTZ subset. Two variables that fall along a line like this carry largely the same information; including both in a regression forces the model to partition the same signal between two parameters.	336
16.2	A confounding situation where nutrient concentration is associated with both temperature and productivity.	341
16.3	A) shows growth regressed against the true temperature, with a tight, steeply rising relationship. B) uses the same response but replaces the true predictor with a version with measurement noise of comparable magnitude to the signal itself. The slope is reduced and the cloud is visibly broader in the x-direction. The biology has not changed; only the measurement quality has.	346
17.1	Standard diagnostic panels for the selected three-predictor seaweed model. Clockwise from top left: residuals versus fitted values (linearity and constant variance), normal Q-Q plot (residual normality), residuals versus leverage with Cook's distance contours (influential observations), and scale-location plot (homoscedasticity).	354
17.2	Ten-fold cross-validation RMSE for the selected three-predictor seaweed model and a larger five-predictor alternative.	358
19.1	Landing distance by beach site for the sampled shorebirds.	383

19.2	Landing distance differs strongly among shorebird species, but values also vary among sites.	384
19.3	Standard residual diagnostics for the mixed model.	387
20.1	Presence and absence of acorn barnacles across the shore-position gradient.	398
20.2	Fitted logistic-regression curves for acorn-barnacle presence along the shore-position gradient.	400
20.3	Snail counts across the shore-position gradient.	403
20.4	Observed snail counts and fitted quasi-Poisson means across the shore-position gradient.	405
21.1	Monthly sea temperature at Port Nolloth.	412
21.2	Seasonal pattern in monthly temperature (across years).	413
21.3	Standard <code>mgcv</code> diagnostic checks for the fitted GAM.	416
21.4	Estimated smooth terms for the GAM: long-term trend and cyclic monthly seasonality.	417
21.5	Observed temperature and fitted GAM values through time.	417
22.1	Examples of curved models fitted to simulated data. A polynomial can capture modest curvature, a GAM can adapt flexibly to the data, and a mechanistic nonlinear model can represent a known saturating biological response.	424
22.2	Plot of uptake rate, V , as a function of substrate concentration, $[S]$, for a multiple flask experiment. The relationship is clearly curved and tends towards saturation.	426
22.3	Michaelis-Menten model fitted to the uptake data. The dashed lines indicate the estimated K_m and V_{max} values.	428
22.4	Basic diagnostics for the Michaelis-Menten fit. The residuals show no serious trend with the fitted values, and the distribution is approximately symmetric.	429
22.5	Simulated nutrient uptake curves for three treatments. The pattern suggests treatment-specific nonlinear responses rather than a single shared curve.	431
22.6	Repeated-measures uptake data from the perturbation experiment. Each dashed line traces one flask through time as substrate concentration declines.	433
22.7	Basic diagnostics for the nonlinear mixed-effects model. The residual pattern is acceptable for a teaching example, with no strong trend against the fitted values.	437

22.8	Treatment-specific nonlinear mixed-effects fits for the perturbation experiment.	438
23.1	Blade length as a function of stipe length in adult <i>Laminaria pallida</i> collected across 13 sites.	446
23.2	Ordinary least squares and quantile-regression fits for the 10th, 50th, and 90th quantiles.	449
23.3	Estimated slope of blade length on stipe length across quantiles from 0.1 to 0.9.	450
25.1	Cross-validation statistics for ridge regression applied to the seaweed data.	464
25.2	Coefficient paths for ridge regression. As the penalty increases, all coefficients are shrunk towards zero, but none are removed completely.	464
25.3	Coefficient paths for lasso regression. As the penalty increases, some coefficients are shrunk all the way to zero.	465
25.4	Cross-validation statistics for lasso regression applied to the seaweed data.	468
25.5	Cross-validation statistics for elastic net regression applied to the seaweed data.	471
26.1	A reproducibly generated figure showing blade length as a function of stipe length in the <i>Laminaria</i> dataset.	481

List of Tables

2.1	Data summary	31
2.2	Mean $\pm$ SD for the <code>ChickWeight</code> dataset as a function of Diet and Time.	33
8.1	Probability of at least one Type I error when all pairwise t -tests are run among K groups.	153
9.1	A pairwise Pearson correlation matrix of the <i>Ecklonia</i> dataset. . . .	182
12.1	Sparrow wing-length data used in the introductory simple linear regression example.	228
12.2	Size measurements for adult foraging Adelie penguins near Palmer Station, Antarctica.	235
13.1	A subset of the <code>C02</code> dataset showing chilled Quebec plants used in the polynomial-regression example.	255
14.1	Summary statistics for the five simple linear regressions of mean Sørensen dissimilarity (γ) against individual climate predictors in the East Coast Transition Zone. Slope and SE are the partial coefficient and its standard error. R^2 is the coefficient of determination; p is the overall model p -value from the F -test against the intercept-only null.	275
14.2	Dummy-variable coding for the four seaweed bioregions when AMP is the reference level. Each row shows the values of the three indicator variables for that bioregion. AMP is coded entirely as zeros, so its contribution is captured by the model intercept.	293
14.3	Dummy-variable coding, effective intercepts, and estimates for each bioregion in <code>mod2</code> , with AMP as the reference level. The full predicted $\hat{Y}$ for any observation adds the effective intercept to $\hat{\beta}_1 \cdot \text{dist}$, which is identical for all bioregions.	297
15.1	Dummy-variable coding for the interaction model $\gamma \sim \text{dist} * \text{bio}$, with AMP as the reference level. For each non-reference bioregion, exactly one intercept dummy and one slope dummy are switched on.	323

15.2	Symbolic effective intercepts and slopes for each bioregion in <code>mod_int</code> . The interaction coefficients $\hat{\delta}_i$ are slope <i>adjustments</i> relative to the reference slope, not bioregion slopes in their own right.	324
15.3	Numerical effective intercepts and slopes for each bioregion, computed from <code>summary(mod_int)</code> . Read the <code>dist</code> row for $\hat{\beta}_{\text{dist}}$ and the <code>dist:bioX</code> rows for the $\hat{\delta}_i$ adjustments, then add. Do not read the interaction rows as slopes — they are deviations from the AMP slope.	324
19.1	A subset of the <code>BeachBirds.csv</code> data used in the mixed-model example.	382
20.1	A subset of the sessile-benthic invertebrate data used for the logistic-regression example.	398
20.2	A subset of the small mobile invertebrate data used for the count-data GLM example.	402
21.1	A subset of the Port Nolloth temperature time series used in the GAM example.	412
22.1	Preview of the simulated multiple-flask dataset (first and last four rows).	426
22.2	Preview of the perturbation dataset (first eight rows).	433
23.1	A subset of the <code>laminaria.csv</code> data used in the quantile-regression example.	445
23.2	Coefficient summaries for the ordinary least squares fit and the 0.1, 0.5, and 0.9 quantile-regression fits.	448
26.1	A subset of the <code>laminaria.csv</code> dataset used in the reproducible-workflow example.	479
26.2	Grouped summary statistics for the Laminaria dataset, generated directly from code in the chapter source.	480
26.3	Minimal project components needed to regenerate the chapter from source.	482

Preface

Biostatistics is a work in progress. It is based on a collection of expanding lecture notes on statistics as used by biologists such as botanists, zoologists, and ecologists. It is intended to be a resource for students and researchers who are interested in learning more about the subject and applying it to their data. The material is presented in a way that is accessible to those with a thorough understanding of biology and a willingness and interest to learn the application of statistical analyses to their data.

Although the book is primarily aimed at biologists, it is also suitable for individuals in other fields who require subjecting their data to inferential statistical analyses. The material is presented in a way that is accessible even to those who might not have a deep interest or background in mathematics or statistics.

The examples in this book require that the reader already possesses a good working knowledge of the R programming language. If you are not familiar with R, I recommend that you take a look at the excellent book [R for Data Science](#) by Hadley Wickham and Garrett Grolemund. You may also study the material provided on [The Tangled Bank](#) website.

You are also required to already have a good grasp of exploratory data analysis (EDA), descriptive (summary) statistics, and data visualisation. If you are not familiar with these topics, I recommend that you take a look at the excellent book [Exploratory Data Analysis with R](#) by Roger D. Peng. [The Tangled Bank](#) is also there to help you.

The book is divided into several chapters, each of which covers a different aspect of statistics. The chapters are organised in a logical order, starting with the basics of probability and moving on to more advanced topics such as hypothesis testing, regression analysis, and multivariate statistics (eventually, maybe, and possible, but there are already excellent texts available on the topic so it is not a priority for me to write this material yet). Each chapter is accompanied by a set of exercises that are designed to help the reader understand the material and apply it to real-world problems.

Biostatistics shows you to connect i) the biological question, ii) the data you collect, and iii) the model you use to analyse it. When those three components fit together, your analysis becomes coherent and defensible. Throughout the book, therefore, you will see how to move from a question about a biological system to a structured analysis. This includes summarising data, selecting appropriate methods,

fitting models, checking whether those models are adequate, and communicating your findings to a scientific audience.

Part I

Foundations

This opening part establishes the theory that supports every later method. I begin with the relationship between a biological question, the data collected to answer it, and the statistical structure used to interpret those data. That means learning how to describe variation, how to visualise structure, and how to think about sampling uncertainty before any formal test is attempted.

The purpose is not to linger in preliminaries, but to make the later inferential and modelling chapters easier to read correctly. If the early descriptive work is weak, later inference becomes mechanical and unreliable. If the foundations are clear, the later methods become extensions of the same statistical reasoning rather than a list of unrelated procedures.

By the end of this part, you should be able to recognise what kind of data you have, describe their main features responsibly, and connect those features to the language of variability, sampling, and uncertainty.

1

The Statistical Landscape

We use statistics to describe variation, quantify uncertainty, compare groups, assess relationships, and build models that link observed patterns to underlying processes. We do this to *explain* natural phenomena.¹ In biology and ecology, statistical thinking *should* begin at the moment you ask a question. It shapes how you design a study, what data you collect, and how you interpret the results. The adage “rubbish in, rubbish out” applies here too. A rubbish experiment or sampling design will give you a rubbish interpretation of biological phenomena.

Answering questions about the natural world using the scientific method requires that we draw on many years of accumulated knowledge and experience. A scientific workflow practised by most biologists unpacks into roughly the following sequence:

1. **Observe a pattern or phenomenon.** Look around you at the world. Be curious about it. Ask questions to figure out an explanation for the pattern or phenomenon that tickled your interest. Children do this naturally and sadly this curiosity is often lost as we grow older. Maybe we are unfortunate to have parents that don’t encourage questions. Or school education is sufficiently soul-destroying, and the creative, questioning mindset is discouraged. But, as a scientist, you need to keep this curiosity alive. Scientists have the privilege of being paid to be curious, to ask questions, to encounter things which we cannot explain, and to find answers. To always occupy a position of uncertainty. This is a wonderful thing, and it is a privilege that should not be squandered.
2. **Questions and hypotheses.** Create an unambiguous statement of the question you want to answer, think about what is causing the pattern or phenomenon you observed, and consider how you might measure the response, the thing you

¹ It is physicist David Deutsch’s opinion that the primary purpose of science is to offer explanations for how things are.

observed initially. This leads to the first *formal* step in the scientific method; that is, the formulation of a hypothesis. A hypothesis is not a question. It is an unambiguous statement that can be disproved (in the frequentist view), or for which we can amass evidence in support of prior beliefs (the Bayesian interpretation of probability). The philosopher Sir Karl Popper argued that the key feature of a scientific hypothesis is its *falsifiability*, the capacity for it to be empirically tested and potentially refuted. For Popper, the progress of science depended not on the accumulation of irrefutable truths but on the iterative process of conjecture and refutation, thus subjecting hypotheses to testing and discarding them when contradicted by evidence. He emphasised the *fallibility* of scientific knowledge, asserting that no hypothesis can ever be definitively proven true, only provisionally accepted until it is falsified. This provisional nature of all scientific theories is central to scientific practice today. In Popper's view, the scientist's superpower lies in their willingness to embrace uncertainty and systematically eliminate errors, not in the infallibility of their conclusions.

3. **Design a study that can produce relevant data.** Design an experiment or sampling campaign to collect data that will allow you to test this hypothesis.
4. **Identify the variables and experimental units.** Clearly understand what the data you'll collect will look like, both for the response and the explanatory variables. For example, do you have a categorical or continuous predictor, is the response continuous, binary, ordinal, etc.? For this, you should have a firm grasp of the various kinds of [Data Classes and Structures in R](#).
5. **Consider alternative explanations and potential confounders.** Think deeply about any confounding influences that might affect your data, and specify exactly what additional data you will have to collect to isolate the hypothesised influence in your analysis. You need to fully understand all the ways that factors not considered in your hypothesis might affect your study's outcome. Omissions cannot be rectified after the fact without repeating the entire experiment or sampling work. It requires knowledge and experience to avoid confounding influences ruining your work.
6. **Select an analysis that matches the design and data structure.** Depending on your experiment's design (3) and the nature of the data you'll obtain (3, 4), choose the appropriate statistical methods to analyse them. You should be able to develop a good idea of what statistical methods you'll use, even before the experiment has been done! Decide on the parametric test, or, should the statistical god with the die not provide an outcome that favours your expectations, you can also decide upfront on a non-parametric equivalent. It is important not to decide on the statistical method after you've collected the data. This is called *p*-hacking, and it is almost a cardinal sin in science.

7. Do the experiment or go out into the world to sample, and collect the data. Have fun. This is why we do science, afterall!
8. Go have a few drinks after a hard day's work and celebrate your success.
9. **Analyse your newly-collected data.** This will include exploratory data analyses (see [Exploring With Summaries and Descriptions](#) and [Exploring With Figures](#)), and then the application of the statistical methods you chose in step 6.
10. **Communicate your results.** Your science will be pointless unless you communicate your findings and explanations to an audience. The first step in that process is to develop the necessary in tables and figures for inclusion in your manuscript.

So, what is clear is that the analysis does not start at the end. It begins at the design stage. If the design does not align with the question, the analysis will struggle to produce meaningful conclusions.

This textbook deals with many of these steps. Much of this knowledge is codified in the form of the statistical method, which provides a systematic framework for collecting,² analysing, interpreting, and presenting data. In this chapter, I will introduce the foundational concepts of *inferential statistics*, which allow you to make inferences about populations based on sample data.

You will encounter a whole list of statistical tests with seemingly arcane names. The emphasis here is not on memorising these named tests (although, eventually you'll get to know them). Instead, the goal is that you understand how different methods fit into a broader statistical analysis framework and how to choose among them in a principled way.

DO IT NOW!

A researcher notices that bird nestlings in urban parks weigh less than those in rural landscapes. Working in pairs, sketch the full sequence of steps (from initial observation to final communication) that would constitute a defensible investigation of this finding. For each step, write one decision that, if made incorrectly, could undermine the whole analysis. How many steps did you identify? Compare your sequence with another pair.

² Yes, statistics also informs us about how to collect data.

DO IT NOW!

Choose a research question you find interesting. Write down one clear research question about it. For that question, identify: (1) the **response variable** (what you would measure), (2) the **predictor(s)** (what you would manipulate or observe), and (3) whether each variable is continuous, discrete, or categorical. Compare your question with the person sitting next to you — do you agree on the variable types?

1.1 IMPORTANT CONCEPTS

A small number of ideas recur throughout the module:

- Statistical reasoning links the question, the nature of the study system, the data, and the model.
- Inference depends on how well the design, analysis, and interpretation align.
- Models provide structured descriptions of how a response varies with predictors.
- Explanation and prediction serve different goals and should be treated separately.
- Assumptions are part of the analysis and must be examined explicitly.

These ideas will reappear in different contexts.

DO IT NOW!

Look at the five key concepts listed above. For each one, write a single sentence giving a concrete example from biology or ecology (real or plausible). For instance, what does it mean in practice for a model to provide a “structured description of how a response varies with predictors”? What does it look like when assumptions are “examined explicitly”? You do not need to use R — just think through what each concept means in the context of real data.

1.2 CORE PRINCIPLES

The following practical rules guide analysis across all chapters:

- Start with a clear biological question before choosing a method.
- Understand how the data were generated and what the experimental units represent.
- Choose a model that reflects the process you aim to explain.
- Check assumptions directly and treat deviations as informative.
- Decide whether your goal is explanation or prediction before interpreting results.

These rules help you avoid treating statistical methods as isolated tools and instead use them as part of a coherent series of analytical decisions.

1.3 CORE SKILLS

By the end of this module, you should be able to:

- summarise and visualise biological data in a way that reveals structure and variability;
- select appropriate statistical methods for common biological questions;
- fit and interpret linear and generalised models;
- evaluate model assumptions and explain how violations affect inference;
- integrate analysis into a reproducible workflow;
- communicate methods and results clearly.

These skills are cumulative and each chapter adds a layer. Your goal is to develop your understanding into a consistent way of thinking about your data, your models, and your biological questions. You should be able to put what you learned into action, and you will be assessed in the capstone module and in the research project (and BCB744-specific assessments too, of course).

DO IT NOW!

Rate your current confidence in each of the six core skills listed above on a scale from 1 (cannot do at all) to 5 (fully confident). Be honest — this is for your own reference. Write the date next to your ratings. At the end of the module, revisit this list and re-rate yourself. What changed? Which skills improved the most?

1.4 SECTION STRUCTURE

If you are new to statistics, follow the parts in order. If you have prior training, you may use early chapters as review before focusing on inference and modelling. The material is organised into five parts with each addressing a different type or stage of statistical application:

1.4.1 *Part I. Foundations*

These are the fundamental concepts of data description, variation, and uncertainty:

- **The Statistical Landscape**
- **Summarising Biological Data**
- **Visualising Data**
- **Distributions, Sampling, and Uncertainty**

1.4.2 *Part II. Inference*

The formal inference tests for comparing groups and assessing associations is introduced:

- [Statistical Inference](#)
- [Raw-Data Assumptions and Transformations](#)
- [t-Tests](#)
- [ANOVA](#)
- [Correlation and Association](#)
- [Choosing the Right Test](#)

1.4.3 *Part III. Modelling*

This part is about models that explain variation or to predict from our data:

- [Residuals and Model-Based Diagnostics](#)
- [Simple Linear Regression](#)
- [Polynomial Regression](#)
- [Multiple Regression and Model Specification](#)
- [Interaction Effects](#)
- [Collinearity, Confounding, and Measurement Error](#)
- [Model Checking and Evaluation](#)

1.4.4 *Part IV. Extensions*

Here I address non-independence, flexible model structures, and extended modelling frameworks — these are beyond the basic toolkit:

- [Pseudoreplication](#)
- [Dependence and Mixed Models](#)
- [Generalised Linear Models](#)
- [Generalised Additive Models](#)
- [Nonlinear Regression](#)
- [Quantile Regression](#)
- [Prediction and Explanation](#)
- [Regularisation](#)

1.4.5 *Part V. Reproducibility*

Statistical analysis ties to transparent and repeatable scientific workflows (a part of reproducible research), which is discussed here:

- [Reproducible Workflow](#)

1.5 ASSESSMENT AND PRACTICE

Use the assessment materials alongside the chapter sequence:

1.5.1 [Biostatistics Self-Assessment](#)

Use this to check conceptual understanding after completing each major block.

1.5.2 [Biostatistics Example 1](#)

Use this to practise integrated analysis and interpretation across multiple concepts.

Consistent practice and reflection are essential for developing reliable statistical judgement.

2

Summarising Biological Data

The First Steps of Exploratory Data Analysis

“I think it is much more interesting to live with uncertainty than to live with answers that might be wrong.”

- — Richard Feynman

IN THIS CHAPTER

- Data summaries
- Measures of central tendency
- Measures of dispersal
- Descriptive statistics by group

TASKS TO COMPLETE IN THIS CHAPTER

- Self-Assessment Task 2-1 (/3)
- Self-Assessment Task 2-2 (/10)
- Self-Assessment Task 2-3 (/5)
- Self-Assessment Task 2-4 (/10)
- Self-Assessment Task 2-5 (/5)
- Self-Assessment Task 2-6 (/3)
- Self-Assessment Task 2-7 (/2)
- [Self-Assessment instructions and full task overview](#)

Exploratory data analysis (EDA) establishes what kind of data we actually have before we test hypotheses or fit models. It tells us how many observations we have, what variables were measured, how much variation is present, whether missing

values or outliers need attention, and whether the data contain obvious grouping structure. Poor summaries at this stage usually lead to poor inference later.

In this chapter, I introduce the numerical summaries used at the start of that process. I focus on measures of centre, measures of spread, and a small set of tools for inspecting the structure of a dataset. These summaries work alongside the figures in [Chapter 3](#) and the treatment of distributions and uncertainty in [Chapter 4](#).

2.1 KEY CONCEPTS

The ideas that organise the rest of the chapter are:

- **EDA:** The first analytical pass through a dataset. Its job is to reveal structure, variation, missingness, and potential problems before formal inference begins.
- **Centre and spread:** Most numerical summaries describe either where the data cluster or how widely they vary. They form part of summary (descriptive) statistics.
- **Distribution:** The shape of a distribution determines which summaries describe it well. Symmetric data and skewed data are not best summarised in the same way, and they will be analysed using different tests or models.
- **Grouped summaries:** Biological datasets often contain treatments, species, sites, or times. Summaries by group may reveal structure hidden in pooled data.

2.2 FOUNDATIONAL DEFINITIONS

A few definitions that recur throughout the module are presented next.

2.2.1 *Variables and Parameters*

A **parameter** is a fixed but usually unknown quantity that describes a population or probability distribution. Examples include the true population mean, variance, or regression slope. A **variable** is the measured characteristic itself: body mass, bill length, temperature, salinity, presence or absence, and so on. We observe variables in a sample and use them to estimate parameters.

2.2.2 *Samples and Populations*

The **population** is the full set of units about which we want to make a claim: all trees in a forest, all quadrats in a marsh, all fish in an estuary, or all patients in a trial. A **sample** is the subset we actually measure. Because we rarely have access to the whole population, we rely on a **random** sample to estimate population-level quantities. The question is therefore whether the sample is informative about the

population we care about. As sample size increases, and as the sample better represents the population, the sample mean becomes a more stable estimate of the population mean (Figure 2.1).

```
set.seed(666)

# pre-allocate the tibble
normal_takes_shape <- tibble(number_draws = c(), draws = c())

# simulate increasingly larger samples
for (i in c(2, 5, 10, 50, 100, 500, 1000, 10000, 100000)) {
  draws_i <-
    tibble(
      number_draws = c(rep.int(
        x = paste(as.integer(i), " draws"),
        times = i
      )),
      draws = c(rnorm(
        n = i,
        mean = 13,
        sd = 1
      ))
    )

  normal_takes_shape <- rbind(normal_takes_shape, draws_i)
  rm(draws_i)
}

normal_takes_shape >
  mutate(number_draws = as_factor(number_draws)) >
  ggplot(aes(x = draws)) +
  geom_density(colour = "indianred3") +
  theme_grey() +
  facet_wrap(
    vars(number_draws),
    scales = "free_y"
  ) +
  labs(
    x = "Mean",
    y = "Density"
  )
```

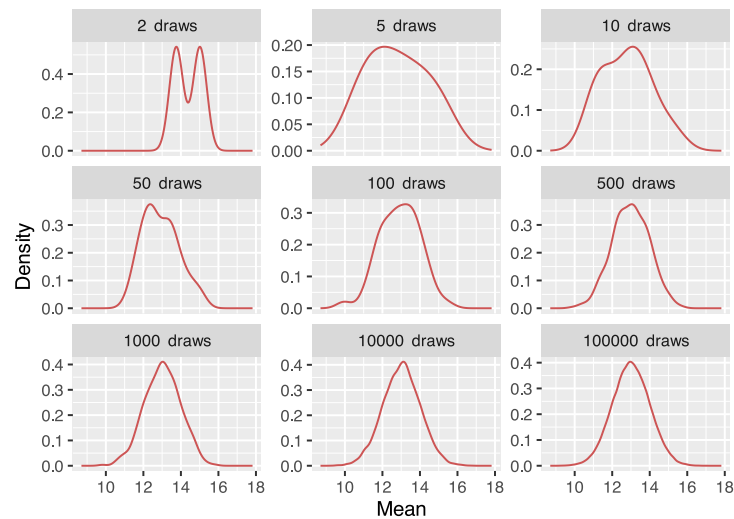


FIGURE 2.1. Drawing increasingly larger sample sizes from a population with a true mean of 13 and an SD of 1.

2.2.3 When Is Something Random?

In statistics, **random** means that outcomes cannot be predicted exactly in advance, even when the process generating them is understood. Random sampling and random assignment use this idea to reduce bias and to justify probability-based inference.

The term **stochastic** is closely related but usually refers to a process rather than to a single outcome. Population growth under fluctuating weather, disease transmission, and dispersal all have deterministic components, but they also include variation that is modelled probabilistically. In practice, both terms point us to the same issue: biological systems often contain uncertainty that must be described rather than ignored.

DO IT NOW!

Run the code for Figure 2.1 and compare the panels for 2, 10, and 10 000 draws in terms of their mean and SD. If you forgot the functions, you can find them below. In two or three sentences, explain what changes as sample size increases and why that is important when we use a sample to estimate a population mean.

2.3 DESCRIPTIVE STATISTICS

Now to the summaries used most often in EDA. These summaries answer three basic questions: where is the centre of the data, how much do the values vary, and which summaries remain sensible when the data are skewed or contain outliers.

2.3.1 *The Core Equations*

The sample mean is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (2.1)$$

In Equation 2.1, $x_1, x_2, \dots, x_n$ are the observations and n is the sample size. The mean is therefore the total of the observations divided by the number of observations.

The sample variance is:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.2)$$

Equation 2.2 measures the average squared deviation from the sample mean. The divisor is $n - 1$ rather than n because we are estimating population variance from a sample.

The sample standard deviation is the square root of the variance:

$$S = \sqrt{S^2} \quad (2.3)$$

The standard deviation is usually easier to interpret than the variance because it is expressed on the original measurement scale of the data.

For robust summaries, the interquartile range is:

$$\text{IQR} = Q_3 - Q_1 \quad (2.4)$$

In Equation 2.4, Q_1 is the first quartile and Q_3 is the third quartile, so the IQR measures the spread of the middle 50% of the data.

2.3.2 *Measures of Central Tendency*

Statistic	Function	Package
Mean	<code>mean()</code>	base

Statistic	Function	Package
Median	median()	base
Mode	Do it!	
Skewness	skewness()	e1071
Kurtosis	kurtosis()	e1071

Measures of central tendency describe where the data cluster. The mean and standard deviation work well when the data are roughly symmetric and not dominated by extreme values. The median and IQR are usually better when the data are skewed or contain outliers.

Before discussing each statistic, I will generate several simple datasets with different shapes. These give us a controlled way to compare the summaries.

```
# Generate random data from a normal distribution
set.seed(666)
n <- 5000 # Number of data points
mean <- 0
sd <- 1
normal_data <- rnorm(n, mean, sd)

# Generate random data from a slightly
# right-skewed beta distribution
alpha <- 2
beta <- 5
right_skewed_data <- rbeta(n, alpha, beta)

# Generate random data from a slightly
# left-skewed beta distribution
alpha <- 5
beta <- 2
left_skewed_data <- rbeta(n, alpha, beta)

# Generate random data with a bimodal distribution
mean1 <- 0
mean2 <- 10
sd1 <- 3
sd2 <- 4

# Generate data from two normal distributions
data1 <- rnorm(n, mean1, sd1)
data2 <- rnorm(n, mean2, sd2)

# Combine the data from both distributions to
# create a bimodal distribution
bimodal_data <- c(data1, data2)

make_hist_plot <- function(x, title, fill_col) {
  stat_lines <- tibble(
    statistic = c("Mean", "Median"),
    xint = c(mean(x), median(x))
  )

  ggplot(tibble(value = x), aes(x = value)) +
    geom_histogram(bins = 30, fill = fill_col, colour = "black", linewidth = 0.3) +
    geom_vline(
```

```

    data = stat_lines,
    aes(xintercept = xint, colour = statistic, linetype = statistic),
    linewidth = 0.5,
    show.legend = TRUE
  ) +
  scale_colour_manual(values = c("Mean" = "red", "Median" = "blue")) +
  scale_linetype_manual(values = c("Mean" = "solid", "Median" = "dashed")) +
  labs(
    title = title,
    x = "Value",
    y = "Frequency"
  ) +
  theme_grey() +
  theme(
    legend.position = "bottom",
    plot.title = element_text(size = 9)
  )
}

plt_normal <- make_hist_plot(normal_data, "Normal Distribution", "grey80")
plt_right <- make_hist_plot(right_skewed_data, "Right-Skewed Distribution", "grey80")
plt_left <- make_hist_plot(left_skewed_data, "Left-Skewed Distribution", "grey80")
plt_bimodal <- make_hist_plot(bimodal_data, "Bimodal Distribution", "grey80")

ggpubr::ggarrange(
  plt_normal, plt_right, plt_left, plt_bimodal,
  ncol = 2, nrow = 2,
  labels = c("A", "B", "C", "D"),
  common.legend = TRUE,
  legend = "bottom"
)

```

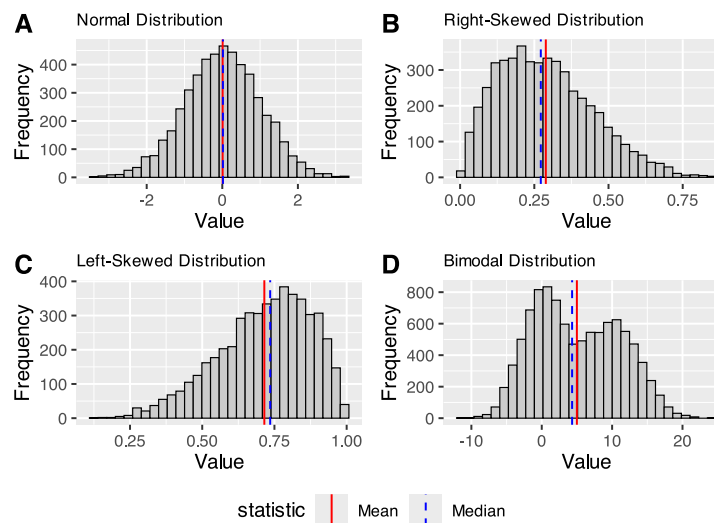


FIGURE 2.2. Generated normal, right-skewed, left-skewed, and bimodal distributions, shown in panels A-D, with the mean and median indicated in each panel.

a) The Sample Mean. The mean is the arithmetic average of the data. As shown in Equation 2.1, it is calculated by summing the observations and dividing by the sample size.

We calculate it with `mean()`:

```
round(mean(normal_data), 3)
```

```
R> [1] 0.009
```

SELF-ASSESSMENT TASK 2-1

How would you manually calculate the mean value for the `normal_data` we generated in the lecture? (/3)

The mean uses all observed values, which makes it informative but also sensitive to skew and outliers. That sensitivity does not make it invalid for non-normal data, but it does make it a poor summary when a few extreme values dominate the result.

In panel A of Figure 2.2, the normal distribution is centred cleanly around its mean. In panels B and C, tail asymmetry pulls the mean away from the bulk of the data.

b) The Median. The median is the middle value after the data have been ordered. With an odd number of observations, it is the single central value. With an even number, it is the mean of the two central values.

The median divides the ordered data into two equal halves. In a symmetric distribution it will often be close to the mean. In skewed data it usually gives a more stable description of the centre because extreme values have little influence on it.

That contrast is visible in panels B and C of Figure 2.2 where the median sits closer to the main cluster of values than the mean does.

We calculate the median with `median()`:

```
round(median(normal_data), 3)
```

```
R> [1] 0.017
```

It is easier to see the calculation on a small dataset:

```
set.seed(123) # for reproducibility
small_normal_data <- round(rnorm(11, 13, 3), 1)
sort(small_normal_data)
```

```
R> [1] 9.2 10.9 11.3 11.7 12.3 13.2 13.4 14.4 16.7 17.7 18.1
```

```
median(small_normal_data)
```

```
R> [1] 13.2
```

DO IT NOW!

Use Figure 2.2 to decide which measure of centre you would report for each of the four panels. For each panel, write down either mean or median and give one short reason for your choice.

WHAT IS THE RELATIONSHIP BETWEEN THE MEDIAN AND QUANTILES?

The median is the 50th percentile, or second quartile (Q_2). More generally, quantiles divide the ordered data into specified proportions. Quartiles split them into four parts, deciles into ten, and percentiles into one hundred.

c) *The Mode.* The mode is the most frequent value or values in a dataset. It is useful for categorical data and for identifying whether a distribution is unimodal, bimodal, or multimodal. For continuous numerical data, exact repeated values are often less informative, so the mode is usually assessed visually from a histogram or density plot rather than calculated directly.

Panel D of Figure 2.2 shows why visual inspection is useful here: one mean can be calculated, but the figure shows two clear peaks.

Base R does not provide a standard `mode()` function for this purpose. In practice, visual inspection is often the more useful route.

d) *Skewness.* Skewness describes asymmetry in a distribution. A symmetric distribution has skewness close to zero. Positive skewness means the right tail is longer; negative skewness means the left tail is longer.

Skewness is often easiest to understand by comparing the mean and median. In a right-skewed distribution the mean is usually greater than the median. In a left-skewed distribution the mean is usually smaller.

You can see both patterns in Figure 2.2, where panel B has a longer right tail, and panel C has a longer left tail.

```
library(e1071)
# Positive skewness
skewness(right_skewed_data)
```

```
R> [1] 0.5453162
```

```
# Is the mean larger than the median?
mean(right_skewed_data) > median(right_skewed_data)
```

```
R> [1] TRUE
```

```
# Negative skewness
skewness(left_skewed_data)
```

```
R> [1] -0.5790834
```

```
# Is the mean less than the median?
mean(left_skewed_data) < median(left_skewed_data)
```

```
R> [1] TRUE
```

e) Kurtosis. Kurtosis describes tail heaviness relative to a normal distribution. A normal distribution has close to zero kurtosis (called mesokurtic). Negative kurtosis indicates data with a thin-tailed (platykurtic) distribution and positive kurtosis indicates a fat-tailed distribution (leptokurtic).

```
kurtosis(normal_data)
```

```
R> [1] -0.01646261
```

```
kurtosis(right_skewed_data)
```

```
R> [1] -0.1898941
```

```
kurtosis(left_skewed_data)
```

```
R> [1] -0.1805365
```

SELF-ASSESSMENT TASK 2-2

Find the *faithful* dataset and describe both variables in terms of their measures of central tendency. Include graphs in support of your answers (use `ggplot()`), and conclude with a brief statement about the data distribution. (/10)

Skewness and kurtosis can be informative, but they do not replace visual inspection or later assumption checks. They give a first numerical impression of distribution shape.

2.3.3 Measures of Variance or Dispersion Around the Centre

Statistic	Function
Variance	<code>var()</code>
Standard deviation	<code>sd()</code>
Minimum	<code>min()</code>
Maximum	<code>max()</code>
Range	<code>range()</code>
Quantile	<code>quantile()</code>
Inter Quartile Range	<code>IQR()</code>

Measures of dispersion describe how widely the values are spread. Two samples can have the same mean but very different biological interpretations if one is tightly clustered and the other highly variable.

a) Variance and Standard Deviation. Variance and standard deviation are measures of dispersion. The sample variance is given in Equation 2.2, and the standard deviation in Equation 2.3. We can calculate them with `var()` and `sd()`:

```
var(normal_data)
```

```
R> [1] 1.002459
```

```
sd(normal_data)
```

```
R> [1] 1.001229
```

SELF-ASSESSMENT TASK 2-3

Manually calculate the variance and SD for the `normal_data` we generated in the lecture. Make sure your answer is the same as those reported there. (/5)

The standard deviation is easier to interpret than the variance because it is measured on the same scale as the data. If temperature is measured in degrees Celsius, the standard deviation is also measured in degrees Celsius.

For roughly normal data, the 68-95-99.7 rule gives a useful approximation: about 68% of observations lie within 1 SD of the mean, about 95% within 2 SD, and about 99.7% within 3 SD (Figure 4.5).

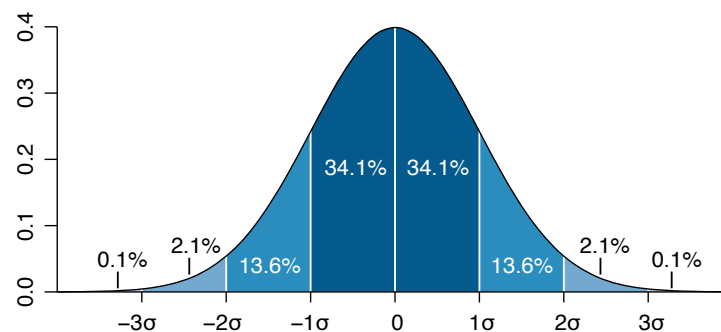


FIGURE 2.3. The idealised Normal distribution showing the proportion of data within 1, 2, and 3SD from the mean.

Like the mean, the standard deviation is sensitive to extreme values. For skewed data or data with strong outliers, the IQR is often more informative.

b) The Minimum, Maximum, and Range. `min()`, `max()`, and `range()` give the extremes of the data:

```
min(normal_data)
R> [1] -3.400137
max(normal_data)
R> [1] 3.235566
range(normal_data)
R> [1] -3.400137 3.235566
```

`range()` returns the minimum and maximum as a pair. If we want the numerical width of the range, we subtract the minimum from the maximum:

```
range(normal_data)[2] - range(normal_data)[1]
R> [1] 6.635703
```

These are simple summaries, but they are often the first place to look for impossible values, obvious outliers, or coding errors.

c) Quartiles and the Interquartile Range. Quartiles divide the ordered data into quarters. The first quartile (Q_1) marks the point below which 25% of the data fall, the second quartile (Q_2) is the median, and the third quartile (Q_3) marks the point below which 75% fall.

The IQR measures the spread of the middle 50% of the data. Because it ignores the tails, it is much less sensitive to outliers than the standard deviation. It is often the better description of spread for skewed data.

We obtain quartiles with `quantile()`:

```
# Look at the normal data
quantile(normal_data, p = 0.25)
R>      25%
R> -0.6597937

quantile(normal_data, p = 0.75)
R>      75%
R> 0.6840946

# Look at skewed data
quantile(left_skewed_data, p = 0.25)
R>      25%
R> 0.6133139

quantile(left_skewed_data, p = 0.75)
R>      75%
R> 0.8390202
```

We calculate the IQR with `IQR()`:


```
IQR(normal_data)
```

```
R> [1] 1.343888
```

DO IT NOW!

Calculate the range and IQR for `left_skewed_data`. The code is just above Figure 2.2. Then add one very large value to that vector and calculate both summaries again. Which measure changes more, and what does that tell you about when range or IQR is the better description of spread?

SELF-ASSESSMENT TASK 2-4

Write a few lines of code to demonstrate that the $(0 - 0.25]$, $(0.25 - 0.5]$, $(0.5 - 0.75]$, $(0.75 - 1]$ quantiles of the `normal_data` we generated in the lecture indeed conform to the formal definition for what quantiles are. I.e., show manually how you can determine that 25% of the observations indeed fall below -0.66 for the `normal_data`. Explain the rationale to your approach. (/10)

The choice between mean and SD on the one hand, and median and IQR on the other, depends on the data. Symmetric distributions are often well described by the first pair. Skewed distributions or data with strong extremes are usually better described by the second.

The contrast among panels A-D in Figure 2.2 shows why that choice depends on the distribution rather than on habit.

2.4 THE PALMER PENGUIN DATASET

The [Palmer penguin dataset](#) in the **palmerpenguins** package is a widely used teaching dataset for data exploration, visualisation, and modelling. It contains measurements from three penguin species in the Palmer Archipelago: Adélie, Chinstrap, and Gentoo.

The variables include bill length (`bill_length_mm`), bill depth (`bill_depth_mm`), flipper length (`flipper_length_mm`), body mass (`body_mass_g`), species, island, and sex. The dataset is rich enough to illustrate grouping structure, missingness, and both numerical and categorical variables without being unnecessarily large.

Let us start by loading the data:

```
library(palmerpenguins)
```

2.5 EXPLORING THE DATA STRUCTURE

Now we go from statistical definitions to implementation. These functions answer three practical questions: what variables exist, how large the dataset is, and whether the data contain missing values or unexpected structure.

2.5.1 *Inspecting Type and Layout*

Several functions give a quick overview of the dataset itself rather than of the values inside it:

Purpose	Function
The class of the dataset	<code>class()</code>
The head of the dataframe	<code>head()</code>
The tail of the dataframe	<code>tail()</code>
Printing the data	<code>print()</code>
Glimpse the data	<code>glimpse()</code>
Show number of rows	<code>nrow()</code>
Show number of columns	<code>ncol()</code>
The column names	<code>colnames()</code>
The row names	<code>row.names()</code>
The dimensions	<code>dim()</code>
The dimension names	<code>dimnames()</code>
The data structure	<code>str()</code>

First, check the class of the object. The penguins dataset is a tibble, which is the tidyverse version of a data frame:

```
class(penguins)
R> [1] "tbl_df"      "tbl"        "data.frame"
```

We can convert between tibbles and data frames with `as.data.frame()` and `as_tibble()`:

```
penguins_df <- as.data.frame(penguins)
class(penguins_df)
R> [1] "data.frame"

penguins_tbl <- as_tibble(penguins_df)
class(penguins_tbl)
R> [1] "tbl_df"      "tbl"        "data.frame"
```

The print methods differ. Data frames print more bluntly; tibbles are more compact and readable:

```
print(penguins_df[1:5,])
```

```
R> species island bill_length_mm bill_depth_mm flipper_length_mm
R> 1 Adelie Torgersen 39.1 18.7 181
R> 2 Adelie Torgersen 39.5 17.4 186
R> 3 Adelie Torgersen 40.3 18.0 195
R> 4 Adelie Torgersen NA NA NA
R> 5 Adelie Torgersen 36.7 19.3 193
R> body_mass_g sex year
R> 1 3750 male 2007
R> 2 3800 female 2007
R> 3 3250 female 2007
R> 4 NA <NA> 2007
R> 5 3450 female 2007
```

```
print(penguins_tbl)
```

```
R> # A tibble: 344 x 8
R> species island bill_length_mm bill_depth_mm flipper_length_mm
R> <fct> <fct> <dbl> <dbl> <int>
R> 1 Adelie Torgersen 39.1 18.7 181
R> 2 Adelie Torgersen 39.5 17.4 186
R> 3 Adelie Torgersen 40.3 18 195
R> 4 Adelie Torgersen NA NA NA
R> 5 Adelie Torgersen 36.7 19.3 193
R> 6 Adelie Torgersen 39.3 20.6 190
R> 7 Adelie Torgersen 38.9 17.8 181
R> 8 Adelie Torgersen 39.2 19.6 195
R> 9 Adelie Torgersen 34.1 18.1 193
R> 10 Adelie Torgersen 42 20.2 190
R> # i 334 more rows
R> # i 3 more variables: body_mass_g <int>, sex <fct>, year <int>
```

`glimpse()` gives similar information in a horizontal layout:

```
glimpse(penguins)
```

```
R> Rows: 344
R> Columns: 8
R> $ species <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie,~
R> $ island <fct> Torgersen, Torgersen, Torgersen, Torgersen, Tor~
R> $ bill_length_mm <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 3~
R> $ bill_depth_mm <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 1~
R> $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190~
R> $ body_mass_g <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3~
R> $ sex <fct> male, female, female, NA, female, male, female,~
R> $ year <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,~
```

2.5.2 Inspecting Size and Names

Use `nrow()`, `ncol()`, and `dim()` to check dataset size:

```
nrow(penguins)
```

```
R> [1] 344
```

```
ncol(penguins)
```

```
R> [1] 8
```

```
dim(penguins)
```

```
R> [1] 344 8
```

Use `colnames()` to inspect variable names:

```
colnames(penguins)
```

```
R> [1] "species"      "island"        "bill_length_mm"
R> [4] "bill_depth_mm" "flipper_length_mm" "body_mass_g"
R> [7] "sex"          "year"
```

Tibbles do not use row names in the same way as traditional data frames, but `row.names()` and `dimnames()` are still worth recognising:

DO IT NOW!

Explain the output of `dimnames()` when applied to the penguins dataset.

SELF-ASSESSMENT TASK 2-5

- Explain the output of `dimnames()` when applied to the penguins dataset. (/2)
- Explain the output of `str()` when applied to the penguins dataset. (/3)

2.5.3 Previewing the Data

`head()` and `tail()` let us inspect the first and last rows:

```
head(penguins)
```

```
R> # A tibble: 6 x 8
R>   species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
R>   <fct>   <fct>         <dbl>         <dbl>           <int>     <int>
R> 1 Adelie  Torge~         39.1           18.7             181       3750
R> 2 Adelie  Torge~         39.5           17.4             186       3800
R> 3 Adelie  Torge~         40.3           18              195       3250
R> 4 Adelie  Torge~         NA              NA              NA         NA
R> 5 Adelie  Torge~         36.7           19.3             193       3450
R> 6 Adelie  Torge~         39.3           20.6             190       3650
R> # i 2 more variables: sex <fct>, year <int>
```

```
tail(penguins, n = 3)
```

```
R> # A tibble: 3 x 8
R>   species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
R>   <fct>   <fct>         <dbl>         <dbl>           <int>     <int>
R> 1 Chinst~ Dream         49.6           18.2             193       3775
R> 2 Chinst~ Dream         50.8           19              210       4100
R> 3 Chinst~ Dream         50.2           18.7             198       3775
R> # i 2 more variables: sex <fct>, year <int>
```

You can wrap them in `print()` if you want more control over display:

```
print(head(penguins))
```

```
R> # A tibble: 6 x 8
R>   species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
R>   <fct>   <fct>         <dbl>         <dbl>           <int>     <int>
R> 1 Adelie  Torge~         39.1           18.7             181       3750
R> 2 Adelie  Torge~         39.5           17.4             186       3800
R> 3 Adelie  Torge~         40.3           18              195       3250
```

```
R> 4 Adelie Torge~      NA      NA      NA      NA
R> 5 Adelie Torge~     36.7     19.3     193     3450
R> 6 Adelie Torge~     39.3     20.6     190     3650
R> # i 2 more variables: sex <fct>, year <int>

print(tail(penguins, n = 3))

R> # A tibble: 3 x 8
R>   species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
R>   <fct>   <fct>       <dbl>         <dbl>           <int>         <int>
R> 1 Chinst~ Dream      49.6          18.2             193         3775
R> 2 Chinst~ Dream      50.8          19              210         4100
R> 3 Chinst~ Dream      50.2          18.7             198         3775
R> # i 2 more variables: sex <fct>, year <int>
```

`str()` is often the most compact first inspection because it shows object type, variable classes, and a preview of values:

```
str(penguins)

R> tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
R> $ species      : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1
...
R> $ island       : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 ...
R> $ bill_length_mm : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
R> $ bill_depth_mm  : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
R> $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
R> $ body_mass_g    : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
R> $ sex           : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
R> $ year          : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

DO IT NOW!

Explain the output of `str()` when applied to the penguins dataset.

2.6 DATA SUMMARIES

Once we understand the structure of the dataset, we summarise the values it contains. The tools in this section automate the numerical descriptions introduced above.

Use them for slightly different purposes:

- `summary()` for a quick overview of variable types and basic summaries.
- `skim()` for a broader inspection that includes missingness and type-specific summaries.
- `describe()` for more detailed descriptive statistics on numerical variables.
- `descriptives()` and `dfSummary()` when you want more elaborate tabular output.

Purpose	Function	Package
Summary of the data properties	<code>summary()</code>	base
	<code>describe()</code>	psych

Purpose	Function	Package
	skim()	skimr
	descriptives()	jmv
	dfSummary()	summarytools

2.6.1 summary()

`summary()` is the standard quick overview in base R. For data frames and tibbles, it reports variable classes and a small set of numerical summaries:

```
summary(penguins)
```

```
R>      species      island bill_length_mm bill_depth_mm
R> Adelie   :152   Biscoe   :168   Min.    :32.10   Min.    :13.10
R> Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
R> Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
R>                                     Mean    :43.92   Mean    :17.15
R>                                     3rd Qu.:48.50   3rd Qu.:18.70
R>                                     Max.    :59.60   Max.    :21.50
R>                                     NA's    :2      NA's    :2
R> flipper_length_mm body_mass_g sex      year
R> Min.    :172.0     Min.    :2700 female:165 Min.    :2007
R> 1st Qu.:190.0     1st Qu.:3550 male  :168  1st Qu.:2007
R> Median :197.0     Median :4050 NA's  : 11  Median :2008
R> Mean    :200.9     Mean    :4202          Mean :2008
R> 3rd Qu.:213.0     3rd Qu.:4750          3rd Qu.:2009
R> Max.    :231.0     Max.    :6300          Max.    :2009
R> NA's    :2        NA's    :2
```

SELF-ASSESSMENT TASK 2-6

Explain the output of `summary()` when applied to the penguins dataset. (/3)

2.6.2 psych::describe()

`psych::describe()` provides a more detailed numerical summary:

```
psych::describe(penguins)
```

```
R>      vars  n   mean    sd median trimmed   mad   min
R> species*  1 344   1.92  0.89   2.00   1.90   1.48   1.0
R> island*   2 344   1.66  0.73   2.00   1.58   1.48   1.0
R> bill_length_mm 3 342  43.92  5.46  44.45  43.91   7.04  32.1
R> bill_depth_mm  4 342  17.15  1.97  17.30  17.17   2.22  13.1
R> flipper_length_mm 5 342 200.92 14.06 197.00 200.34  16.31 172.0
R> body_mass_g    6 342 4201.75 801.95 4050.00 4154.01 889.56 2700.0
R> sex*          7 333   1.50  0.50   2.00   1.51   0.00   1.0
R> year         8 344 2008.03  0.82 2008.00 2008.04   1.48 2007.0
R>                                     max range skew kurtosis se
R> species*    3.0   2.0  0.16   -1.73  0.05
R> island*     3.0   2.0  0.61   -0.91  0.04
```

```

R> bill_length_mm      59.6   27.5  0.05   -0.89  0.30
R> bill_depth_mm       21.5    8.4 -0.14   -0.92  0.11
R> flipper_length_mm  231.0   59.0  0.34   -1.00  0.76
R> body_mass_g        6300.0 3600.0  0.47   -0.74 43.36
R> sex*                2.0    1.0 -0.02   -2.01  0.03
R> year                2009.0  2.0 -0.05   -1.51  0.04

```

2.6.3 skimr::skim()

skim() adds type-specific summaries and a clearer account of missingness:

```

library(skimr)
skim(penguins)

```

TABLE 2.1. Data summary

Name	penguins
Number of rows	344
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

Variable type: factor

skim_vari- able	n_miss- ing	com- plete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_vari- able	n_miss- ing	com- plete_rate	mean	sd	po	p25	p50	p75	p100	hist
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6	- ##*.
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5	++##:

skim_vari- able	n_miss- ing	com- plete_rate	mean	sd	po	p25	p50	p75	p100	hist
flip- per_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0	:#- +:
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0	-#* :
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0	#.#.#

2.6.4 `summarytools::dfSummary()`

`dfSummary()` from **summarytools** produces a richer tabular report:

The `dfSummary()` example is omitted from the PDF edition because its HTML-style rendered table does not convert reliably to LaTeX in a collected book build.

No single tool is best in every situation. The important point is to know what each one is for and to use it deliberately rather than mechanically.

2.7 DESCRIPTIVE STATISTICS BY GROUP

DO IT NOW!

Before looking at the grouped `ChickWeight` summaries, identify the variables that define the grouping structure in that dataset. Then propose one grouped table and one grouped figure that would help you compare chicks in a biologically sensible way.

SELF-ASSESSMENT TASK 2-7

Why is it important to consider the grouping structures that might be present within our datasets? (/2)

Whole-dataset summaries are only a starting point. Biological data are often structured by species, treatment, site, sex, season, or year, and those groupings are often more informative than the overall mean. Once we acknowledge that structure, the descriptive question changes from “What is the average?” to “Average for whom, under what condition, and with what spread?”

The `ChickWeight` data make the point clearly. A single mean across all chicks, all diets, and all sampling days hides the fact that the birds were measured repeatedly and assigned to different diets. It is much more informative to summarise weight within diet groups and at specific time points, for example at the start and end of the experiment. That lets us compare means with standard deviations, medians

TABLE 2.2. Mean $\pm$ SD for the `ChickWeight` dataset as a function of Diet and Time.

Diet	Time	Mean $\pm$ SD
1	0	41.4 $\pm$ 1
1	21	177.8 $\pm$ 58.7
2	0	40.7 $\pm$ 1.5
2	21	214.7 $\pm$ 78.1
3	0	40.8 $\pm$ 1
3	21	270.3 $\pm$ 71.6
4	0	41 $\pm$ 1.1
4	21	238.6 $\pm$ 43.3

with interquartile ranges, and then relate those numerical summaries to figures that make the group differences visible.

An analysis of the `ChickWeights` dataset that recognises the effect of diet and time (start and end of experiment) might reveal something like this:

```
R> # A tibble: 8 x 10
R>   Diet   Time mean_wt sd_wt min_wt qrt1_wt med_wt qrt3_wt max_wt n_wt
R>   <fct> <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <int>
R> 1 1      0    41.4    1     39    41     41     42     43    20
R> 2 1     21   178.   58.7    96   138.   166   208.   305    16
R> 3 2      0    40.7    1.5    39   39.2   40.5    42     43    10
R> 4 2     21   215.   78.1    74   169   212.   262.   331    10
R> 5 3      0    40.8    1     39    41     41     41     42    10
R> 6 3     21   270.   71.6   147   229   281   317   373    10
R> 7 4      0     41    1.1    39   40.2    41     42     42    10
R> 8 4     21   239.   43.3   196   204   237   264   322     9
```

We typically report the measure of central tendency together with the associated variation. So, in a table we would want to include the mean $\pm$ SD. For example, this table is almost ready for including in a publication:

Further, we want to supplement this EDA with some figures that visually show the effects. Here I show a few options for displaying the effects in different ways: Figure 2.4 shows the spread of the raw data, the mean, median or as well as the appropriate accompanying indicators of variation around the mean or median. I will say much more about using figures in EDA in [Chapter 3](#).

2.8 REPORTING

We want to communicate our EDA in a report or publication. Here is how we would typically do it:

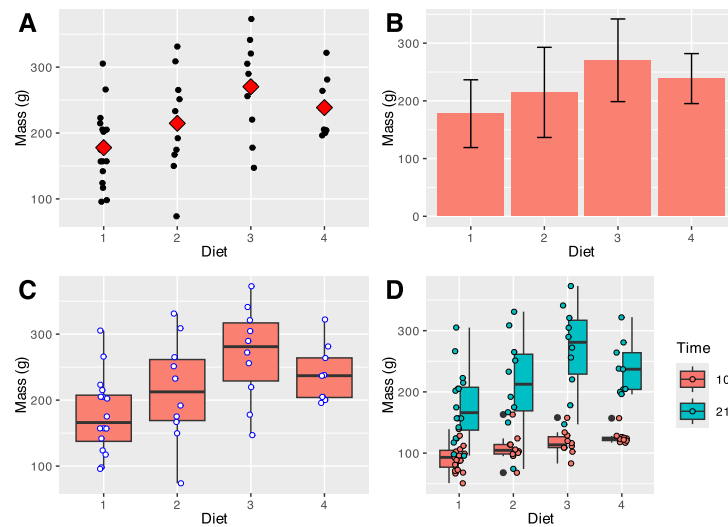


FIGURE 2.4. The figures represent A) a scatterplot of the mean and raw chicken mass values; B) a bar graph of the chicken mass values, showing whiskers indicating $1 \pm SD$; C) a box and whisker plot of the chicken mass data; and D) chicken mass as a function of both Diet and Time (10 and 21 days).

WRITE-UP**Methods**

Chicken body mass was summarised descriptively by diet group at the start of the experiment (Time = 0) and again at day 21. Group means, standard deviations, and sample sizes were calculated, and complementary figures were used to visualise the spread of raw observations within each diet.

Results

At the start of the experiment, mean body mass was similar across diets (about 41 g in all four groups). By day 21, mean mass had increased in every diet group, but the increase differed among diets: Diet 3 had the highest mean final mass (270.3 ± 71.6 g SD, $n = 10$), followed by Diet 4 (238.6 ± 43.3 g, $n = 9$), Diet 2 (214.7 ± 78.1 g, $n = 10$), and Diet 1 (177.8 ± 58.7 g, $n = 16$). The grouped summaries and figures therefore suggest a strong diet-related difference in final body mass, with Diet 3 producing the heaviest chicks in this descriptive comparison.

Discussion

This is still a descriptive result rather than a formal inferential test, but it shows what grouped summaries reveal. Once the data are split by diet and time, biologically important differences become visible that would be hidden in a single overall mean.

DO IT NOW!

Select your favourite three continuous variables in the BCB7342 field trip set of data. Assess the grouping structure, and apply a full set of descriptive analyses. Describe your data's distribution in the light of your finding, and provide visual support.

2.9 CONCLUSION

Numerical summaries are the starting point for any serious data analysis. Measures of centre tell you where a distribution sits; measures of spread tell you how uncertain that location is; quantiles and structure-inspection tools tell you whether the data meet the assumptions you will rely on in later tests. None of these summaries is informative in isolation. For example, a mean without a measure of spread is incomplete, and a spread without context for the sample size can also be misleading. Used together, and always computed within the grouping structures that the study design imposes, they give you a first account of what your data contain.

The numerical summaries covered here are, however, only one half of exploratory data analysis. Tables of means and standard deviations alone can hide the shape of a distribution, mask outliers, and obscure the relationships between variables. [Chapter 3](#) extends the picture by turning exploratory data analysis into graphical evidence. In that chapter, I will show you how to choose the right plot for each

type of question and how to communicate your findings in a form that a scientific audience can immediately interpret.

3

Visualising Data

Extending Exploratory Data Analysis with Figures

IN THIS CHAPTER

- The diversity of graphs used to communicate statistical results
- How to select the right graph for any particular dataset
- Additional packages available to extend ggplot's functionality

TASKS TO COMPLETE IN THIS CHAPTER

- Self-Assessment Task 3-1 (/30)
- [Self-Assessment instructions and full task overview](#)

Figures are not there to look pretty, although they can. In exploratory data analysis, figures are part of the analysis itself. A good figure lets you see shape, spread, clustering, outliers, and relationships that a table of summary statistics cannot show.

In the work below, I will extend the summaries in [Chapter 2](#) by turning them into graphical supporting evidence. The main task is to match the plot to the question. Some plots show the distribution of one variable, some compare groups, and others reveal relationships between variables.

The examples here will help you recognise which figure answers which descriptive question, rather than to exhaust every possible `ggplot2` option. For broader coverage and additional figure types, consult the following sources:

- [ggplot2](#) by Hadley Wickham et al.
- [ggplot2: Elegant Graphics for Data Analysis, v.3](#) by Hadley Wickham, et al.

- [R Graphics Cookbook, 2nd edition](#) by Winston Chang
- [A Comprehensive Guide on ggplot2 in R](#) by Devashree Madhugiri
- [ggplot2 extensions](#)
- [The R Graph Gallery](#) by Yan Holtz

3.1 KEY CONCEPTS

These decision steps guide the figure types you will encounter below:

- **Graphical EDA:** Plots are an essential part of the data exploration and assist in communicating our findings.
- **ggplot2:** A versatile R package for building statistical graphics.
- **Distribution Plots:** Use histograms and density plots to examine the shape of a single variable.
- **Group Comparison Plots:** Use box plots and violin plots to compare distributions across groups; use bar graphs mainly for counts or proportions.
- **Relationship Plots:** Use scatter plots and line graphs to show how variables change together.

3.2 CHOOSING THE RIGHT PLOT

Choose the plot from the question you are asking. So, if you want to:

- know how a single variable is distributed, start with a **histogram** or **density plot**.
- compare groups, use a **box plot** or **violin plot**. Use a **bar graph** only when the quantity of interest is a count or proportion, or when you deliberately want to show a summary such as a mean with error bars.
- examine the relationship between two variables, use a **scatter plot**. If one axis is ordered in time or sequence, a **line graph** may be better.

Summary statistics remain important, but they are not enough on their own because two datasets can share the same mean, standard deviation, and correlation while having completely different structure.

DO IT NOW!

Retrieve the `random_not.csv` and `random_not_either.csv` data files I made for this exercise and calculate the mean and standard deviation of the variables in each file. Then make scatter plots of the two datasets side by side and compare them. The contrast is obvious. What do you see? What do you conclude?

The point of the exercise is that a figure can expose structure that summary statistics conceal.

3.3 VISUALISING DISTRIBUTIONS

We begin with plots that describe the shape of a single variable. These plots answer questions about skewness, spread, outliers, and possible multi-modal distribution.

3.3.1 Histograms

Histograms display the distribution of a continuous variable by dividing the data into bins and counting how many observations fall within each bin. They should be the first plot to make when you want to inspect the shape of a distribution.

We have a choice of **absolute** frequency histograms, where the bar heights are counts, and **relative** frequency histograms, where the heights are proportions. Relative frequency is useful when comparing samples of different sizes. The **empirical cumulative distribution function** (ECDF) is another distribution plot that shows the cumulative proportion of observations below each value. The Old Faithful data illustrate these alternatives in Figure 3.1.

```
hist1 <- ggplot(data = faithful, aes(x = eruptions)) +
  geom_histogram(colour = "black", fill = "salmon", alpha = 0.6) +
  labs(title = "'Vanilla' Histogram",
       x = "Eruption duration (min)",
       y = "Count")

hist2 <- ggplot(data = faithful, aes(x = eruptions)) +
  geom_histogram(aes(y = ..density..),
                position = "identity", binwidth = 1,
                colour = "black", fill = "salmon", alpha = 0.6) +
  labs(title = "Relative Frequency",
       x = "Eruption duration (min)",
       y = "Relative\ncontribution")

hist3 <- ggplot(data = faithful, aes(x = waiting)) +
  geom_histogram(aes(y = 0.5 * ..density..),
                position = "identity", binwidth = 0.5,
                colour = "salmon", fill = "salmon", alpha = 0.6) +
  labs(title = "Relative Frequency",
       x = "Waiting time (min)",
       y = "Relative\ncontribution")

hist4 <- ggplot(data = faithful, aes(x = eruptions)) +
  stat_ecdf() +
  labs(title = "ECDF",
       x = "Eruption duration (min)",
       y = "Relative\ncontribution")

ggarrange(hist1, hist2, hist3, hist4, ncol = 2, nrow = 2, labels = "AUTO")
```

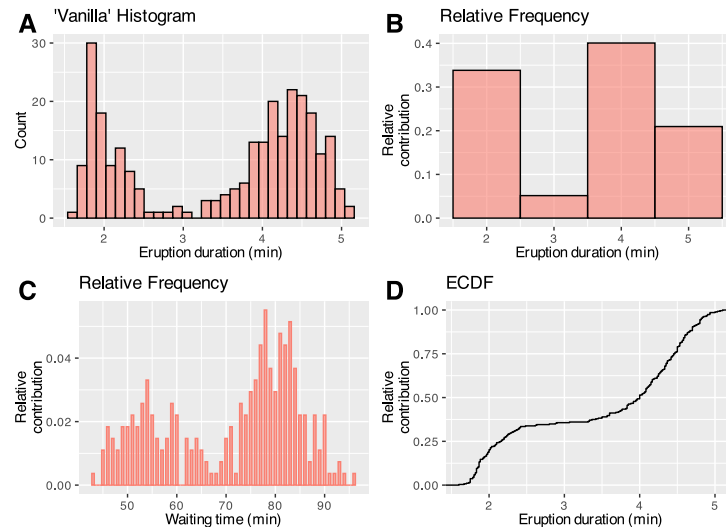


FIGURE 3.1. Example histograms for the Old Faithful data. A) A default frequency histogram. B) A relative frequency histogram with 1-minute bins. C) A relative frequency histogram with 0.5-minute bins. D) An ECDF.

`ggplot2` constructs histograms with `geom_histogram()`. We can also create bins manually with `cut()` if we need full control.

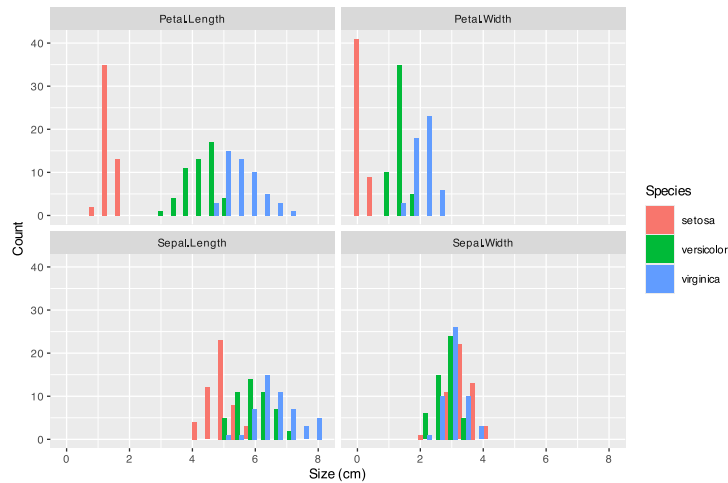
DO IT NOW!

Using with the `cut()` function, recreate Figure 3.1 A-C manually.

What if the same continuous variable is measured in several groups? The `iris` dataset gives a grouped example. These are size measurements for 50 flowers from each of three species of *Iris*. The grouped histograms in Figure 3.2 let us compare the distributions directly.

```
iris.long <- iris %>%
  gather(key = "variable", value = "size", ~Species)

ggplot(data = iris.long, aes(x = size)) +
  geom_histogram(position = "dodge",
    colour = NA, bins = 20,
    aes(fill = Species)) +
  facet_wrap(~variable) +
  labs(x = "Size (cm)",
    y = "Count")
```


FIGURE 3.2. Grouped histograms for the four *Iris* variables.

3.3.2 Density Plots

Density plots provide a smoothed view of a distribution. They emphasise shape rather than bin counts, which makes them useful when you want to compare overall form without committing too strongly to a particular bin width.

Compared with a histogram, a density plot is smoother and often easier to compare across groups. The trade-off is that it does not show counts as directly. Histograms answer “how many observations fall in each interval?” Density plots answer “what is the overall shape of the distribution?”

Figure 3.3 shows several density-plot variations:

```
dens1 <- ggplot(data = faithful, aes(x = eruptions)) +
  geom_density(colour = "black", fill = "salmon", alpha = 0.6) +
  labs(x = "Eruption duration (min)",
       y = "Density")

dens2 <- ggplot(data = faithful, aes(x = eruptions)) +
  geom_density(colour = "black", fill = "salmon", alpha = 0.6) +
  geom_rug(colour = "red") +
  labs(x = "Eruption duration (min)",
       y = "Density")

dens3 <- ggplot(data = faithful, aes(x = eruptions)) +
  geom_histogram(aes(y = ..density..),
                position = "identity", binwidth = 1,
                colour = "black", fill = "turquoise", alpha = 0.6) +
  geom_density(colour = "black", fill = "salmon", alpha = 0.6) +
  labs(x = "Eruption duration (min)",
       y = "Density")
```

```
dens4 <- ggplot(data = faithful, aes(x = eruptions)) +
  geom_histogram(aes(y = ..count..),
    binwidth = 0.2, colour = "black", fill = "turquoise", alpha = 0.6) +
  geom_density(aes(y = ..density.. * nrow(datasets::faithful) * 0.2),
    position = "identity",
    colour = "black", fill = "salmon", alpha = 0.6) +
  labs(x = "Eruption duration (min)",
    y = "Count")

ggarrange(dens1, dens2, dens3, dens4, ncol = 2, nrow = 2, labels = "AUTO")
```

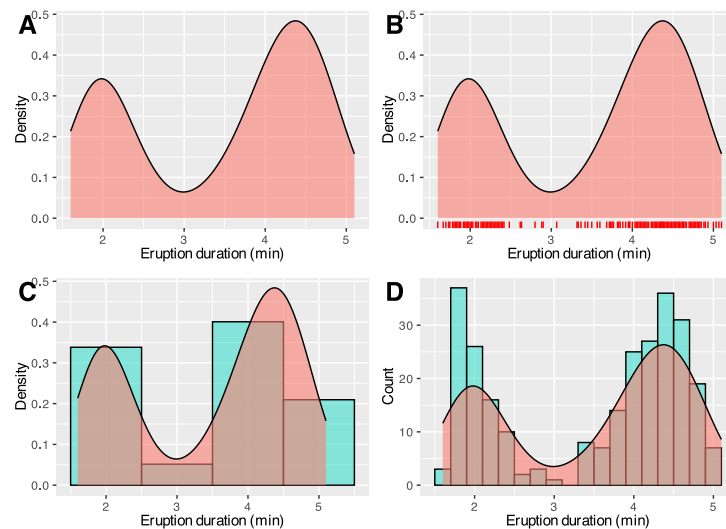


FIGURE 3.3. Density-plot variations based on the Old Faithful data. A) A basic density plot. B) A density plot with a rug. C) A histogram overlaid with a density curve. D) A count histogram with a rescaled density curve.

DO IT NOW!

Using the `faithful` dataset, produce two side-by-side plots: a histogram (choose a sensible bin width) and a density plot of eruptions. Look at the shapes. Do you see one peak or two? What does that tell you biologically? Now change the bin width in the histogram — does the apparent number of peaks change? Discuss with a partner whether the histogram or density plot is more revealing for this dataset and why.

3.4 COMPARING GROUPS

We now compare distributions across categories. The main choice here is whether you want to show the full distribution within each group or only a summary. In most biological applications, the distribution is revealing.

3.4.1 Box Plots

Box plots provide a compact summary of the distribution within each group. They show the median, quartiles, whiskers, and potential outliers, which makes them useful for comparing several groups at once.

Box plots are often more informative than mean-only bar graphs because they show spread and asymmetry directly. A useful variation is to add jittered raw observations on top of the boxes.

I create a simple example using the `msleep` dataset (Figure 3.4). Additional examples are provided in [Chapter 2](#).

```
msleep >
  ggplot(aes(x = vore, y = sleep_total)) +
    geom_boxplot(colour = "black", fill = "salmon", alpha = 0.4,
                 outlier.colour = "red3", outlier.fill = "red",
                 outlier.alpha = 1.0, outlier.size = 2.2) +
    geom_jitter(width = 0.10, fill = "blue", alpha = 0.5,
                col = "navy", shape = 21, size = 2.2) +
    labs(x = "'-vore'",
         y = "Sleep duration (hr)")
```

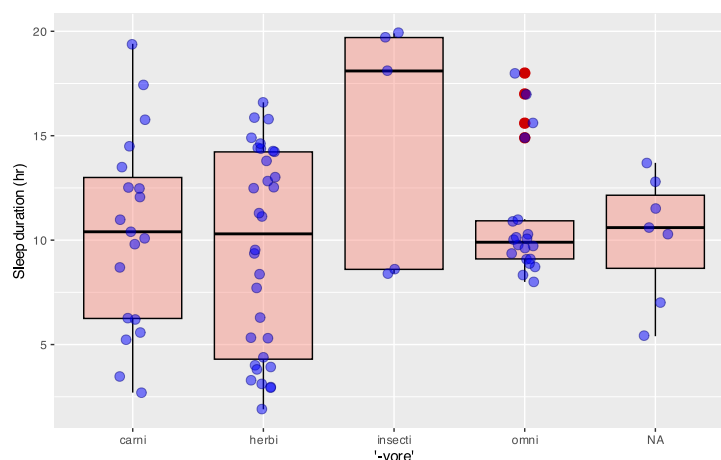


FIGURE 3.4. Box plot summarising the amount of sleep required by different 'vores'.

Box plots are sometimes called **box-and-whisker plots**. The `geom_boxplot()` documentation gives the clearest definition of the parts:

The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles).

The upper whisker extends from the hinge to the largest value no further than $1.5 \times \text{IQR}$ from the hinge (where IQR is the inter-quartile range, or distance between the first, third quartiles). The lower whisker extends from the hinge to the smallest value at most $1.5 \times \text{IQR}$ of the hinge. Data beyond the end of the whiskers are called 'outlying' points and are plotted individually.

In a notched box plot, the notches extend approximately $1.58 \times \text{IQR} / \sqrt{n}$. This gives a roughly 95% confidence interval for comparing medians.

Here is a second example, this time with notched box plots (Figure 3.5):

```
library(ggsci)

ggplot(data = iris.long, aes(x = Species, y = size)) +
  geom_boxplot(alpha = 0.4, notch = TRUE) +
  geom_jitter(width = 0.1, shape = 21, fill = NA,
              alpha = 0.4, aes(colour = as.factor(Species))) +
  facet_wrap(~variable, nrow = 2) +
  scale_color_npg() +
  labs(y = "Size (cm)") +
  guides(colour = FALSE) +
  theme(axis.text.x = element_text(face = "italic"))
```

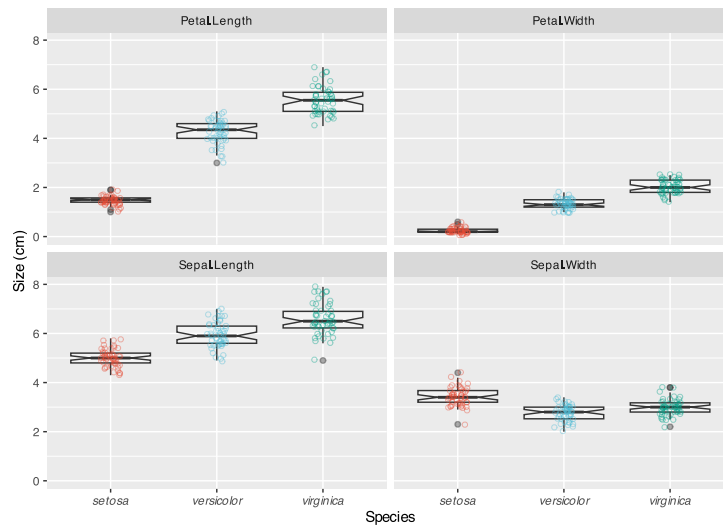


FIGURE 3.5. A panelled collection of box plots, one for each of the four variables, with a scatter plot to indicate the spread of the raw data points.

3.4.2 Violin Plots

Violin plots show the same grouped comparisons as box plots, but they display the full smoothed distribution within each group. That makes them useful when the shape of the data is interesting, especially if the distributions may be skewed or multimodal.

A box plot is better when you want a compact summary. A violin plot is better when you want to show the internal shape of the distribution. Here are the iris data shown with violin plots (Figure 3.6):

```
vio1 <- ggplot(data = iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_violin() +
  labs(title = "Basic violin plot", y = "Sepal length (cm)") +
  theme(axis.text.x = element_text(face = "italic"),
        legend.position = "none")

vio2 <- ggplot(data = iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_violin(show.legend = FALSE, draw_quantiles = c(0.25, 0.5, 0.75)) +
  labs(title = "With quartiles", y = "Sepal length (cm)") +
  theme(axis.text.x = element_text(face = "italic"),
        legend.position = "none")

vio3 <- ggplot(data = iris, aes(x = Species, y = Sepal.Length, colour = Species)) +
  geom_violin(fill = "grey70") +
  geom_boxplot(width = 0.1, colour = "grey30", fill = "white") +
  labs(title = "Nested box plots", y = "Sepal length (cm)") +
```

```

theme(axis.text.x = element_text(face = "italic"),
      legend.position = "none")

vio4 <- ggplot(data = iris, aes(x = Species, y = Sepal.Length, colour = Species)) +
  geom_violin(fill = "grey70") +
  geom_boxplot(width = 0.1, colour = "black", fill = "white") +
  geom_jitter(shape = 1, width = 0.1, colour = "red", alpha = 0.7, fill = NA) +
  labs(title = "Boxes and jittered data", y = "Sepal length (cm)") +
  theme(axis.text.x = element_text(face = "italic"),
        legend.position = "none")

ggarrange(vio1, vio2, vio3, vio4, ncol = 2, nrow = 2, labels = "AUTO")

```

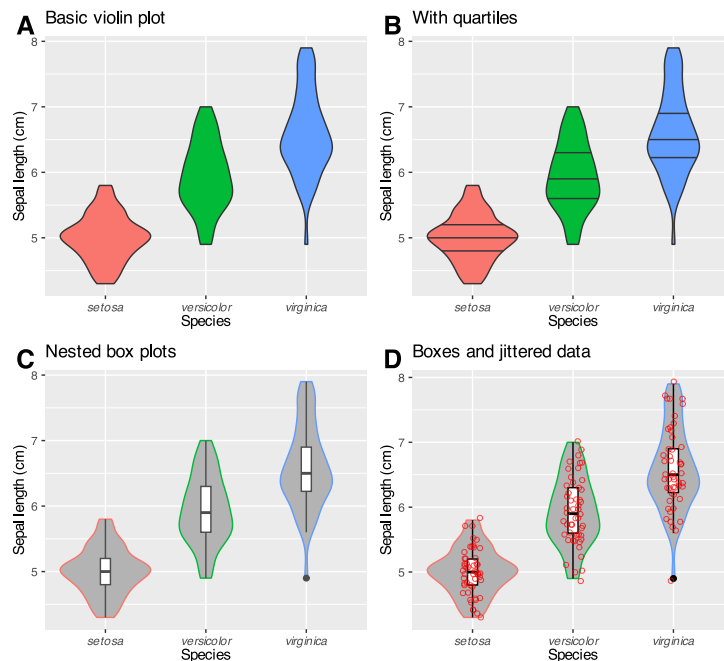


FIGURE 3.6. Variations of violin plots applied to the *Iris* data.

The **ggpubr** package also provides many convenience functions for publication-quality graphs, including violin plots.

DO IT NOW!

For each of the following scenarios, decide whether a box plot or a violin plot is more appropriate, and briefly justify your choice:

- You have body-mass data for three penguin species ($n \approx 150$ per species) and want to check whether the distributions are symmetric or skewed.
- You are comparing gill-net catch-per-unit-effort across five coastal sites, but $n = 8$ per site.
- A reviewer asks you to show the raw spread of pH measurements at two sites in a manuscript figure.

After deciding individually, compare your choices with a partner. Are there cases where both would be acceptable?

SELF-ASSESSMENT TASK 3-1

- Using a tidy workflow, assemble a summary table of the **palmerpenguins** dataset that has a similar appearance as that produced by `psych::describe(penguins)`. (/5)
 - For bonus marks (which will not count anything), apply a beautiful, creative styling to the table using the **kable** package. Try and make it as publication ready as possible. Refer to a few journal articles to see how to professionally typeset tables.
- Still using the **palmerpenguins** dataset, perform an exploratory data analysis to investigate the relationship between penguin species, their morphological traits (bill length and bill depth, flipper length, and body mass). Employ the tidyverse approaches learned earlier in the module to explore the data, account for the grouping structures present within the dataset. (/10)
- Provide visualisations (use Figure 4 as inspiration) and summary statistics to support your findings and elaborate on any observed patterns or trends. (/10)
- Ensure your presentation is professional and adhere to the standards required by scientific publications. State the major aims of your analysis and the patterns you seek. Using the combined findings from the EDA and the figures produced here, discuss the findings in a formal Results section. (/5)

3.4.3 Bar Graphs

Bar graphs need stricter handling than they usually get. Use them confidently for **counts** and **proportions** of categorical data. Use them more cautiously for **means with error bars**, because that display hides the shape of the underlying distribution. When the data are continuous and the distribution is important, box plots or violin plots are usually better.

The first use is counts or proportions in non-overlapping categories. Here the bar height is the quantity of interest. In the second use, a bar may represent a summary such as a mean with an error bar. That can be acceptable, but it should be a deliberate choice rather than the default.

In Figure 3.7 I demonstrate count and proportion bar graphs using the built-in `iris` data:

```
iris.cnt <- iris %>%
  count(Species) %>%
  mutate(prop = n / sum(n))

plt1 <- ggplot(data = iris.cnt, aes(x = "", y = n, fill = Species)) +
  geom_bar(width = 1, stat = "identity") +
  labs(title = "Stacked Bar Graph", subtitle = "Counts",
       x = NULL, y = "Count") +
  theme_pubclean() + scale_color_few() +
  scale_fill_few()

plt2 <- ggplot(data = iris.cnt, aes(x = "", y = prop, fill = Species)) +
  geom_bar(width = 1, stat = "identity") +
  scale_y_continuous(breaks = c(0.00, 0.33, 0.66, 1.00)) +
  labs(title = "Stacked Bar Graph", subtitle = "Proportions",
       x = NULL, y = "Proportion") +
  theme_pubclean() + scale_color_few() +
  scale_fill_few()

plt3 <- plt1 + coord_polar("y", start = 0) +
  labs(title = "Pie Chart", subtitle = "Same proportions, weaker comparison",
       x = NULL, y = NULL) +
  scale_fill_brewer(palette = "Blues")

plt4 <- ggplot(data = iris, aes(x = Species, fill = Species)) +
  geom_bar(show.legend = FALSE) +
  labs(title = "Side-by-Side Bars", subtitle = "Counts per species", y = "Count") +
  theme_pubclean() + scale_color_few() +
  scale_fill_few()

ggarrange(plt1, plt2, plt3, plt4, nrow = 2, ncol = 2,
          labels = "AUTO", common.legend = TRUE)
```

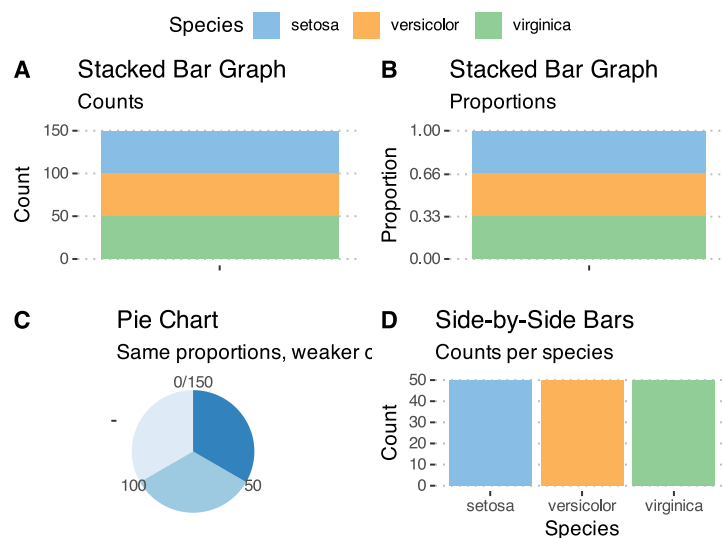



FIGURE 3.7. Bar graphs for categorical summaries in the built-in *Iris* data. A) A stacked count bar graph. B) A stacked proportion bar graph. C) A pie chart using the same proportions. D) A side-by-side count bar graph.

Here is the second use, where the bars represent means with standard deviations, as shown in Figure 3.8:

```
iris >
  pivot_longer(cols = Sepal.Length:Petal.Width,
               names_to = "variable",
               values_to = "size") >
  group_by(Species, variable) >
  summarise(mean = round(mean(size), 1),
            sd = round(sd(size), 1), .groups = "drop") >
  ggplot(aes(x = Species, y = mean)) +
  geom_bar(position = position_dodge(), stat = "identity",
           col = "black", fill = "salmon", alpha = 0.4) +
  geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd),
               width = .2) +
  facet_wrap(~variable,
            scales = "free") +
  ylab("Size (mm)") +
  theme_grey()
```

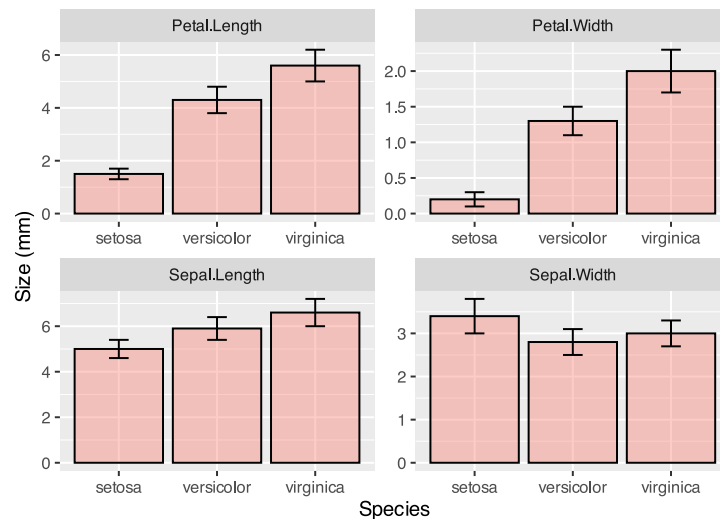


FIGURE 3.8. Bar graphs indicating the mean size ($\pm$ SD) for various flower features of three species of *Iris*.

The rule to keep in mind is to use bar graphs for counts or proportions, and to use box plots or violin plots when you need to show how continuous data are distributed within groups.

DO IT NOW!

Open a recent biology paper (your own reading list, or ask me for one). Find a bar graph that shows mean $\pm$ SD (or $\pm$ SE) for a continuous variable across treatment groups. Ask yourself: (1) does the bar graph reveal the distribution shape or just a summary? (2) What plot type would be more informative? Can you redraw the same comparison as a box plot or violin plot using the approximate data you can read from the original figure (you'll need n , the mean, and the SD)? If you can, would your interpretation of the results change?

3.5 VISUALISING RELATIONSHIPS

We now examine relationships between variables. Here the choice depends on whether the observations are scattered in two dimensions or ordered along time or some other continuous sequence.

3.5.1 Scatter Plots

Scatter plots are the standard choice for two continuous variables. Each point is one observation, and the cloud of points shows association, spread, outliers, clustering, and possible nonlinearity. Because the points are not connected, the plot does not imply continuity between neighbouring observations.

Scatter plots are especially useful for:

- identifying positive, negative, or absent trends;
- assessing whether the relationship is roughly linear or clearly nonlinear;
- revealing clusters or subgroup structure;
- checking for outliers that could influence later analyses.

All of these are illustrated in Figure 3.9. The example shows the relationship between two continuous variables in the *iris* data. The grouping structure is shown with colour, and the second panel adds fitted lines.

```
plt1 <- ggplot(data = iris, aes(x = Petal.Length, y = Petal.Width, colour = Species)) +
  geom_point() +
  labs(x = "Petal length (cm)", y = "Petal width (cm)") +
  scale_color_fivethirtyeight() +
  scale_fill_fivethirtyeight()

plt2 <- ggplot(data = iris, aes(x = Petal.Length, y = Petal.Width, colour = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, show.legend = FALSE) +
  scale_color_fivethirtyeight() +
  scale_fill_fivethirtyeight() +
  labs(x = "Petal length (cm)", y = "Petal width (cm)")

ggarrange(plt1, plt2, ncol = 2, nrow = 1, labels = "AUTO", legend = "right",
  common.legend = TRUE)
```

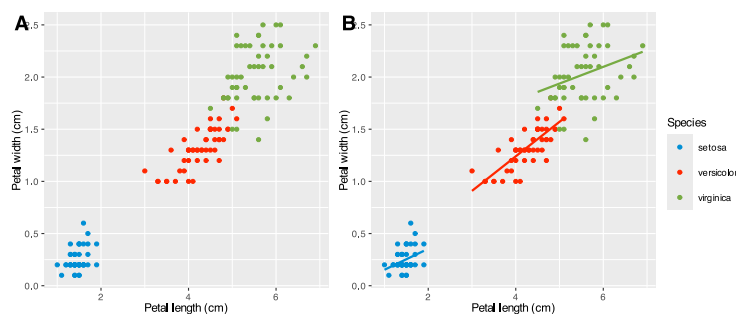


FIGURE 3.9. Examples of scatter plots made for the *Iris* data. A) A default scatter plot showing the relationship between petal length and width. B) The same data with fitted lines added.

These same features make scatter plots useful later as diagnostic tools for modelling.

DO IT NOW!

Using the built-in `msleep` dataset, make a scatter plot of `bodywt` (x-axis) against `brainwt` (y-axis). Colour the points by `vore`. Then add a separate linear trend line for each feeding group using `geom_smooth(method = "lm", se = FALSE)`.

Describe what you see: do the feeding groups appear to follow the same body-mass to brain-mass relationship, or do some groups differ? Are there any obvious outliers, and do the points cluster more strongly in some groups than others? Does adding grouped trend lines change your interpretation compared with looking at the scatter alone? You will revisit this kind of plot when we fit regression models in [Chapter 12](#).

3.5.2 Line Graphs

A line graph connects data points with lines and is most useful when the order of the observations must be considered. The usual case is time, but any ordered continuous scale can serve the same role. A line graph therefore implies continuity from one observation to the next in a way that a scatter plot does not.

We typically encounter line graphs in time-series displays. Points may be shown or omitted. Importantly, the line connects consecutive observations and makes trends, cycles, and seasonality easier to see. The monthly climatology in [Figure 3.10](#) provides a simple example.

```
library(lubridate)
read_csv(here::here("data", "BCB744", "SACTN_SAWS.csv")) >
  mutate(month = month(date)) >
  group_by(site, month) >
  dplyr::summarise(temp = mean(temp, na.rm = TRUE), .groups = "drop") >
  ggplot(aes(x = month, y = temp)) +
  geom_step(colour = "red4") +
  scale_x_continuous(breaks = c(1, 3, 5, 7, 9, 11)) +
  xlab("Month") + ylab("Temperature (°C)") +
  facet_wrap(~site, ncol = 5) +
  theme_grey()
```

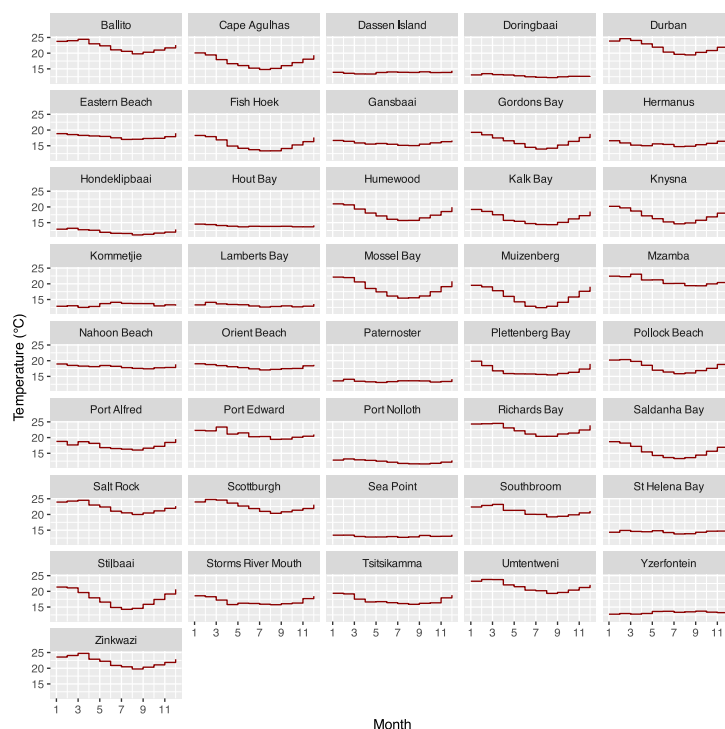


FIGURE 3.10. A time-series plot showing the monthly climatology for several sites around South Africa. The specific line drawn here forms a stepped graph.

3.6 EXTENSIONS

These plots extend the same principles to higher-dimensional data. They are still descriptive tools, but they are designed for cases where a simple one-dimensional distribution or two-variable relationship is no longer enough.

3.6.1 Heatmaps

Heatmaps encode values in a matrix as colour. They are useful when the structure of interest lies across two ordered dimensions at once, such as time by variable, site by species, or a correlation matrix. We will see another heatmap in Chapter 9 on [Correlation and association](#). A calendar heatmap is a specialised form that lays time out in calendar space rather than on a single axis. The Western Australian example

in Figure 3.11 shows how this reveals clustering in time. The Western Australian example in Figure 3.11 shows how this reveals clustering in time.

```
source("https://raw.githubusercontent.com/iascchen/VisHealth/master/R/calendarHeat.R")
```

```
temps <- heatwaveR::sst_WA >
  filter(t >= "2010-01-01" & t <= "2019-12-31") >
  mutate(weekday = lubridate::wday(t),
         weekday_f = lubridate::wday(t, label = TRUE),
         week = lubridate::week(t),
         month = lubridate::month(t, label = TRUE),
         year = lubridate::year(t)) >
  group_by(year, month) >
  mutate(monthweek = 1 + week - min(week))

ggplot(temps, aes(monthweek, weekday_f, fill = temp)) +
  geom_tile(colour = "white") +
  facet_grid(year(t) ~ month) +
  scale_x_continuous(breaks = c(1, 3, 5)) +
  scale_y_discrete(breaks = c("Sun", "Wed", "Sat")) +
  scale_fill_viridis_c() +
  xlab("Week of Month") +
  ylab("") +
  ggtitle("Time-Series Calendar Heatmap: Western Australia SST") +
  labs(fill = "[°C]")
```

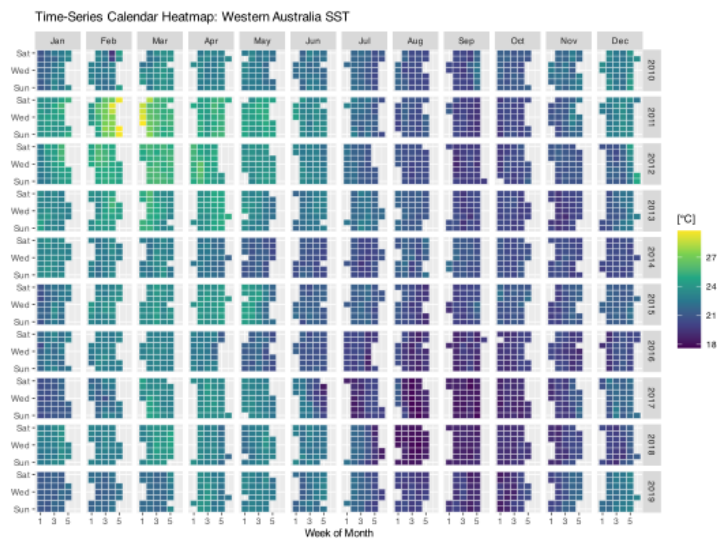


FIGURE 3.11. A calendar heatmap showing a time series of SST for Western Australia. The marine heatwave of summer 2011 is visible in the warm colours.

3.6.2 Hovmöller Diagrams

A Hovmöller diagram is a specialised heatmap used widely in oceanography and atmospheric science. One axis is usually time and the other is a spatial coordinate such as latitude or longitude. This makes it useful for tracking how spatial structure changes through time.

A **horizon plot** is a related way to compress long time-series displays while preserving anomaly structure. I provide more detail in the [vignette](#), where the method is used for extreme temperature events.

A Hovmoller diagram such as Figure 3.12 makes the latitude-by-time structure explicit.

```
library(data.table)
library(colorspace)

NWA <- fread(here::here("data", "BCB744", "NWA_Hovmoller.csv"))

NWA >
  mutate(anom = zonal_sst - mean(zonal_sst)) >
  ggplot(aes(x = t, y = lat, fill = anom)) +
  geom_tile(colour = "transparent") +
  scale_fill_binned_diverging(palette = "Blue-Red 3", n_inter = 21) +
  xlab("") + ylab("Latitude [°N]") + labs(fill = "[°C]") +
  theme_grey()
```

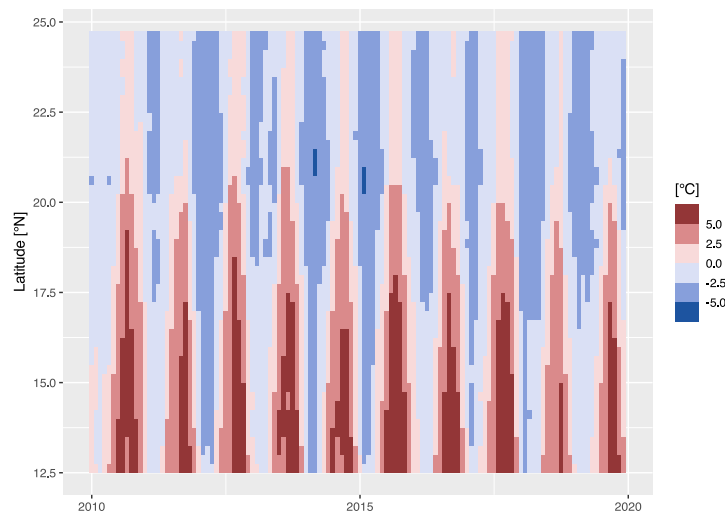


FIGURE 3.12. A Hovmöller diagram of zonally averaged SST for a region off North-west Africa in the Canary upwelling system.

```

library(ggHoriPlot)

cutpoints <- NWA %>%
  mutate(
    outlier = between(
      zonal_sst,
      quantile(zonal_sst, 0.25, na.rm = TRUE)-
        1.5*IQR(zonal_sst, na.rm = TRUE),
      quantile(zonal_sst, 0.75, na.rm = TRUE)+
        1.5*IQR(zonal_sst, na.rm=TRUE))) %>%
    filter(outlier)

ori <- round(sum(range(cutpoints$zonal_sst))/2, 2)

sca <- round(seq(range(cutpoints$zonal_sst)[1],
  range(cutpoints$zonal_sst)[2],
  length.out = 7)[-4], 2)

NWA %>% ggplot() +
  geom_hORIZON(aes(t,
    zonal_sst,
    fill = after_stat(Cutpoints)),
    origin = ori, horizonscale = sca) +
  scale_fill_hcl(palette = "RdBu", reverse = TRUE) +
  facet_grid(lat~.) +
  theme_few() +
  theme(
    panel.spacing.y = unit(0, "lines"),
    strip.text.y = element_text(size = 7, angle = 0, hjust = 0),
    axis.text.y = element_blank(),
    axis.title.y = element_blank(),
    axis.ticks.y = element_blank(),
    panel.border = element_blank()
  ) +
  scale_x_date(expand = c(0,0),
    date_breaks = "1 year",
    date_labels = "%Y") +
  xlab("Year") +
  ggtitle("Canary current system zonal SST")

```

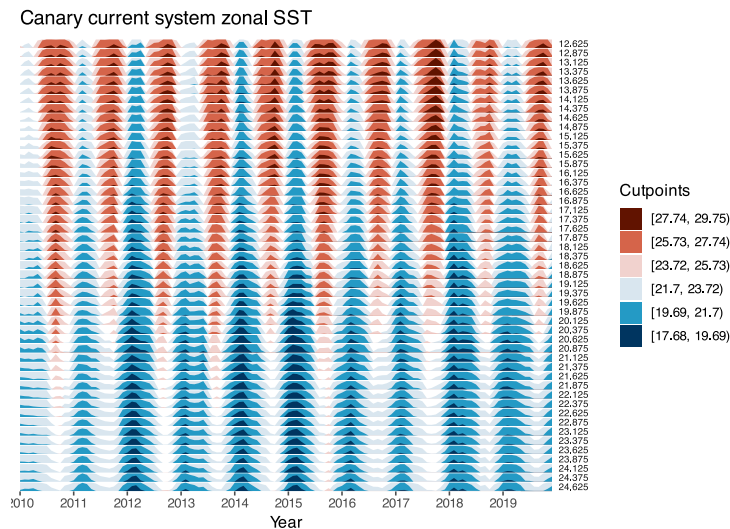



FIGURE 3.13. Zonally averaged time series of SST in the Canary Current system displayed as a horizon plot.

The same zonal SST structure can also be compressed into the horizon display in Figure 3.13.

3.6.3 Other Examples (To Develop)

(Incomplete.)

3.7 CONCLUSION

Figures extend statistical summaries by showing what the summaries hide. The main task is to match the plot to the question: distributions with histograms or density plots, group comparisons with box or violin plots, relationships with scatter plots or line graphs, and more complex structures with heatmaps or related displays.

The list in this chapter is not exhaustive, but it covers the main decisions you will make during EDA. For more examples and implementation detail, return to the resources listed in the introduction.

4

Distributions, Sampling, and Uncertainty

IN THIS CHAPTER

- What data distributions are and why they matter for statistical analysis
- How discrete and continuous distributions differ
- Common discrete distributions: Bernoulli, Binomial, Negative Binomial, Geometric, Poisson
- Common continuous distributions: Normal, Log-Normal, Gamma, Exponential, Beta
- Sampling variation and the central limit theorem
- Standard errors and confidence intervals as measures of uncertainty
- How to identify the distribution that best describes your data

TASKS TO COMPLETE IN THIS CHAPTER

- None

SOFTWARE USED IN THIS CHAPTER

- All distribution generation and visualisation in this chapter use `library(TidyDensity)` and `library(tidyverse)`. Install TidyDensity with `install.packages("TidyDensity")` if needed.

A good grasp of data distributions is a prerequisite for any statistical analysis. Before you can choose an appropriate test, fit a model, or interpret inferential output, you need to understand what kind of variation your data exhibit and what kind of

stochastic process most plausibly generated them. Distributions encode biological reality, their shape (seen in a figure) reflects the mechanism behind the data, and therefore choosing the wrong distributional model leads to invalid inference even when every other step of the analysis is correct.

The material in this chapter is the bridge between data description ([Chapter 2](#) and [Chapter 3](#)) and formal inference ([Chapter 5](#)). I cover three interconnected ideas:

1. the **distribution of observed data** and how biological processes shape that distribution;
2. the **distribution of estimates across repeated samples** (sampling distributions); and
3. the **uncertainty** attached to those estimates, expressed as standard errors and confidence intervals.

4.1 WHAT ARE DATA DISTRIBUTIONS?

A data distribution is a mathematical function that describes how probable each possible value of a variable is. When we collect measurements from any biological system (such as the heights of individuals, counts of organisms in quadrats, the proportion of habitat patches occupied by a species, for example) those values spread across a range and concentrate in particular regions. The pattern of that spread is the distribution, and the type of plot that reveals the shape is called a histogram.

The key questions are “what shape does my histogram have?” and “what kind of process most plausibly generated these values?” A count of animals can only be zero or a positive integer, so distributions that allow negative or fractional values are inappropriate. A proportion is bounded between zero and one. A waiting time cannot be negative. So, the data-generating process constrains the distribution, and recognising that constraint guides model choice.

Distributions fall into two broad categories based on the nature of the variable:

- **Discrete distributions** apply when the variable can only take countable values; they are typically non-negative integers. Examples include counts of events, individuals, or successes are discrete.
- **Continuous distributions** apply when the variable can take any value within an interval. Examples are measurements of length, mass, concentration, or time are continuous.

4.2 DISCRETE DISTRIBUTIONS

4.2.1 Mathematical Foundations

Discrete random variables have a finite or countable set of possible values. Two constraints govern any valid discrete probability distribution:

- Each probability $P(X = k)$ must lie in the interval $[0, 1]$: no outcome can have negative probability or a probability exceeding certainty.
- The probabilities across all possible outcomes must sum to exactly 1: $\sum_k P(X = k) = 1$.

The **expected value** (mean) of a discrete random variable is the probability-weighted average of all possible outcomes:

$$E(X) = \sum_k k \cdot P(X = k)$$

For a fair six-sided die, each face has probability $\frac{1}{6}$, so $E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = 3.5$. The die will never land on 3.5, but that is the long-run average over many throws.

4.2.2 The Bernoulli Distribution

MATHEMATICAL DETAIL

The Bernoulli distribution:

$$P(X = k) = \begin{cases} p, & k = 1 \text{ (success)} \\ 1 - p, & k = 0 \text{ (failure)} \end{cases} \quad (4.1)$$

p is the probability of success; $1 - p$ is the probability of failure.

Mean: $\mu = p$

Variance: $\sigma^2 = p(1 - p)$

The Bernoulli distribution is the simplest discrete distribution. It models a single binary trial with exactly two possible outcomes: success ($X = 1$, with probability p) and failure ($X = 0$, with probability $1 - p$). Every more complex trial-based distribution is built on this foundation.

Biological context. Any single binary observation follows a Bernoulli distribution when there is a fixed, constant probability of success:

- *Is this particular seed viable?* (germination: yes or no)
- *Is this study site occupied by the focal species?* (occupancy: yes or no)
- *Did this individual survive to the next census?* (survival: yes or no)

- *Does this offspring inherit the dominant allele?* (genetic outcome: yes or no)
- *Did the predator successfully capture its prey in this hunt?* (predation: yes or no)

The Bernoulli distribution itself describes a single event. In practice we nearly always record multiple events and care about the aggregate, which brings us to the Binomial distribution.

4.2.3 The Binomial Distribution

MATHEMATICAL DETAIL

The Binomial distribution:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (4.2)$$

n = number of independent trials; k = number of successes; p = probability of success per trial; $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient (the number of ways to arrange k successes in n trials).

Mean: $\mu = np$

Variance: $\sigma^2 = np(1 - p)$

Example: Planting 50 seeds with germination probability $p = 0.6$. The probability that exactly 35 germinate: $P(X = 35) = \binom{50}{35} (0.6)^{35} (0.4)^{15}$.

The Binomial distribution extends the Bernoulli to n independent trials, each with the same probability of success p . It tells us the probability of observing exactly k successes in those n trials. The important assumptions are independence between trials and a constant success probability.

Biological context. The Binomial distribution arises whenever we count successes in a fixed number of binary trials:

- **Seed germination:** Plant 100 seeds of a species in a greenhouse experiment. Each seed either germinates or does not, with probability p . The number that germinate follows a Binomial distribution. This lets us test whether germination rates differ between seed sources or treatments.
- **Disease prevalence:** A field epidemiologist collects a random sample of 500 individuals from a wildlife population. Each individual is either infected or not. The total number of infected individuals in the sample is Binomially distributed and enables estimation of population-level prevalence.
- **Species occupancy:** During a biodiversity assessment, 40 forest landscape patches are surveyed for the presence of a particular tree frog. If the species

occupies each patch independently with probability p , the count of occupied patches is Binomially distributed.

- **Mendelian genetics:** In a monohybrid cross, each offspring inherits either the dominant or recessive allele with probabilities following Mendel's ratios. The number of offspring showing the dominant phenotype in a litter of n is Binomially distributed.

Figure 4.1 illustrates the Binomial distribution for $p = 0.75$ across 100 trials.

```
binom <- tidy_binomial(.n = 40, .num_sims = 10, .prob = 0.75, .size = 100)

pqq <- ggplot(binom, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(binom, .plot_type = "density") + theme_grey(base_size = 8),
  pqq,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

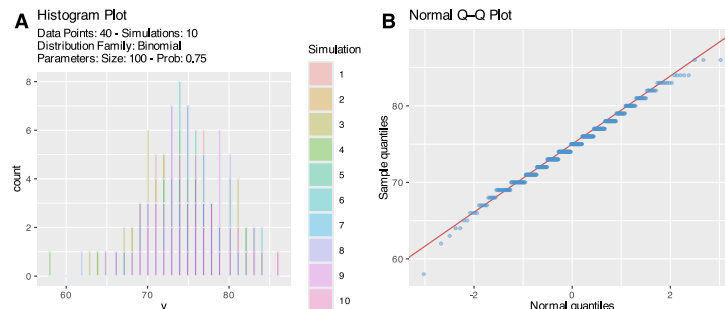


FIGURE 4.1. Binomial distribution with 40 observations per simulation, 10 simulations, 100 trials, and $p = 0.75$. This could represent 10 separate germination experiments, each planting 100 seeds with a 75% germination probability. The Normal Q-Q plot reveals the characteristic staircase pattern of a discrete distribution and the mild right-skew: points curve away from the reference line at both ends.

Figure 4.1 A shows a near-symmetric, bell-shaped density centred around 75 successes out of 100 trials, consistent with the expected value $np = 100 \times 0.75 = 75$. The spread is modest, reflecting the relatively low variance $np(1 - p) = 18.75$.

Figure 4.1 B reveals the characteristic staircase pattern of a discrete distribution: instead of a smooth curve of points, they fall in horizontal bands corresponding to each integer value. Despite this discreteness, the points track the reference line

closely across the bulk of the distribution, indicating that this near-symmetric Binomial is well approximated by a Normal. Minor departures at the tails reflect the finite support of the Binomial.

HOW TO READ THE TWO DIAGNOSTIC PANELS

Each distribution figure in this chapter shows two panels side by side: a **density plot** on the left and a **Normal Q–Q plot** on the right. Together they reveal the shape of the distribution and how far it departs from normality.

Panel A — Density Plot

This panel shows how values are spread along the x-axis.

- The **x-axis** is the variable (*e.g.*, counts, waiting times, or proportions).
- The **y-axis** is the probability density — how often values near a given point occur relative to others.
- A single tall peak indicates that most values cluster around a central value.
- A long tail to the right (right-skewed) or left (left-skewed) indicates that extreme values are more common in that direction.
- Multiple lines correspond to independent simulations; if they overlap closely the distribution is stable across runs.

Panel B — Normal Q–Q Plot

A **Quantile–Quantile plot** (Q–Q plot) compares the *quantiles* of the simulated data against the quantiles of a theoretical Normal distribution. It answers the question: *how closely does this distribution resemble a Normal?*

- The **x-axis** shows the theoretical quantiles you would expect if the data were perfectly Normal.
- The **y-axis** shows the actual quantiles observed in the simulated data.
- If the data are approximately Normal, the points will lie along the diagonal reference line (shown in red).
- Systematic departures from the line indicate non-normality:
 - Points curving **upward at the right tail** indicate **right skewness** (more large values than a Normal would have).
 - Points curving **downward at the left tail** indicate **left skewness**.
 - Points diverging at **both tails** indicate **heavy tails** (more extreme values than a Normal).
 - A nearly straight line indicates approximate normality, even if the underlying distribution is not Normal by construction.

The Normal distribution is always the **reference** in these Q–Q plots. A distribution that plots as a straight line here can be treated as approximately Normal for the purposes of parametric inference.

In Figure 4.1 A (density plot), we see a discrete distribution with a clear peak near the expected number of successes. Most values cluster tightly around this peak, with a slight extension toward higher values. This reflects repeated trials with a high success probability: most experiments produce similar counts, with moderate variability.

Figure 4.1 B (Normal Q–Q plot) shows a stepped pattern rather than a smooth line. This arises because the data are discrete counts rather than continuous values. The mild curvature away from the reference line indicates slight skewness. As the number of trials increases, the points move closer to the line, reflecting that the Binomial becomes approximately Normal for large sample sizes.

So, the distribution is governed by repeated independent trials with fixed probability, producing clustered counts and approximate normality only at larger scales.

Note that both the Bernoulli and Binomial distributions assume independence between trials and a fixed, constant success probability, but in real ecological data these assumptions are not always fully met. I will revisit this in Chapter 20 when we discuss quasi-binomial models for overdispersed proportions.

DO IT NOW!

A marine biologist surveys 30 rock pools. In each pool, she records whether a sea urchin (*Parechinus angulosus*) is present (1) or absent (0). Suppose the true probability of presence is $p = 0.4$.

- What is the expected number of occupied pools? Calculate this by hand: $E(X) = np$.
- Simulate this scenario in R with `rbinom(1000, size = 30, prob = 0.4)`. Plot a histogram of the 1000 outcomes. Where does the histogram peak? Does it match your calculation?
- Now change `size` to 5 and then to 100. How does the shape of the distribution change?

4.2.4 The Poisson Distribution

MATHEMATICAL DETAIL

The Poisson distribution:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (4.3)$$

$\lambda > 0$ is the mean rate of occurrence (events per unit time, area, or effort); $k = 0, 1, 2, \dots$; $e \approx 2.718$ is Euler's number; $k!$ is the factorial of k .

Mean: $\mu = \lambda$

Variance: $\sigma^2 = \lambda$

A defining feature of the Poisson: mean equals variance. When $\text{Var}(X) \gg \lambda$ in your data, the Poisson model is inadequate, use the Negative Binomial instead.

The Poisson distribution models the number of events or counts occurring in a fixed interval of time, area, or effort, given that:

1. events or counts occur independently of one another,
2. two events/counts cannot occur simultaneously,
3. the average rate λ is constant over the interval.

Biological context. The Poisson distribution arises from processes where events/counts are rare relative to the number of “opportunities” and are spatially or temporally independent:

- **Quadrat counts of randomly distributed organisms:** If individuals are scattered uniformly and randomly across a landscape (no aggregation, no repulsion), the count per quadrat follows a Poisson distribution. In practice, organisms are rarely perfectly random (some level of aggregation is more common), but Poisson remains the baseline model for count data.
- **Mutation rates:** The number of de novo mutations per genome per generation is approximately Poisson when mutations occur independently and at a constant low rate per base pair.
- **Rare events in time:** The number of disease outbreaks per year in a region, the number of storms exceeding a given intensity per decade, or the number of vertebrate extinctions per century are often Poisson-modelled when events are rare and independent.
- **Radioactive decay or photon counts:** In experimental biology, the number of photons detected per time interval in a fluorescence assay follows a Poisson distribution.

Figure 4.2 shows the Poisson distribution for $\lambda = 2$, a plausible mean density for a moderately common invertebrate species per 1 m² quadrat.

```
pois <- tidy_poisson(.n = 50, .lambda = 2, .num_sims = 10)

pqq_pois <- ggplot(pois, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(pois, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_pois,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

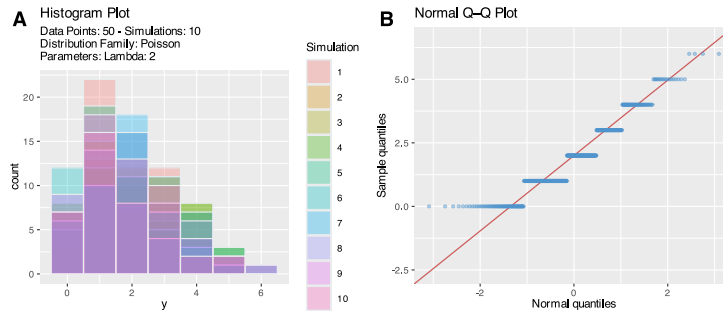


FIGURE 4.2. Poisson distribution with 50 observations per simulation, 10 simulations, and $\lambda = 2$. This could represent the expected number of a particular invertebrate species per 1 m² quadrat in a habitat where the mean density is 2 individuals per quadrat.

Figure 4.2 A shows a strongly right-skewed distribution. Most observations are small counts (including zeros), with a rapid decline in frequency as counts increase. This reflects a process where events occur independently at a constant rate: low counts are common, high counts are rare.

In Figure 4.2 B, we see clear curvature away from the Normal reference line, especially in the upper tail. This indicates that the distribution has more small values and fewer large values than a Normal would predict. The pattern reflects random events occurring in space or time at a constant rate; the strong skewness signals that a Normal model is inappropriate for such data.

4.2.5 The Negative Binomial Distribution

MATHEMATICAL DETAIL

The Negative Binomial distribution:

$$P(X = k) = \binom{k+r-1}{k} p^r (1-p)^k \quad (4.4)$$

r = number of successes before the experiment stops; k = number of failures; p = probability of success per trial. As a count model, r is a dispersion parameter controlling how much the variance exceeds the mean.

Mean: $\mu = \frac{r(1-p)}{p}$

Variance: $\sigma^2 = \frac{r(1-p)}{p^2}$

Since $\sigma^2 > \mu$ always, the Negative Binomial accommodates **overdispersion**, which is the single most common feature of ecological count data.

In ecology, the Negative Binomial is best understood as a Poisson distribution whose rate parameter λ itself varies randomly across observations. When some individuals or patches are intrinsically more likely to accumulate events than others (because of unmeasured heterogeneity), counts become more variable than Poisson. This is called **overdispersion**.

Biological context. Overdispersed counts are the rule, not the exception, in field biology:

- **Parasite loads:** Most host individuals carry few or no parasites, while a small number carry very large parasite burdens. This pattern of “aggregation” (*i.e.*, the excess of zeros and a long right tail) is characteristic of the Negative Binomial. The vast majority of helminths, ectoparasites, and pathogens follow this distribution in their hosts.
- **Plant seed bank counts:** Seeds in the soil are clumped near parent plants and in microhabitats that favour retention. Counts per soil core are thus overdispersed relative to Poisson.
- **Insect counts per plant:** Many insect species aggregate on certain host plants due to patch quality, host chemistry, or conspecific attraction. Counts per plant show far more variation than a Poisson model would predict.
- **Gene expression counts in RNA-seq:** Read counts per gene per sample exhibit overdispersion due to biological variability between individuals and makes the Negative Binomial the standard model for RNA-seq differential expression analysis.

Figure 4.3 illustrates the characteristic long right tail of the Negative Binomial, reflecting the typical pattern of overdispersed ecological count data.

```
negbinom <- tidy_negative_binomial(.n = 50, .size = 5, .prob = 0.7, .num_sims = 10)

pqq_nb <- ggplot(negbinom, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(negbinom, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_nb,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

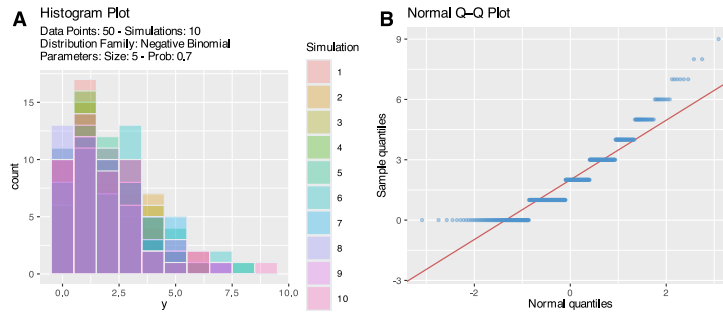


FIGURE 4.3. Negative Binomial distribution with 50 observations per simulation, 10 simulations, size parameter $r = 5$, and $p = 0.7$. The long right tail reflects the typical pattern of overdispersed ecological count data.

Figure 4.3 A shows a right-skewed distribution with a longer tail than the Poisson. Many observations are small, but there are more large values than expected under a Poisson process and this indicates aggregation. So, some observations accumulate far more counts than others.

Figure 4.3 B shows obvious deviation from the Normal reference line, especially in the upper tail. The spread of points indicates substantial variability beyond what a Normal model would capture and this is consistent with heterogeneous processes where some locations or individuals have higher underlying rates. This is called overdispersion, and it is a property of ecological data.

4.2.6 The Geometric Distribution

MATHEMATICAL DETAIL

The Geometric distribution:

$$P(X = k) = p(1 - p)^k \quad (4.5)$$

k = number of failures before the first success; p = probability of success per trial.

Mean: $\mu = \frac{1-p}{p}$

Variance: $\sigma^2 = \frac{1-p}{p^2}$

The Geometric distribution is a special case of the Negative Binomial with $r = 1$: the number of failures before achieving a *single* success.

The Geometric distribution models the number of failures that occur before the first success in a sequence of independent Bernoulli trials. It is a “waiting time” distribution in the discrete setting.

Biological context. The Geometric distribution applies whenever we want to know how many unsuccessful trials we will need to endure before the first success:

- **Rare species detection:** An ecologist surveys forest patches one at a time, searching for evidence of a rare mammal. If the probability of detecting the species in any single patch is $p = 0.15$, the number of patches visited before the first detection follows a Geometric distribution.
- **Pathogen screening:** A scientist tests animals from a herd for a particular infection. If prevalence is 10%, the number of negative tests before the first positive is Geometrically distributed. This information can inform decisions about how many animals need screening.
- **Breeding success:** If a bird pair has probability p of successfully fledging at least one chick per breeding attempt, the number of failed attempts before their first success follows a Geometric distribution.

Figure 4.4 shows the Geometric distribution for $p = 0.7$.

```
geom <- tidy_geometric(.n = 50, .prob = 0.7, .num_sims = 10)

pqq_geom <- ggplot(geom, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(geom, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_geom,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

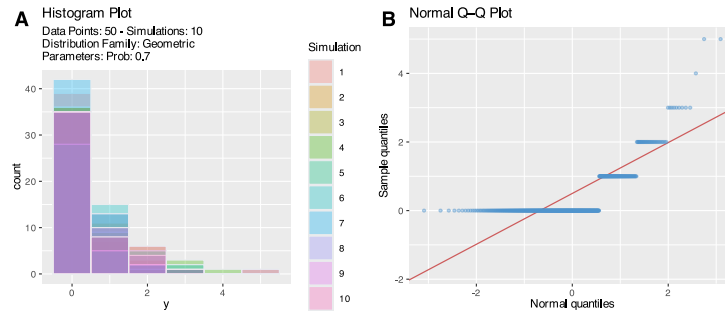


FIGURE 4.4. Geometric distribution with 50 observations per simulation, 10 simulations, and $p = 0.7$. Most observations cluster near zero (few failures before success), but the long tail captures the occasional run of bad luck.

We see in Figure 4.4 A a sharp concentration of values near zero, but with frequencies dropping quickly as values increase. This indicates that success typically occurs after few failures, but occasional long sequences of failures still occur.

Figure 4.4 B shows strong curvature away from the Normal reference line, which confirms that the distribution is highly skewed with a long right tail. Success often occurs quickly, but infrequent long waits are possible. So, these long waits pull the upper tail far beyond what any Normal model would predict.

4.3 CONTINUOUS DISTRIBUTIONS

Continuous random variables can take any value within a range, and their distributions are described by **probability density functions (PDFs)** rather than probability mass functions. For a continuous variable, $P(X = x) = 0$ for any single point x ; instead, we work with probabilities over intervals: $P(a \leq X \leq b) = \int_a^b f(x) dx$.

4.3.1 The Normal Distribution

MATHEMATICAL DETAIL

The Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (4.6)$$

μ = mean (location); σ = standard deviation (scale); σ^2 = variance.

Support: $(-\infty, +\infty)$

Symmetric around μ ; tails thin out rapidly.

Rule of thumb: approximately 68%, 95%, and 99.7% of observations fall within 1, 2, and 3 standard deviations of the mean, respectively.

NOTE

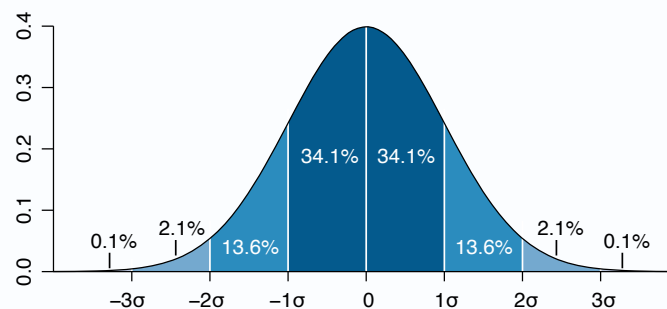


FIGURE 4.5. The idealised Normal distribution showing the proportion of data within 1, 2, and 3 SDs from the mean.

The Normal distribution is the most widely used distribution (and the most desired!) in statistics. The **central limit theorem** (see below) guarantees that averages of large samples tend toward normality regardless of the underlying distribution. It is characterised by its bell shape, perfect symmetry around the mean, and the fact that the mean, median, and mode all coincide (see Figure 4.5 for the idealised proportions).

Biological context. The Normal distribution arises from processes where many small, independent factors contribute additively to the outcome:

- **Morphological traits:** Body length, leaf area, wing span, and shell diameter in natural populations are often approximately normally distributed. Each individual's phenotype is the product of many small genetic and environmental contributions adding together. By the central limit theorem, this additive structure produces approximate normality.
- **Measurement error:** Repeated measurements of the same object differ due to instrument noise and small random errors. These differences are approximately normally distributed around the true value.

- **Physiological variables:** Resting heart rate, core body temperature, and blood oxygen levels in a population of healthy individuals tend to cluster symmetrically around a species-typical mean.
- **Residuals in linear models:** Even when the raw data are not normal, the residuals from a well-specified linear model often are, satisfying a key assumption for inference.

Figure 4.6 shows five simulations of the standard normal ($\mu = 0, \sigma = 1$).

```
norm <- tidy_normal(.n = 100, .num_sims = 5)
pqq_norm <- ggplot(norm, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)
ggarrange(
  tidy_autoplot(norm, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_norm,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

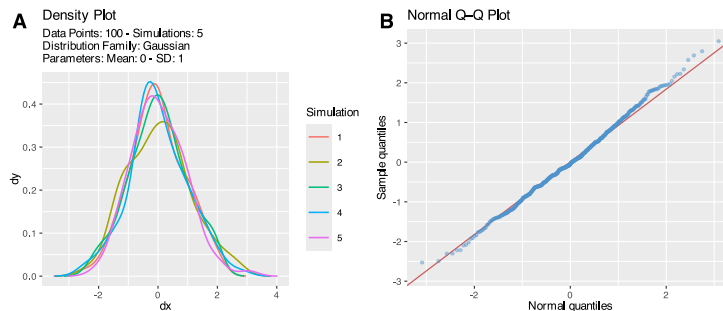


FIGURE 4.6. Normal distribution with 100 observations per simulation and 5 simulations ($\mu = 0, \sigma = 1$, i.e. the standard normal). In practice, biological variables will have different means and standard deviations, but the symmetric bell shape will be the same. The Q–Q plot points follow the reference line closely, confirming normality.

In Figure 4.6 A we see a symmetric, bell-shaped distribution centred around zero. Values are evenly distributed on both sides, with fewer observations in the tails. Multiple simulations overlap closely, which indicates a stable, well-behaved distribution.

Figure 4.6 B shows points lying almost exactly along the reference line. This is what normality looks like. Points track the diagonal smoothly from lower-left to

upper-right, with only minor scatter at the extremes. This is the reference pattern against which all other Q–Q plots in this chapter should be compared.

THE CENTRAL LIMIT THEOREM

The **central limit theorem (CLT)** states that the sampling distribution of the mean of n independent, identically distributed observations approaches a Normal distribution as $n \rightarrow \infty$, regardless of the shape of the underlying population distribution. Three conditions must hold:

- Observations must be **independent**: one observation must not influence another.
- Observations must come from the **same distribution** with the same finite mean and variance.
- Sample size must be **sufficiently large**: a common rule of thumb is $n \geq 30$, though heavily skewed or heavy-tailed populations require larger samples.

The CLT is why Normal-based tests and confidence intervals work reasonably well in many practical situations, because we are relying on the behaviour of sample means, not the raw data. It also explains why checking normality of *residuals* matters more than checking normality of the raw response variable in most regression contexts.

DO IT NOW!

Generate 200 observations from a Normal distribution with mean 50 and SD 10 (`rnorm(200, mean = 50, sd = 10)`). Then:

- Compute the mean and median. Are they close? They should be — why?
- Compute the proportion of observations that fall within 1, 2, and 3 SDs of the mean. Compare to the 68–95–99.7 rule.
- Make a Q–Q plot (`qqnorm()`; `qqline()`). Do the points follow the reference line?

4.3.2 The Log-Normal Distribution

MATHEMATICAL DETAIL

The Log-Normal distribution:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \quad (4.7)$$

μ = mean of $\ln(X)$; σ = SD of $\ln(X)$.

Support: $(0, +\infty)$

Right-skewed for most parameter values.

If X is log-normally distributed, then $Y = \ln(X)$ is normally distributed.

If a variable is the product of many small, independent, positive multiplicative factors rather than additive ones, its distribution will be log-normal rather than normal. The log-normal is always positive, always right-skewed, and produces the characteristic pattern of many small values with a long right tail.

Biological context. Multiplicative processes occur very frequently in biology and ecology:

- **Body mass and biomass:** Growth over time is fundamentally multiplicative (proportional growth rates), producing right-skewed size distributions. Log-transformation typically normalises mammal body masses across species.
- **Chlorophyll-*a* concentration and phytoplankton biomass:** Nutrient availability, light, temperature, and grazing all interact multiplicatively, producing log-normally distributed algal biomass in lakes and coastal waters. This is why water quality monitoring data are routinely log-transformed.
- **Species abundance distributions:** In most communities, many species are rare (small populations) and a few are very common. Across a community, species abundance values are approximately log-normally distributed.
- **Latency times for diseases:** The incubation periods of many infectious diseases, which depend on multiplicative amplification of pathogens within a host, are log-normally distributed.

Figure 4.7 shows the characteristic right-skewed shape of the Log-Normal.

```
lnorm <- tidy_lognormal(.n = 100, .meanlog = 0, .sdlog = 0.5, .num_sims = 5)

pqq_ln <- ggplot(lnorm, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(lnorm, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_ln,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

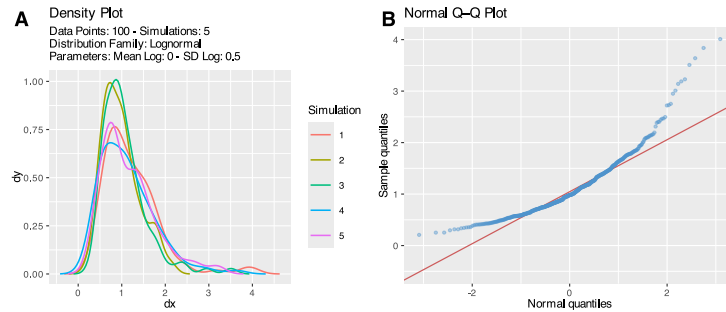


FIGURE 4.7. Log-Normal distribution with 100 observations per simulation and 5 simulations ($\mu = 0$, $\sigma = 0.5$). The strong right skew in the density plot is echoed by the upward curve of Q-Q points away from the reference line in the upper tail. On a log scale, this distribution becomes Normal.

In Figure 4.7 A, we see a strongly right-skewed distribution. There are many small values cluster near zero while a long tail of large values extends to the right. This is typical of multiplicative biological processes such as body mass or chlorophyll-*a* concentration.

In the Normal-QQ plot, seen in Figure 4.7 B, there is a clear upward curvature in the upper tail. Large observed values far exceed what a Normal distribution would predict. The left tail falls below the reference line to produce the characteristic S-shape of a positively skewed distribution. Log-transforming the data would collapse this curvature and bring the points onto the reference line.

4.3.3 The Gamma Distribution

MATHEMATICAL DETAIL

The Gamma distribution:

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad (4.8)$$

$\alpha > 0$ = shape parameter; $\beta > 0$ = scale parameter; $\Gamma(\alpha)$ is the gamma function (a generalisation of the factorial).

Support: $(0, +\infty)$

Mean: $\mu = \alpha\beta$

Variance: $\sigma^2 = \alpha\beta^2$

When $\alpha = 1$, the Gamma reduces to the Exponential distribution.

The Gamma distribution is a flexible two-parameter family for positive, right-skewed continuous variables. It is the natural model for quantities that represent accumulated waiting times or the sum of independent exponential random variables. The shape parameter α controls the degree of skewness, so as α increases, the distribution becomes more symmetric and approaches a Normal.

Biological context.

- **Time until death or recovery:** In survival analysis, the time from infection to death or recovery is often modelled with a Gamma distribution, especially when the hazard rate is not constant.
- **Precipitation and environmental data:** The amount of rainfall per month or per event, and many other right-skewed environmental variables, are frequently modelled with a Gamma distribution.
- **Enzyme kinetics:** The time for an enzyme to complete a catalytic cycle can be modelled as the sum of several Exponential waiting times, giving a Gamma distribution.
- **Individual resource acquisition:** The amount of food ingested or energy harvested per foraging bout, which is always positive and right-skewed, is often Gamma-distributed.

Figure 4.8 shows the Gamma distribution with shape $\alpha = 1$ (equivalent to an Exponential). Increasing α shifts mass to the right and progressively reduces skewness toward a symmetric, bell-shaped form.

```
gam <- tidy_gamma(.n = 100, .shape = 1, .scale = 0.4, .num_sims = 5)

pqq_gam <- ggplot(gam, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(gam, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_gam,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

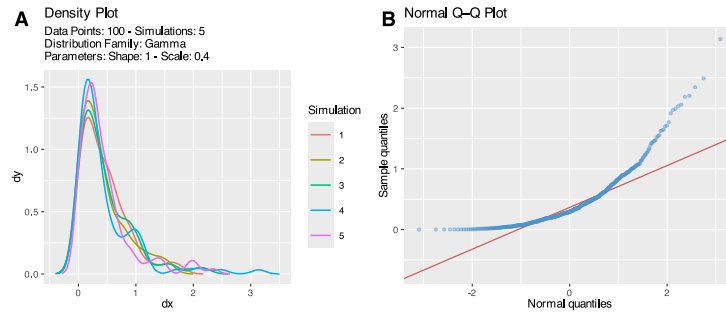


FIGURE 4.8. Gamma distribution with 100 observations per simulation, 5 simulations, shape $\alpha = 1$, and scale $\beta = 0.4$. With $\alpha = 1$, this reduces to an Exponential distribution. Increasing α shifts the mass rightward and reduces skewness. The Q–Q plot shows the pronounced right-skew departure from normality.

We see in Figure 4.8 A a right-skewed distribution: with $\alpha = 1$, most observations are near small values, with a gradual decline toward larger values. This is identical to the Exponential shape.

Figure 4.8 B shows consistent deviation from the Normal line, especially in the upper range. The upward curvature confirms positive skewness. As the shape parameter α increases, this curvature would progressively straighten, and the Q–Q plot would approach the reference line.

4.3.4 The Exponential Distribution

MATHEMATICAL DETAIL

The Exponential distribution:

$$f(x) = \lambda e^{-\lambda x} \quad (4.9)$$

$\lambda > 0$ = rate parameter (events per unit time).

Support: $[0, +\infty)$

Mean: $\mu = \frac{1}{\lambda}$

Variance: $\sigma^2 = \frac{1}{\lambda^2}$

Key property: memorylessness — the probability of waiting an additional time t does not depend on how long you have already waited.

The Exponential distribution models the time between successive events in a Poisson process. It is the only continuous distribution with the memoryless property, meaning past waiting time gives no information about future waiting time.

Biological context.

- **Inter-arrival times on camera traps:** If wildlife move through a survey area according to a Poisson process (independently, at a constant rate), the time between successive detections of the same species follows an Exponential distribution.
- **Radioactive decay:** The time until a single radioactive atom decays is exponentially distributed. This principle underlies radiometric dating used in palaeontology and geochronology.
- **Pathogen transmission:** In simple epidemic models (SIR models), the time from infection to becoming infectious is often modelled as Exponential, corresponding to a constant per-unit-time probability of progressing to the infectious stage.
- **Lifespan under constant hazard:** If an organism faces a constant per-unit-time mortality risk (e.g., random predation), its lifespan is exponentially distributed. Most real organisms do not have constant hazard (risk increases with age), so the Weibull or Gamma is more realistic, but the Exponential serves as the simplest baseline.

Figure 4.9 shows the steep decline from zero that is the hallmark of the Exponential.

```
expon <- tidy_exponential(.n = 100, .rate = 1, .num_sims = 5)

pqq_exp <- ggplot(expon, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(expon, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_exp,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

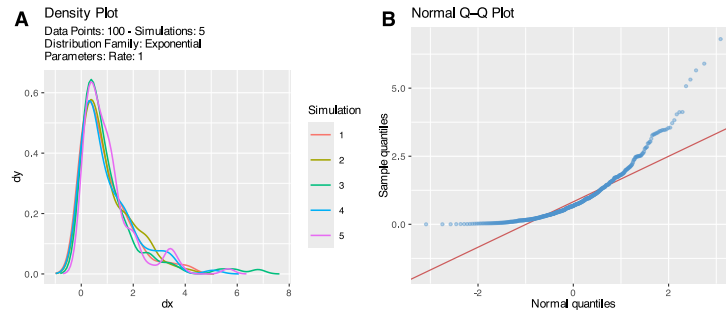


FIGURE 4.9. Exponential distribution with 100 observations per simulation, 5 simulations, and rate $\lambda = 1$. The steep decline from zero reflects the high probability of short waiting times and the low probability of long ones. The Q–Q plot curves strongly upward, indicating heavy right skewness.

Figure 4.9 A shows a steep drop from zero, with most values very small and few large values. This indicates that short waiting times are common and long waiting times are rare. This is a direct consequence of the memoryless property.

In Figure 4.9 B you can see a strong upward curvature away from the Normal line, reflecting extreme right skewness. The departure is among the most pronounced of any distribution in this chapter, reflecting the single-parameter simplicity of the Exponential and its inability to produce anything close to symmetric variation. The lack of symmetry shows that Normal-based methods are unsuitable for raw data of this form.

4.3.5 The Beta Distribution

MATHEMATICAL DETAIL

The Beta distribution:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (4.10)$$

$\alpha, \beta > 0$ are shape parameters; $B(\alpha, \beta)$ is the beta function.

Support: $[0, 1]$

Mean: $\mu = \frac{\alpha}{\alpha + \beta}$

Variance: $\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

When $\alpha = \beta = 1$, the Beta distribution is Uniform on $[0, 1]$. When $\alpha = \beta > 1$, it is symmetric and bell-shaped around 0.5.

The Beta distribution is the natural distribution for proportions and probabilities, so this includes any variable strictly bounded between 0 and 1. Unlike the Normal, it can take a wide variety of shapes depending on its two parameters. These include U-shaped, J-shaped, bell-shaped, or uniform, and they are therefore quite flexible.

Biological context.

- **Vegetation cover:** Percentage vegetation cover per quadrat is a proportion and cannot be below 0% or above 100%. Beta regression, which uses the Beta distribution as the response distribution, is the appropriate model for cover data.
- **Dietary composition:** The proportion of a predator's diet comprised by a particular prey type is bounded between 0 and 1 and is well-modelled by a Beta distribution.
- **Allele frequencies:** The frequency of a particular allele in a population is a proportion and, under neutral drift in a finite population, follows a Beta distribution (the Beta being the stationary distribution of the Wright-Fisher diffusion).
- **Survival and detection probabilities:** In mark-recapture models, detection probabilities and apparent survival rates are bounded in $[0, 1]$ and are typically modelled with Beta priors in a Bayesian framework.

Figure 4.10 shows a right-skewed Beta ($\alpha = 2, \beta = 5$).

```
bet <- tidy_beta(.n = 100, .shape1 = 2, .shape2 = 5, .ncp = 0, .num_sims = 5)

pqq_bet <- ggplot(bet, aes(sample = y)) +
  stat_qq_line(colour = "indianred", linewidth = 0.4) +
  stat_qq(colour = "steelblue3", alpha = 0.4, size = 0.8) +
  labs(x = "Normal quantiles", y = "Sample quantiles",
       title = "Normal Q-Q Plot") +
  theme_grey(base_size = 8)

ggarrange(
  tidyautoplot(bet, .plot_type = "density") + theme_grey(base_size = 8),
  pqq_bet,
  ncol = 2, nrow = 1, labels = "AUTO"
)
```

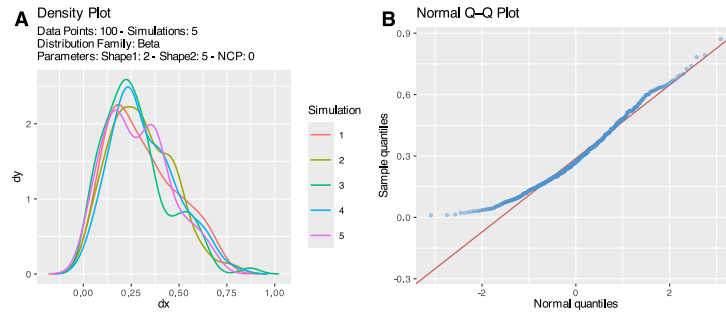


FIGURE 4.10. Beta distribution with 100 observations per simulation, 5 simulations, $\alpha = 2$, and $\beta = 5$. This right-skewed shape on the $[0, 1]$ interval could represent, for example, occupancy probabilities for a moderately rare species across a set of habitat patches. The Q–Q plot shows moderate departure from normality in the upper tail.

Figure 4.10 A shows a distribution confined strictly between 0 and 1, with most values concentrated below 0.5, which is consistent with a detection probability that is more often low than high. The bounded support is an important and defining feature. So, unlike the Normal, the Beta cannot produce values outside $[0, 1]$.

Figure 4.10 B shows deviation from the Normal line, particularly in the upper range. This reflects both the mild right skew and the bounded support. The departure is more moderate than for the Exponential or Log-Normal and illustrates that not all non-normal distributions are equally far from normality.

4.4 IDENTIFYING YOUR DATA'S DISTRIBUTION

Traditionally, visualising the data with a histogram, empirical density, empirical CDF, and QQ plot together gives the most comprehensive picture (Figure 4.11). The QQ plot is especially revealing, because if the data come from the reference distribution, the points should fall on the 45° diagonal line.

```
set.seed(123)
y <- rnorm(n = 200, mean = 13, sd = 2)
old_par <- par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))
hist(y, main = "Histogram", xlab = "y")
plot(density(y), main = "Kernel density", xlab = "y")
plot(ecdf(y), main = "Empirical CDF", xlab = "y")
z <- (y - mean(y)) / sd(y)
qqnorm(z, main = "Normal QQ plot"); abline(0, 1, col = "red")
par(old_par)
```

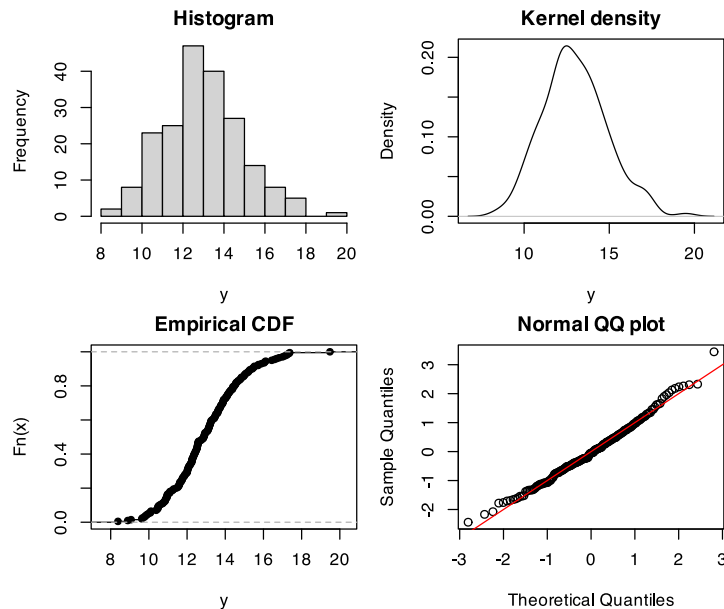


FIGURE 4.11. Four-panel diagnostics for 200 normally distributed observations: histogram, kernel density estimate, empirical CDF, and Normal QQ plot. The close adherence to the 45° line in the QQ plot confirms approximate normality.

The red 45° line marks where points should fall if the data are normal. Systematic departures from the line (such as S-curves indicating heavy or light tails, J-curves indicating skew) point toward alternative distributions.

Here is a practical decision guide to finding your data's distribution. It combines your expert knowledge about the biological process that generated your data with the numerical and visual properties:

Step 1. Identify the support

Eliminate the inappropriate candidates:

- integers only → discrete
- strictly positive → Gamma / Log-normal candidates
- bounded $[0,1]$ → Beta
- unbounded → Normal candidate

Step 2. Identify the process

Most importantly, map the variable to a mechanism:

- counts of independent events → Poisson
- counts with aggregation → Negative Binomial

- repeated success/failure → Binomial
- waiting times → Exponential / Gamma
- multiplicative growth → Log-normal
- additive effects → Normal

Step 3. Check shape visually

Use histogram and Q-Q:

- symmetric → Normal plausible
- right-skewed → Gamma / Log-normal
- heavy tails → consider transformations or alternative families
- many zeros → consider zero-inflation or count models

Step 4. Compare mean–variance relationship (for counts)

- Poisson: mean $\approx$ variance
- Negative Binomial: variance $>$ mean

Step 5. Consider transformation vs model

Transform your data to change the scale of the model, if required:

- log-transform → tests whether Log-normal is appropriate
- square-root → stabilises variance in counts

Step 6. Treat models as approximations

- no real dataset follows a distribution exactly,
- the goal is compatibility, not exact match.

DO IT NOW!

Select your favourite three continuous variables in the BCB7342 field trip set of data. For each one, identify: (1) whether it is discrete or continuous, (2) what biological process generates it, (3) which distribution from this chapter most plausibly describes it (use visualisations + summary statistics), and (4) what would make that distribution choice wrong (e.g., overdispersion, bounded range, always positive). Share your choices with a partner and see if you agree.

Write up your findings and include them in your Progress Portfolio.

4.5 INFERENCE-RELATED DISTRIBUTIONS

Three continuous distributions arise primarily in hypothesis testing and confidence interval construction rather than as direct models of biological data-generating processes. You will encounter all three in later chapters, but a short introduction is useful here.

The Student *t*-distribution arises when estimating the mean of a normally distributed population with an unknown variance from a small sample. It has heavier tails than the Normal, reflecting increased uncertainty from small samples. As

sample size grows, the t -distribution converges to the Normal. We use it in [Chapter 7](#) for one- and two-sample t -tests and in regression to test whether coefficients differ from zero.

The Chi-squared (χ^2) distribution is the sum of the squares of ν independent standard Normal variables. It arises in tests of categorical data ([Chapter 7](#)), in comparing variances, and in model goodness-of-fit tests. It is always positive and right-skewed.

The F -distribution is the ratio of two independent chi-squared variables divided by their degrees of freedom. It arises in ANOVA ([Chapter 8](#)), where we compare a ratio of variances (mean squares) to test whether group means differ. It also appears in the overall F -test of regression models.

None of these three distributions is typically used to model raw biological data; they are the theoretical distributions of test statistics under null hypotheses.

4.6 SAMPLING-LEVEL THINKING

4.6.1 *Samples, Populations, and Sampling Variation*

Everything covered so far describes variation in the observed data themselves. Now, instead of asking how individual measurements vary, I ask how a quantity calculated from those measurements (for example, a sample mean, a proportion, a regression slope) varies when we imagine repeating the study many times. This distinction between the variability of raw observations and the variability of derived estimates is what makes statistical inference possible.

Statistical inference depends on the fact that we almost always observe a **sample** rather than the full **population**. Different samples from the same population will differ from one another. This is **sampling variation**, and it is a fact of nature. If we repeatedly sampled the same population and calculated the sample mean each time, those means would themselves form a distribution; this is the **sampling distribution of the mean**.

This distinction matters greatly:

- The **data distribution** describes the values of individual observations.
- The **sampling distribution** describes the values of a statistic (e.g., the mean) across many hypothetical repeated samples.

Most misunderstandings about p -values and confidence intervals arise from confusing these two distributions.

4.6.2 The Central Limit Theorem

The **central limit theorem (CLT)** says that the sampling distribution of the sample mean tends toward normality as sample size increases, even when the underlying data are not normal. Figure 4.12 demonstrates this nicely: the underlying population is strongly right-skewed (a Gamma distribution), yet the distribution of sample means becomes progressively more symmetric and bell-shaped as n grows.

```
set.seed(42)
population <- tibble(x = rgamma(100000, shape = 2, rate = 1))

simulate_means <- function(sample_n, reps = 4000) {
  set.seed(42)
  tibble(
    n = paste("n =", sample_n),
    mean_x = replicate(reps, mean(sample(population$x, size = sample_n)))
  )
}

bind_rows(
  simulate_means(2),
  simulate_means(5),
  simulate_means(25),
  simulate_means(99)
) >
mutate(n = factor(n, levels = c("n = 2", "n = 5", "n = 25", "n = 99"))) >
ggplot(aes(x = mean_x)) +
  geom_histogram(bins = 30, fill = "salmon", colour = "white") +
  facet_wrap(~n, scales = "free_y") +
  labs(x = "Sample mean", y = "Count")
```

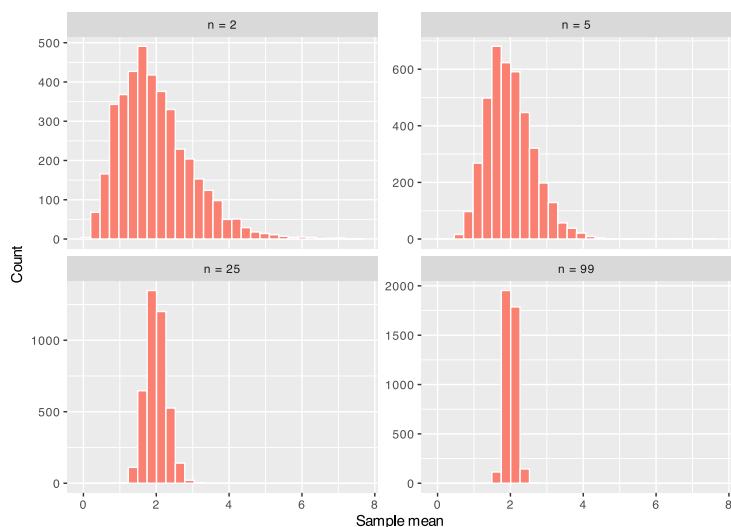


FIGURE 4.12. A simulation demonstrating the central limit theorem. The underlying population is strongly right-skewed (Gamma distribution), yet the sampling distribution of the sample mean becomes increasingly normal as sample size grows from $n = 2$ to $n = 99$.

The CLT is why Normal-based tests work well for means even when raw data are skewed. The important idea is whether the *mean* is approximately normally distributed, not whether the raw data are. Poor experimental design, extreme skewness in very small samples, and non-independence between observations remain real problems that the CLT cannot rescue.

4.6.3 Standard Error

The **standard deviation** (SD) describes variability among individual observations in a sample. The **standard error** (SE) describes variability in an estimate (such as the sample mean) across repeated samples.

For the sample mean:

$$SE_{\bar{x}} = \frac{SD}{\sqrt{n}}$$

As sample size n increases, the standard error decreases proportionally to $1/\sqrt{n}$. So, doubling your sample size reduces the SE by a factor of $\sqrt{2} \approx 1.41$. This formal-

ises the intuitive idea that larger samples produce more precise estimates. The SE is not a property of the raw data; it is a property of the estimation procedure.

→

4.6.4 Confidence Intervals

A **confidence interval (CI)** gives a range of plausible values for an unknown population parameter. The frequentist interpretation of a 95% CI is precise but often misunderstood: if we were to repeat the same sampling procedure many times and calculate a 95% CI each time, approximately 95% of those intervals would contain the true parameter. The interval either does or does not contain the true value, so the “95%” is a statement about the long-run behaviour of the procedure.

Confidence intervals are more informative than bare p -values because they reveal:

- the **scale** of the effect (is the difference biologically meaningful?);
- the **precision** of the estimate (is the interval narrow or wide?);
- **whether biologically important values remain plausible** (does the interval include zero? does it include the value that would matter for management?).

A simple example computing a CI for a small sample:

```
x <- c(12.1, 11.8, 12.4, 12.0, 11.9, 12.3, 12.2, 11.7)

n      <- length(x)
mean_x <- mean(x)
sd_x   <- sd(x)
se_x   <- sd_x / sqrt(n)
t_crit <- qt(0.975, df = n - 1) # two-sided 95% CI uses 97.5th percentile

tibble(
  n      = n,
  mean   = round(mean_x, 3),
  sd     = round(sd_x, 3),
  se     = round(se_x, 3),
  ci_low = round(mean_x - t_crit * se_x, 3),
  ci_high = round(mean_x + t_crit * se_x, 3)
)

R> # A tibble: 1 x 6
R>   n mean sd se ci_low ci_high
R>   <int> <dbl> <dbl> <dbl> <dbl> <dbl>
R> 1     8 12.0 0.245 0.087 11.8 12.3
```

Read the estimate and its interval together: the mean locates the result while the CI conveys its precision.

4.7 LINKING DISTRIBUTIONS TO LATER CHAPTERS

Understanding distributions sits at the junction between data description and formal inference:

- **Chapter 2** and **Chapter 3** describe what the data look like numerically and graphically.
- This chapter adds probabilistic structure: why data have the shapes they do and what process likely generated them.
- **Chapter 5** uses sampling distributions and uncertainty to formalise hypothesis testing.
- **Chapter 7** and **Chapter 8** assume approximately normal residuals, which is justified by the CLT for means.
- **Chapters 14–15** extend regression to multiple predictors, where residual distributions remain central.
- **Chapter 20** onwards addresses **generalised linear models**, which explicitly relax the normality assumption and model responses using Binomial, Poisson, Negative Binomial, Gamma, or Beta distributions directly.

4.8 SUMMARY

- Distributions describe the pattern of variation in data and reflect the biological processes that generated them.
- **Discrete distributions** (Bernoulli, Binomial, Poisson, Negative Binomial, Geometric) apply to countable outcomes such as presence/absence, counts per quadrat, or trials until success.
- **Continuous distributions** (Normal, Log-Normal, Gamma, Exponential, Beta) apply to measurements that can take any value within a range.
- The Normal distribution arises from additive processes; the Log-Normal arises from multiplicative ones; Gamma and Exponential model waiting times and accumulated events; Beta is the natural choice for bounded proportions.
- Statistical inference depends on **sampling variation**: different samples give different estimates.
- The **central limit theorem** explains why sampling distributions of means are approximately Normal even when raw data are not.
- **Standard errors** quantify precision of estimates; **confidence intervals** express that precision on the scale of the parameter of interest.
- The Cullen-Frey graph and QQ plots help identify which distribution is most compatible with observed data.

Part II

Inference

This part moves from describing data to drawing formal conclusions from them. Here the question changes from “what do the data look like?” to “what can I reasonably claim beyond this sample?” That shift introduces hypotheses, decision rules, assumptions, and the standard inferential tools used to compare groups or assess association.

The emphasis is not on memorising named tests. It is on learning how inferential questions differ, how assumptions shape what a result means, and how to match a method to the design and data structure that generated the evidence. The chapters therefore build from general inferential logic into specific families of tests and then close with a decision guide that shows how these pieces fit together.

Once this logic is established, the book can move beyond isolated tests and into models that describe responses, predictors, and structure more explicitly.

5

Statistical Inference

An Introduction to Inferential Statistics

IN THIS CHAPTER

- The concept of inferential statistics
- Hypothesis testing
- p -values and decision thresholds
- Assumptions
- How inferential results are reported in a journal article

TASKS TO COMPLETE IN THIS CHAPTER

- None

In Chapter 5, we arrive at the core purpose of this book: a practical exploration of the *inferential statistics* essential to modern biology. Using this toolkit, we can transform individual data points into broader scientific insights for the best possible explanations for the phenomenon being studied. These statistical methods are fundamental to hypothesis-driven research and provide the means to *generalise findings* from a sample to a population or predict future outcomes. The reasoning is that sample estimates vary from one sample to the next, thus offering only a partial glimpse into the broader population it represents. In Chapter 4 I showed how to describe that variation. Through inferential statistics, we can draw conclusions that extend beyond the sample by assessing how robustly the sample supports hypotheses about biological processes and populations.

Inferential statistics build on the foundation of *exploratory data analysis (EDA)*, which leans heavily on descriptive, or summary, statistics. These initial, descriptive methods summarise the core features of data (their distribution described by a central tendency and spread) and offer a clear, distilled view of the data's characteristics. Descriptive statistics capture a snapshot of the dataset, but they stop short of allowing us to deduce broader truths about the population from which the data arose. Inferential statistics then takes this groundwork and advances it within the framework of *probability theory*. In doing so, we draw conclusions and quantify the uncertainty around these deductions.

Which probability model best represents a dataset is not an arbitrary choice; it flows from the biological mechanisms that generate the data. Confident, effective use of inferential statistics therefore requires an appreciation of both biological theory and the unique characteristics of the data at hand. The probability distribution of data is closely tied to the natural processes behind it, a connection that is often underappreciated. Biological data are shaped by a multitude of factors, partly deterministic, but very much affected by genetics, environmental variables, and stochastic variations among them, each leaving an imprint on the underlying statistical distribution. For example:

1. **Plant Height (Normal Distribution):** The heights of individual plants in a population typically follow a *normal distribution*. This distribution arises from the combined effects of genetic factors and environmental conditions that influence plant growth, such as soil quality, light, and water availability.
2. **Litter Size in Mammals (Poisson Distribution):** In many mammal species, the number of offspring per litter may follow a *Poisson distribution*, which is common for count data. This distribution reflects the biological processes involved in reproduction, where most females have an average litter size, and larger litters are progressively rarer.

In each case, the mechanism that produces the data also determines its distribution, and that distribution is what we assess when constructing probability statements about the world.

The central idea of inference is to ask how surprising our observed result would be if a stated null model were true. Knowing which probability distribution governs data under that null is therefore what makes the comparison meaningful and the conclusion defensible. This thinking is the foundation of hypothesis testing, which structures every analytical step that follows.

5.1 THE BASIC WORKFLOW OF HYPOTHESIS TESTING

Hypothesis testing is best understood as a sequence of linked steps:

1. state a **null hypothesis** (H_0) and an **alternative hypothesis** (H_a);
2. calculate a **test statistic** from the sample data;
3. evaluate how compatible that statistic is with the null model;
4. compare the result with a chosen decision threshold, usually α ;
5. interpret the outcome in scientific context.

This workflow recurs throughout the rest of the module whether we are comparing means, testing associations, or fitting models.

5.2 HYPOTHESES

Hypothesis testing starts by setting up two competing explanations for the data. The **null hypothesis** (H_0) usually represents no effect, no difference, or no association. The **alternative hypothesis** (H_a) represents the pattern we would regard as evidence against that null model. Only one of them can be retained after running the statistical test.

Hypothesis testing evaluates whether the observed data are consistent with the null model. It does not prove a hypothesis true, but it asks whether the evidence is strong enough to reject H_0 .

At the heart of many basic scientific questions is the simple comparison “Is A different from B?” In statistical notation:

H_0 : Group A is not different from Group B

H_a : Group A is different from Group B

More formally:

$$H_0 : \bar{A} = \bar{B}$$

$$H_a : \bar{A} \neq \bar{B}$$

$$H_0 : \bar{A} \leq \bar{B}$$

$$H_a : \bar{A} > \bar{B}$$

$$H_0 : \bar{A} \geq \bar{B}$$

$$H_a : \bar{A} < \bar{B}$$

NOTE

The hypothesis pair written with $\neq$ is **two-sided** because it asks whether the two means differ in either direction. The two hypothesis pairs written with $>$ and $<$ are **one-sided** because they specify direction explicitly. This distinction becomes important when we formalise these tests in [Chapter 7](#).

We should also be precise about the language of decisions. In frequentist hypothesis testing, we either **reject** H_0 or **fail to reject** H_0 . We do not “prove” H_a , and we do not “accept” H_0 as true.

DO IT NOW!

For each of the following scenarios, write the appropriate null and alternative hypotheses and state whether the test should be one-sided or two-sided. Justify your choice.

- A pharmaceutical company tests whether a new drug *reduces* blood pressure more than a placebo.
- An ecologist measures whether a marine reserve has *changed* average fish size and abundance (either increased or decreased) after five years.
- A food scientist checks whether a new preservation method *increases* shelf life beyond the industry standard of 14 days.

After writing your hypotheses, discuss with a partner: why does the directionality matter for the p -value you will eventually compute?

5.3 TEST STATISTICS, P-VALUES, AND DECISION RULES

5.3.1 *Test Statistics*

A **test statistic** is a numerical summary of the evidence in the sample. Different methods use different statistics: a t -test uses a t statistic, ANOVA uses an F statistic, correlation uses the correlation coefficients ρ or τ , and regression uses slopes, R^2 , and related summaries.

The common purpose is that all these statistics reduce the sample evidence to a quantity that can be evaluated under the null model.

5.3.2 *The p-Value*

The **p -value** is the probability of obtaining a result at least as extreme as the one observed, assuming that H_0 is true. A p -value is *not* the probability that H_0 is true. It

is a measure of how incompatible the data are with the null model. Smaller p -values indicate stronger incompatibility.

For example, assume our statistical test results in a p -value of 0.03, then, within the standard frequentist interpretation of a p -value we say:

“If H_0 is true, there is a 3% probability of obtaining a result at least as extreme as ours.”

That statement has two parts.

1. First, “If H_0 is true”. This is the condition. We temporarily assume the null hypothesis is correct. For example, if H_0 says there is no treatment effect, no association, or no difference between groups, we calculate probabilities under that assumption.
2. Second, “a result at least as extreme as ours”. This means not only the exact result we observed, but also any result that would provide *equal or stronger* evidence against H_0 , according to the test statistic being used. So the p -value counts your result and more extreme ones in the tail or tails of the null distribution.

So, for a p -value of 0.03, the interpretation is:

If the null hypothesis were true, results this extreme or more extreme would arise about 3 times in 100 repeated samples.

5.3.3 The Significance Threshold, α

Before looking at the result, we usually choose a decision threshold, denoted α . In many biological studies, $\alpha = 0.05$ remains conventional. Other values, such as 0.01 or 0.001, may be used when false positives carry greater cost.

Thresholds are conventions that guide decisions. They do not replace expert judgement.

→

5.3.4 The Decision Rule

The usual rule is:

- if $p \leq \alpha$, reject H_0 ;
- if $p > \alpha$, fail to reject H_0 .

That rule is operational, but it is not the whole interpretation. A result just below 0.05 and one just above 0.05 do not represent radically different biological realities. The threshold helps formalise a decision but it does not remove the need to consider effect size, uncertainty, design quality, and context.



Noah Motion

Statistical te
"approach"
flee from it,
around it, lo
unknowingl
from afar, or
it in remote

810 Likes

24 Jan 2018 at 17:40

5.3.5 Interpretation

When $p \leq \alpha$, the data are commonly described as statistically significant. When $p > \alpha$, the evidence is not strong enough to reject H_0 under the chosen threshold.

In neither case does the p -value tell us whether the effect is biologically important. That judgement depends on the size of the effect, the uncertainty around it, and the scientific context.

TYPE I AND TYPE II ERRORS

A **Type I error** is the false rejection of H_0 : concluding that there is an effect when there is none. A **Type II error** is the false failure to reject H_0 : missing a real effect. The threshold α controls the long-run rate of Type I error if the assumptions of the test hold. A smaller threshold is more conservative, but that gain usually comes at a cost: it becomes harder to detect real effects.

5.4 ASSUMPTIONS

Inferential procedures rely on assumptions about the data, the sampling or experimental design, and the model. For example, many tests assume independence, appropriate distributional form, and constant variance.

Testing these assumptions are part of the process leading to inference. If the assumptions are badly violated, the reported p -values, confidence intervals, and test statistics may not mean what we think they mean.

This is the purpose of [Chapter 6](#), where I will explain how to check assumptions and what to do when they fail.

5.5 PARAMETRIC AND NON-PARAMETRIC METHODS

Many standard inferential procedures are **parametric**. That is, they have assumptions about normality. But this does not mean they require the raw data themselves to be perfectly normal in every case. More often, the assumptions concern the distribution of residuals, the form of the model, or the behaviour of the test statistic under the null hypothesis.

When those assumptions are not reasonable, we often use **non-parametric** or rank-based alternatives. I introduce those alternatives within the relevant method families in [Chapter 7](#), [Chapter 8](#), and [Chapter 9](#), and then brought together in [Chapter 10](#).

5.6 REPORTING INFERENCE RESULTS

Worked examples in the later chapters end with a Reporting section. The purpose of those sections is to demonstrate how the outcome of an analysis would be written in a scientific paper.

In practice, the reporting should follow the normal requirements of a journal article. The details differ from one method to another, but the overall structure remains more-or-less like this:

GENERAL REPORTING REQUIREMENTS

- In the **Methods**, state what was measured, what the response variable was, what the grouping or predictor variables were, and which statistical test or model was used.
- In the **Methods**, state any design features that are essential for interpretation, such as pairing, repeated measures, blocking, or the true experimental unit.
- In the **Methods**, cite the statistical method and, where relevant, the R package used to implement it. Statistical analysis is part of the scholarly record and should be cited as such.
- In the **Results**, report the main biological pattern in plain language before or alongside the formal test result.
- In the **Results**, report the core statistical outcome using the relevant quantities for the method at hand, such as a t statistic, F statistic, correlation coefficient, slope, confidence interval, or similar effect summary.
- In the **Results**, do not confuse statistical significance with biological importance.
- In the **Discussion**, explain what the result means biologically rather than merely restating whether H_0 was rejected.
- In the **Discussion**, note any important limitation, assumption issue, or caution that affects the interpretation of the result.
- Keep the reporting concise and readable. A good Results paragraph should sound like a journal article, not like raw statistical output copied from R.

This is the general reporting approach that I apply throughout the rest of the Biostatistics module. So, in [Chapter 7](#), [Chapter 8](#), [Chapter 9](#), and the later modelling chapters, with each worked example I therefore include a “Reporting” box that separates the **Methods**, **Results**, and **Discussion**.

If you need the correct citation for an R package, ask R for it directly:

```
citation()
citation("stats")
citation("ggplot2")
```

`citation()` with no package name gives the citation for R itself. `citation("package_name")` gives the citation for a specific package. Once

that reference is in your `.bib` file, you can cite it in Quarto with the usual citation syntax, for example `[@R-base]` or `[@R-ggplot2]`, depending on the citation key in the BibTeX entry.

The method itself should also be named clearly in the prose. A Methods sentence can therefore look like this:

```
Bill length was compared between species with Welch's two-sample
*t*-test in R [@R-base].
```

Or, when a contributed package is central to the analysis:

```
Estimated marginal means were obtained with the *emmeans* package
[@R-emmeans].
```

5.7 CONCLUSION

Inferential statistics lets us use a sample to say something disciplined about a larger population. In the workflow, we state hypotheses, calculate a test statistic, evaluate it under a null model, apply a decision threshold, and interpret the result in context.

The harder part is expert scientific judgement. We still need to ask whether the design was sound, whether the assumptions were reasonable, and whether the detected effect matters biologically. Those questions remain central in every chapter that follows.

The next five chapters unpack that general logic into specific inferential families. I begin with assumption checking, then move through *t*-tests, ANOVA, and correlation, and finally close the inferential block with a method-selection guide. The goal is that each later method is read as one expression of the same inferential workflow introduced here, not as a separate ritual with its own isolated rules.

6

Raw-Data Assumptions and Transformations

Diagnostics for Groups, Pairs, and Associations

IN THIS CHAPTER

- how assumption checks depend on data structure
- normality within groups
- variance across groups
- paired differences
- joint structure in associations
- transformations and what to do when assumptions fail

TASKS TO COMPLETE IN THIS CHAPTER

- None

Inferential statistics can be broadly categorised into *parametric* and *nonparametric* methods. The choice between them hinges on understanding the distribution of our data and the assumptions underlying each method. Parametric methods rely on specific assumptions about the underlying probability distribution of the population from which the sample data are drawn. The two key assumptions are *normality*, that the data follow a normal (Gaussian) distribution, and *homoscedasticity*, which requires equal variances across groups or levels of predictors.

Strictly speaking, the core parametric requirement is not normality per se but that the data follow a *known* probability distribution specified in advance. When the response is a count, a proportion, or a binary outcome, other distributions (Poisson, binomial) apply, and methods such as Generalised Linear Models extend the parametric framework accordingly (Part V). For the tests covered in Parts II and III of the book (*t*-tests, ANOVA, regression, and correlation) the relevant distribution is normal, so normality and homoscedasticity are the assumptions that matter here.

Nonparametric methods offer an alternative when data do not conform to any known distribution, or when assumptions cannot be met and the sample is too small for the Central Limit Theorem (CLT) to compensate. They make fewer assumptions, are more robust to non-standard distributions, and trade a modest reduction in statistical power for greater generality. For most biological datasets of adequate size, a well-chosen parametric test is preferred, but nonparametric alternatives become the more defensible choice when distributional requirements clearly cannot be satisfied.

Choosing between the two approaches requires first assessing whether your data meet parametric assumptions. With larger samples, the CLT means that moderate departures from normality become less important, because the sampling distribution of the mean approaches normality regardless. With small samples, the same violations carry more weight and nonparametric alternatives deserve serious consideration. When assumptions are met, parametric tests are more powerful but when they are violated, that power advantage dissipates.

Assumption checks determine whether a planned inferential method is defensible. In this chapter, I apply those checks to **raw data** rather than to model residuals — residual-based diagnostics belong to the later regression chapters, beginning with [Residuals and Model-Based Diagnostics](#). My aim here is to prepare you for the tests immediately ahead: *t*-tests, ANOVA, and [Correlation and Association](#).

The discussion is organised around data structure. Group comparisons require normality checks within groups and homoscedasticity checks across them. Paired designs require checks on the differences. Associations require checks on the joint pattern of both variables.

6.1 ASSUMPTION-CHECKING WORKFLOW

Assumption checking follows a series of steps:

1. identify the structure of the analysis;
2. inspect the raw data visually;
3. assess the assumptions graphically;
4. use formal tests to support judgement;

5. decide whether to proceed, transform, or change method;
6. re-check if the analysis changes.

Different assumptions often fail together which is why I tie the workflow the analysis rather than to a list of tests.

6.2 IDENTIFYING THE DATA STRUCTURE

Before checking assumptions, identify the structure of the data:

- **Are there groups?** Then check each group separately.
- **Are the measurements paired?** Then check the differences.
- **Are both variables continuous?** Then check the joint pattern of the relationship.

Assumptions apply to the structure implied by the analysis, not to the dataset as a whole. I introduced the grouped-data example earlier in [Comparing Groups](#), where I emphasised how plots change once categories are recognised explicitly.

LOOKING AHEAD: RESIDUALS

In the regression chapters, I will apply these same ideas to **residuals**, which are the differences between observed values and model predictions, but here I develop the assumptions at the level of raw data. The extension to residuals belongs with the regression series of chapters.

6.3 GROUP COMPARISONS: NORMALITY WITHIN GROUPS

I begin with a grouped comparison that prepares directly for [t-tests](#) and [ANOVA](#).

```
peng <- penguins %>%
  drop_na(bill_length_mm, body_mass_g, species)
```

I compare bill length across penguin species. The grouping structure is seen in the dataset where you can see observations belonging to several species. This is the same grouped-data idea introduced in [Comparing Groups](#), but here the aim is diagnostic rather than descriptive. The assumption applies within each species, not to the pooled data.

6.3.1 Graphical Checks

Start with plots. Histograms show the shape within each species. A **Q-Q plot** compares the observed quantiles of the data with the quantiles expected from a normal

distribution. If the distribution is close to normal, the points fall roughly along a straight reference line. These grouped and pooled views are shown in Figure 6.1.

```
plt1 <- ggplot(peng, aes(body_mass_g)) +
  geom_histogram(bins = 20, fill = "grey70", colour = "white") +
  facet_wrap(~species, scales = "free_x") +
  labs(x = "Body mass (g)", y = "Count")

plt2 <- ggplot(peng, aes(sample = body_mass_g)) +
  stat_qq(shape = 21, fill = "salmon", colour = "black") +
  stat_qq_line(colour = "red4") +
  labs(x = "Theoretical quantiles", y = "Sample quantiles")

plt3 <- ggplot(peng, aes(sample = body_mass_g)) +
  stat_qq(shape = 21, fill = "salmon", colour = "black") +
  stat_qq_line(colour = "red4") +
  facet_wrap(~species, scales = "free") +
  labs(x = "Theoretical quantiles", y = "Sample quantiles")

ggarrange(plt1, plt2, plt3, ncol = 1, labels = "AUTO")
```

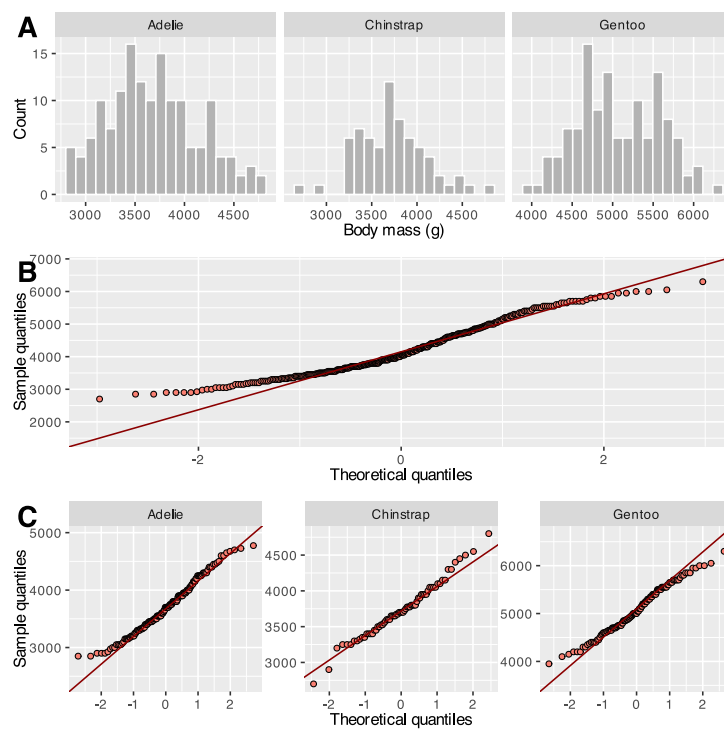


FIGURE 6.1. Normality checks for grouped data. A) Histograms of bill length within each penguin species. B) A pooled Q-Q plot that ignores species. C) Q-Q plots drawn separately within species.

In Figure 6.1, the pooled Q-Q plot bends because it combines Adelie, Chinstrap, and Gentoo penguins into one mixture distribution. The grouped Q-Q plots are the relevant ones because the comparison is made across species. That is the pattern we carry forward into Chapter 7 and Chapter 8.

6.3.2 Formal Test

The most commonly used formal test is the **Shapiro-Wilk test**, `shapiro.test()`. Its null hypothesis is that the data are compatible with normality:

H_0 : the distribution is compatible with normality

H_a : the distribution departs from normality

The test statistic is:

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (6.1)$$

Here, W is the Shapiro-Wilk test statistic, a_i are coefficients that depend on the sample size and expected order statistics under normality, $x_{(i)}$ is the i -th ordered observation, and $\bar{x}$ is the sample mean.

Run the test within species, not on the pooled sample:

```
peng >
  group_by(species) >
  summarise(
    shapiro_w = as.numeric(shapiro.test(bill_length_mm)$statistic),
    shapiro_p = as.numeric(shapiro.test(bill_length_mm)$p.value)
  )
```

```
R> # A tibble: 3 x 3
R>   species    shapiro_w shapiro_p
R>   <fct>      <dbl>      <dbl>
R> 1 Adelie      0.993      0.717
R> 2 Chinstrap   0.975      0.194
R> 3 Gentoo      0.973      0.0135
```

When p is greater than 0.05, the test shows no strong evidence against normality. That result does not prove exact normality, but it shows that the observed departure is not strong enough for this test to detect at the sample size available. There is sufficient evidence against the null, suggesting that Gentoo penguins have non-normal data.

That result does not automatically end the analysis, but it does change the next step. First, inspect the plots in Figure 6.1 and ask whether the departure is mild or severe. Second, continue to the variance check below, because non-normality and unequal variance often appear together. If one or more groups are clearly non-

normal, and especially if variance also differs among groups, the later comparison should use a method that tolerates those features better. In the chapters that follow, that may mean Welch's version of the test, a rank-based alternative, a transformation followed by a fresh round of checks, or a different model altogether. The decision depends on the size of the departure, the sample sizes, and the biological scale on which you need to interpret the result.

DO IT NOW!

Using the penguins dataset (load **palmerpenguins**), run the normality checks for `flipper_length_mm` instead of `bill_length_mm`. For each species, produce a Q-Q plot and apply the Shapiro-Wilk test. Do the conclusions differ from what was found for bill length? Copy the code from above and adapt it. Report your W statistics and p-values and describe in one sentence whether you would consider the flipper length data approximately normal within each species.

6.3.3 Other Tests for Normality

Several other tests can be used to assess whether data are compatible with normality:

- **Kolmogorov-Smirnov test** compares the empirical distribution of a sample with a specified theoretical distribution. In R use `ks.test()`.
- **Anderson-Darling test** is another goodness-of-fit test, available as `ad.test()` in packages such as **nortest**.
- **Lilliefors test** is a modification of the Kolmogorov-Smirnov test for estimated mean and variance. See `lillie.test()`.
- **Jarque-Bera test** is based on skewness and kurtosis. See `jarque.bera.test()` in **tseries**.
- **Cramer-Von Mises test** is another goodness-of-fit approach. See `cvm.test()` in **goftest**.

These tests support the plots. They do not replace them.

LOOKING AHEAD: NORMALITY OF RESIDUALS

In the regression chapters, I check normality with Q-Q plots of residuals rather than with grouped raw-data plots. That extension belongs with fitted models and is introduced in [Chapter 11](#).

6.4 GROUP COMPARISONS: VARIANCE ACROSS GROUPS

Grouped comparisons also require attention to spread. For methods such as one-way ANOVA, large differences in variance across groups can affect the inferential result.

We use penguin body mass because the spread differs more clearly across species:

```
peng_var <- peng %>%
  group_by(species) %>%
  summarise(sample_var = var(body_mass_g), .groups = "drop")

plt1 <- ggplot(peng, aes(species, body_mass_g, fill = species)) +
  geom_boxplot(show.legend = FALSE, alpha = 0.7) +
  labs(x = "", y = "Body mass (g)")

plt2 <- ggplot(peng_var, aes(species, sample_var, fill = species)) +
  geom_col(show.legend = FALSE, alpha = 0.7) +
  labs(x = "", y = "Sample variance")

ggpubr::ggarrange(plt1, plt2, ncol = 2, widths = c(1.4, 1))
```

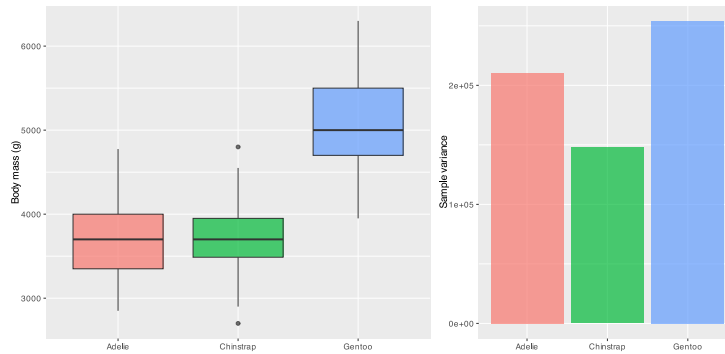


FIGURE 6.2. Variance checks for grouped data. A) Box plots of body mass by penguin species. B) Sample variances by species.

The box plots and variance bars in Figure 6.2 show the spread clearly. Large differences in spread suggest heteroscedasticity.

The most commonly used formal test is **Levene's test**, `car::leveneTest()`. Its null hypothesis is that the group variances are equal:

$$H_0 : \sigma_A^2 = \sigma_B^2$$

$$H_a : \sigma_A^2 \neq \sigma_B^2$$

The Levene test statistic is:

$$W = \frac{(N - k)}{(k - 1)} \cdot \frac{\sum_{i=1}^k n_i (\bar{z}_i - \bar{z})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2} \quad (6.2)$$

Here, N is the total sample size, k is the number of groups, n_i is the sample size in group i , z_{ij} is the absolute deviation of observation j in group i from its group centre, $\bar{z}_i$ is the mean of those deviations within group i , and $\bar{z}$ is the overall mean deviation.

```
car::leveneTest(body_mass_g ~ species, data = peng)
```

```
R> Levene's Test for Homogeneity of Variance (center = median)
R>      Df F value    Pr(>F)
R> group  2  5.1203 0.006445 **
R> ---
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A small p -value indicates evidence against equal variance. This structure is used directly in [ANOVA](#).

```
peng_var
```

```
R> # A tibble: 3 x 2
R>   species    sample_var
R>   <fct>         <dbl>
R> 1 Adelie      210283.
R> 2 Chinstrap   147713.
R> 3 Gentoo      254133.
```

6.4.1 Other Tests for Homogeneity

Several other tests are available for comparing variances:

- **F-test** compares two variances. Use `var.test()`. It assumes normality.
- **Bartlett's test** compares variances across multiple groups. Use `bartlett.test()`. It also assumes normality.
- **Brown-Forsythe test** is a median-based modification of Levene's test and is more robust to non-normality.
- **Fligner-Killeen test** is a robust non-parametric variance test available in base R as `fligner.test()`.

LOOKING AHEAD: RESIDUAL PLOTS

In the regression chapters, I assess changing variance with residual-versus-fitted plots. That diagnostic belongs with fitted models and is introduced in [Chapter 11](#).

6.5 PAIRED DESIGNS: CHECK THE DIFFERENCES

In paired designs, the assumption applies to the **differences**, not to the raw measurements.

```
set.seed(1)
paired <- tibble(
  before = rnorm(20, 10, 2),
  after = before + rnorm(20, -1, 1)
) >
mutate(diff = after - before)
```

The paired structure means that each after value belongs to the same unit as one before value. The analysis will therefore be built on the differences, which are visualised in Figure 6.3.

```
plt1 <- ggplot(paired, aes(diff)) +
  geom_histogram(bins = 12, fill = "grey70", colour = "white") +
  labs(x = "After - before", y = "Count")

plt2 <- ggplot(paired, aes(sample = diff)) +
  stat_qq(shape = 21, fill = "salmon", colour = "black") +
  stat_qq_line(colour = "red4") +
  labs(x = "Theoretical quantiles", y = "Sample quantiles")

ggarrange(plt1, plt2, ncol = 2, labels = "AUTO")
```

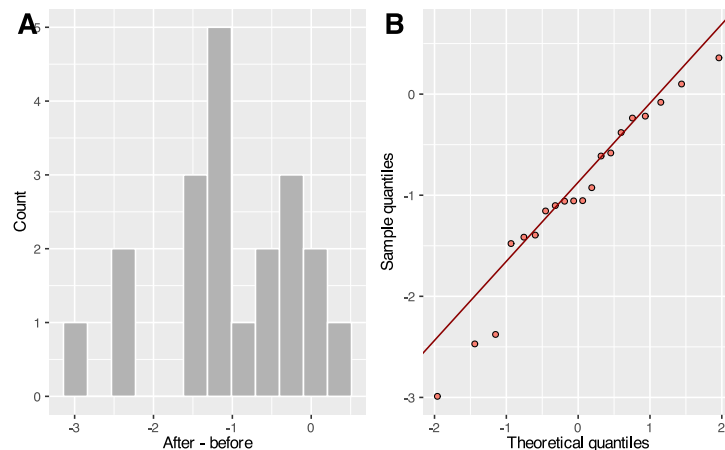


FIGURE 6.3. Checks for paired data. A) Histogram of paired differences. B) Q-Q plot of paired differences.

```
shapiro.test(paired$diff)

R>
R> Shapiro-Wilk normality test
```

```
R>
R> data: paired$diff
R> W = 0.94617, p-value = 0.3127
```

The histogram and Q-Q plot in Figure 6.3 are the relevant diagnostics because the paired t -test is built on the differences, not on the raw before and after values. This is the structure used directly in the paired t -test in Chapter 7.

6.6 ASSOCIATIONS: CHECK THE JOINT STRUCTURE

Association problems have a different geometry. There are no groups, and there are no paired differences. The main diagnostic is the joint structure of the two variables.

```
ggplot(trees, aes(Height, Volume)) +
  geom_point(shape = 21, fill = "salmon", colour = "black") +
  labs(x = "Height", y = "Volume")
```

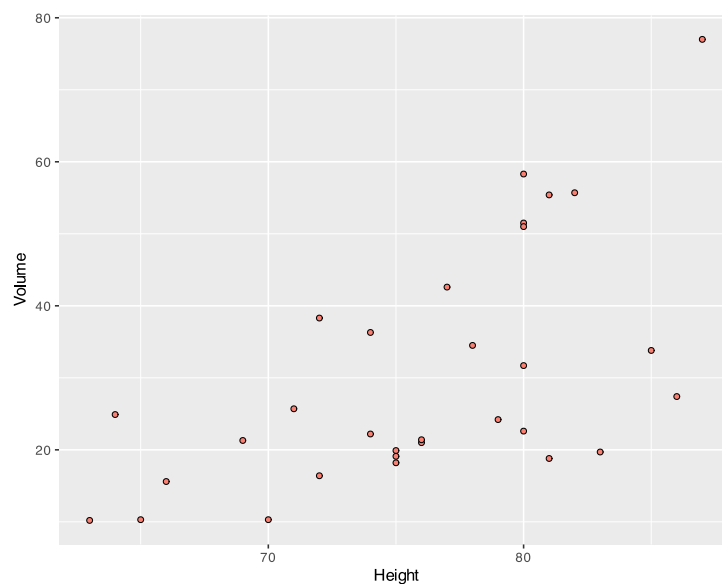


FIGURE 6.4. Scatter plot of tree height and volume from the base R `trees` dataset.

In Figure 6.4, the questions are whether the relationship is approximately linear and whether a few extreme values dominate the pattern. Checking each variable for normality in isolation would miss the main issue. The structure used here leads directly into [Correlation and Association](#).

6.7 INDEPENDENCE

Independence is set by design. If repeated measurements are taken on the same individual, or if samples are nested within a site, tank, quadrat, or transect, then the observations are not independent.

Suppose ten leaves are measured from each of three plants, but the analysis treats the thirty leaves as if they came from thirty independent plants. That is **pseudoreplication**. The apparent sample size is inflated, and the inferential result becomes too optimistic because the real experimental unit was the plant, not the leaf.

Check independence by asking:

1. what is the true experimental unit;
2. are observations repeated, nested, spatially clustered, or temporally linked;
3. does the design require a paired test, repeated-measures method, or mixed model.

We return to these issues in [Pseudoreplication](#) and [Dependence and Mixed Models](#).

DO IT NOW!

For each scenario below, state whether the observations are independent and identify the potential source of non-independence if there is one:

- a. You measure oxygen consumption in 20 individual fish, one per tank, across two temperature treatments.
- b. You measure body mass in 30 individual fish, but six fish share each of five tanks, and tanks are the actual treatment units.
- c. You sample 15 quadrats along a single 100 m transect to estimate cover of an invasive alga.
- d. You measure blood pressure in 10 patients before and after a treatment.

Which of these would call for a paired test, a mixed model, or a simple independent-samples test? Discuss with a partner.

6.8 REPORTING ASSUMPTION CHECKS

WRITE-UP

Methods

Penguin body mass was examined by species before any formal group comparison. Normality was assessed within species using histograms, Q-Q plots, and the Shapiro-Wilk test. Homogeneity of variance across species was assessed with box plots, direct comparison of sample variances, and Levene's test.

Results

Body-mass distributions differed among penguin species in both location and spread. The species-specific histograms and Q-Q plots in Figure 6.1 gave the relevant diagnostic view because the assumption applied within species rather than to the pooled dataset. The box plots and variance summaries in Figure 6.2 reinforced that variance was not uniform across species, so a later comparison of species means would need to take that heterogeneity into account.

Discussion

In a journal-style account, these diagnostics justify the choice of inferential method. Here they show that the grouped structure of the penguin data must be respected and that unequal variance among species would need to be handled explicitly in the later analysis.

6.9 RESPONDING TO ASSUMPTION VIOLATIONS

Once assumptions have been checked, the response is usually one of four:

1. proceed because the departures are minor and the planned method is adequate;
2. change to a method that matches the data more closely;
3. transform the response and then re-check the assumptions;
4. recognise that the design does not support the intended inference.

The best response depends on the structure of the problem. Counts, proportions, and non-independent observations often need a different method rather than a transformed version of the same method.

The non-parametric alternatives for the main inferential methods are introduced within the relevant method chapters: Wilcoxon procedures in Chapter 7, Kruskal-Wallis in Chapter 8, and rank-based correlation in Chapter 9. Chapter 10 then summarises how those choices fit together.

6.10 WHY TRANSFORM DATA?

Transformations change the scale of the response so that the data are more compatible with the assumptions of the planned method. They are often used to reduce right-skew, stabilise variance, or linearise a relationship.

Use a transformation only when it improves the fit between the data and the method in a way that still supports a clear biological interpretation.

After transformation, quantities such as means and confidence intervals usually need to be back-transformed before reporting. Back-transformed uncertainty is often asymmetric.

“Torture numbers and they will confess to anything” — Gregg Easterbrook

6.11 TRANSFORMATION DECISION GUIDE		
Data problem	Usual response	Comment
Strong right-skew in positive continuous data	$\log(x)$ or $\log_{10}(x)$	Often useful for multiplicative processes or long right tails
Count data with variance increasing with the mean	$\sqrt{x}$ or a count model	A Poisson or negative-binomial GLM is often better
Proportions near 0 or 1	Usually a binomial GLM	Arcsine transformations are now often unnecessary
Severe right-skew with difficult scale	$1/x$ in rare cases	Hard to interpret biologically
Negative skew	Reflect, then transform if needed	Often suggests that a different model or scale may be better

If the response has its own natural error structure, such as counts, proportions, presence-absence data, or survival times, a generalised model is often better than transformation.

DO IT NOW!

Use the decision guide above to recommend a transformation (or decide that transformation is not appropriate) for each of the following:

- Seabird nest counts per island (range: 0 to 2500; variance increases with the mean).
- Percentage cover of lichen on rock surfaces (range: 2% to 98%; most sites around 50%).
- Salmon body mass in grams (range: 1.5 to 8.2 kg; slight right-skew).
- Presence or absence of a parasitic worm (binary: 0 or 1).
- Time-to-death of bacteria exposed to an antibiotic (range: 5 to 240 minutes; strongly right-skewed).

For each, state the transformation or model type you would recommend and give a one-sentence justification.

6.12 COMMON TRANSFORMATIONS

6.12.1 Log Transformation

A log transformation is often useful for positive, right-skewed data. Use $\log(x)$ or $\log_{10}(x)$. The statistical conclusion does not depend on whether the base is e or 10 because the two scales differ only by a constant factor.

If zeros are present, a constant is sometimes added first, for example $\log(x + 1)$. Do that for a clear reason, not mechanically.

6.12.2 Square-Root Transformation

The square-root transformation, $\sqrt{x}$, is often used for count-like data or for responses where the variance increases with the mean.

6.12.3 Arcsine Transformation

The arcsine square-root transformation,

$$y' = \arcsin(\sqrt{y}) \quad (6.3)$$

was historically used for proportions between 0 and 1. It still appears in older literature, but a binomial GLM is often the better solution because it respects the mean-variance relationship directly.

The back-transformation is:

$$y = \sin(y')^2 \quad (6.4)$$

6.12.4 Reciprocal Transformation

The reciprocal transformation, $1/x$, can reduce strong right-skew, but it often produces a scale that is hard to explain biologically.

6.12.5 Square and Cube Transformations

Square and cube transformations can be useful after reflecting negatively skewed data, but they also magnify larger values and can make outliers more influential.

6.12.6 Anscombe Transformation

For Poisson counts, the Anscombe transformation is:

$$y' = 2\sqrt{x + \frac{3}{8}} \quad (6.5)$$

It was designed to stabilise variance in count data, but a Poisson or negative-binomial GLM is usually preferable now.

6.13 WORKED EXAMPLE WITH TRANSFORMATIONS

To demonstrate what a transformation is trying to achieve, I use the built-in `airquality` dataset. Here the response is ozone concentration, and I compare **May-June** with **August-September**. This is a two-group comparison, so the relevant questions are the same ones used earlier in the chapter: are the distributions within groups reasonably well behaved, and are the group variances similar enough for a mean-based comparison?

```
ozone_dat <- na.omit(airquality[, c("Ozone", "Month")]) >
as_tibble() >
filter(Month %in% c(5, 6, 8, 9)) >
mutate(
  season = factor(
    ifelse(Month %in% c(5, 6), "May-June", "August-September")
  ),
  log_ozone = log(Ozone)
)

plt1 <- ggplot(ozone_dat, aes(Ozone)) +
  geom_histogram(bins = 20, fill = "grey70", colour = "grey30") +
  facet_wrap(~season, scales = "free_x") +
  labs(x = "Ozone", y = "Frequency")

plt2 <- ggplot(ozone_dat, aes(log_ozone)) +
  geom_histogram(bins = 20, fill = "grey70", colour = "grey30") +
  facet_wrap(~season, scales = "free_x") +
  labs(x = "log(Ozone)", y = "Frequency")

ggarrange(plt1, plt2, ncol = 2, labels = "AUTO")
```

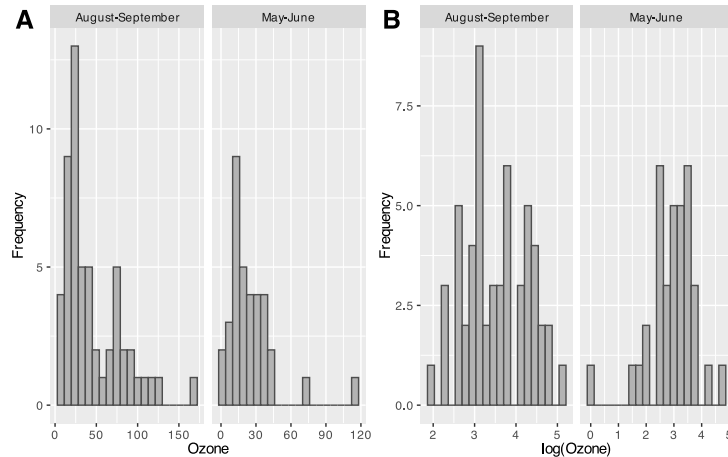


FIGURE 6.5. A real-data example showing how a log transformation changes the distribution of ozone concentrations. A) Raw ozone values by season. B) Log-transformed ozone values by season.

```
ozone_dat >
  group_by(season) >
  summarise(
    raw_w = unname(shapiro.test(Ozone)$statistic),
    raw_p = shapiro.test(Ozone)$p.value,
    log_w = unname(shapiro.test(log_ozone)$statistic),
    log_p = shapiro.test(log_ozone)$p.value,
    .groups = "drop"
  )

R> # A tibble: 2 x 5
R>   season      raw_w      raw_p log_w  log_p
R>   <fct>      <dbl>      <dbl> <dbl> <dbl>
R> 1 August-September 0.858 0.0000112 0.973 0.260
R> 2 May-June        0.765 0.0000448 0.939 0.0537

car::leveneTest(Ozone ~ season, data = ozone_dat)

R> Levene's Test for Homogeneity of Variance (center = median)
R>           Df F value Pr(>F)
R> group 1    6.3393 0.01362 *
R>      88
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

car::leveneTest(log_ozone ~ season, data = ozone_dat)

R> Levene's Test for Homogeneity of Variance (center = median)
R>           Df F value Pr(>F)
R> group 1    0.0964 0.7569
R>      88
```

Figure 6.5 shows the effect of the transformation directly. The raw ozone values show strong right-skew in both seasons. On the raw scale, both groups depart

clearly from normality and the variances differ. On the log scale, the upper tail is compressed, the group spreads are closer, and the Shapiro-Wilk and Levene results are far less problematic. The transformation has changed the scale in a way that makes a mean-based comparison more defensible.

In this chapter, I stop at the diagnostic stage since the inferential step belongs in [Chapter 7](#) where the same ozone data are analysed with a two-sample t -test on the transformed scale before the result is back-transformed for reporting.

6.14 CHECK THE ASSUMPTIONS AGAIN

The data diagnostics do not end after we have transformed the data. The transformed data must be checked again. The second pass decides whether the transformation actually improved the fit between the data and the planned method. It is annoying but necessary.

6.15 WHEN TRANSFORMATION IS NOT THE RIGHT SOLUTION

Transformation is usually the wrong response when:

- the response is fundamentally a count, proportion, presence-absence variable, or survival time with its own natural error structure;
- the observations are not independent;
- the biological interpretation becomes less clear than the original problem;
- a more appropriate method already exists.

DO IT NOW!

The `airquality` dataset (built into R) contains ozone concentrations measured in New York over several months. Run the full assumption-checking workflow on ozone (Ozone) split by Month:

- a. Produce histograms and Q-Q plots grouped by month.
- b. Apply the Shapiro-Wilk test within each month.
- c. Apply Levene's test for equal variance across months (use `car::leveneTest()`).
- d. Based on your diagnostics, would you apply a log transformation? Apply it and re-check.

6.16 SUMMARY

1. Assumptions apply to the structure implied by the analysis.
2. Group comparisons require checks within groups and across group variances.
3. Paired designs require checks on the differences.

4. Associations require checks on the joint pattern of the variables.
5. Independence comes from design, not from a formal test.
6. Transformations should be justified by the data and followed by a second round of diagnostics.

7

t-Tests

One-Sample, Two-Sample, and Paired Mean Comparisons

IN THIS CHAPTER

- One-sample t -tests
- Two-sample t -tests
- Paired t -tests
- Directional tests
- Rank-based alternatives within the t -test family

TASKS TO COMPLETE IN THIS CHAPTER

- Self-Assessment Task 7-1 (/15)
- Self-Assessment Task 7-2 (/10)
- [Self-Assessment instructions and full task overview](#)

t -tests compare means. The correct t -test is determined by the structure of the data:

- one sample and no grouping variable;
- two independent groups defined by a categorical variable with two levels;
- paired observations from the same sampling units.

The statistical question is always something like, is the estimated mean, or the estimated mean difference, large enough relative to sampling variation to justify rejecting the null hypothesis? In practice, the main quantity of interest is the estimated effect itself, together with its confidence interval. The p -value supports that interpretation and should not replace it.

Below, I follow the same order throughout: identify the sampling or experimental design, inspect the data, check the assumptions, run the test, and interpret the effect size, uncertainty, and *p*-value together. With only two groups, the two-sample *t*-test is the two-group special case of one-way ANOVA. In [Chapter 8](#), I extend the same idea to more than two groups and explain why repeated *t*-tests inflate Type I error.

7.1 DESIGN DETERMINES THE TEST

Consideration of the design comes first:

- **One-sample design:** one continuous variable, no grouping factor, and a biologically meaningful reference value μ_0 .
- **Two-sample design:** one continuous response and one categorical grouping variable with exactly two independent levels.
- **Paired design:** two measurements linked within the same sampling unit, so the analysis is done on the within-pair differences.

Checking the assumptions follows:

- The response should be continuous.
- Observations should be independent in the way implied by the sampling or experimental design.
- The relevant distribution should be reasonably well behaved.
- In the two-sample case, the spread in the two groups should be assessed before deciding whether Student's or Welch's formulation is appropriate.

Approximate normality depends on what is being analysed. In a one-sample test, it concerns the sample values. In a two-sample test, it concerns the distribution within each group. In a paired test, it concerns the paired differences. [Chapter 6](#) developed those checks directly on raw data.

Large samples make *t*-tests fairly robust to mild non-normality, especially when no single value dominates the result and the group spreads are not extreme. Independence is different. A large sample does not repair pseudoreplication, repeated use of the same individuals, or a poorly designed sampling scheme.

7.2 R FUNCTIONS

The main function in this chapter is `t.test()`. It handles:

- one-sample tests with `mu = ...`;
- two independent groups using either `x`, `y` or a formula such as `response ~ group`;
- paired tests with `paired = TRUE`.

The most useful arguments are:

- `mu` for the one-sample reference mean;
- `alternative` for "two.sided", "less", or "greater";
- `paired` for matched observations;
- `var.equal` to force Student's two-sample t -test instead of Welch's default.

When assumptions fail badly, `wilcox.test()` provides rank-based alternatives for one-sample, two-sample, and paired designs. The design still determines the correct procedure.

MATHEMATICAL DETAIL

For a one-sample t -test, the test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (7.1)$$

In Equation 7.1, $\bar{x}$ is the sample mean, μ_0 is the hypothesised mean under the null hypothesis, s is the sample standard deviation, and n is the sample size.

For Student's two-sample t -test, where a common variance is assumed:

$$t = \frac{\bar{A} - \bar{B}}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \quad (7.2)$$

with pooled variance

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \quad (7.3)$$

For Welch's two-sample t -test, which does not assume equal variances:

$$t = \frac{\bar{A} - \bar{B}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \quad (7.4)$$

with Welch-Satterthwaite degrees of freedom

$$\text{d.f.} = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{(s_A^2/n_A)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}} \quad (7.5)$$

For a paired t -test, the data are reduced to the within-pair differences and the test becomes:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (7.6)$$

The formulas explain how the statistic is constructed. In practice, R does the arithmetic and we focus on whether the design, assumptions, and interpretation are appropriate.

7.3 A SIMPLE EXPERIMENT

DO THIS NOW!

Work with two friends and devise a real experiment, or a designed sampling exercise, in which you demonstrate the following two principles:

- a. the effect of sample size on the estimate of the mean and variance (use SD),
- b. the effect of repeated sampling on the value of the mean and variance.

In both cases, apply the correct t -test from this chapter to test whether sample size and repeat sampling have a statistically significant effect. Report back comprehensively with Task F.

Many biological questions reduce to a mean comparison. A treatment is compared with a control, a measured trait is compared with a reference value, or the same individuals are measured before and after an intervention. The sampling or experimental design may be determined by scientific question, but the inference always concerns comparison of means or mean differences.

7.4 ONE-SAMPLE T-TESTS

In a one-sample design there is no grouping variable. We have one continuous sample and a reference value, μ_0 , that comes from a threshold, a historical mean, a theoretical expectation, or some other biologically meaningful reference level.

7.4.1 When to Use a One-Sample t -Test

Use a one-sample t -test when:

- you have one sample of observations;
- the response variable is continuous;
- the observations are independent;
- you want to compare the population mean with a known or hypothesised value, μ_0 ;
- the sample is reasonably well behaved, or the sample size is large enough that the test is robust to small departures from normality.

For the standard two-sided form, the hypotheses are:

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

The one-sample t -test then asks whether the observed sample mean is far enough from μ_0 , relative to sampling variation, to justify rejecting H_0 .

7.5 EXAMPLE 1: TREE HEIGHT AGAINST TWO REFERENCE MEANS

The built-in `trees` dataset records black cherry trees, with height given in feet. Here I convert height to metres and evaluate two rounded reference means, 23.0 m and 24.5 m. The example shows how the same sample can lead to different inferences depending on the reference value being tested.

```
tree_one <- as_tibble(trees) >
  transmute(height_m = Height * 0.3048, sample = "Black cherry")
```

7.5.1 Do an Exploratory Data Analysis (EDA)

This is a one-sample problem, so there is no grouping variable. I inspect the sample itself.

```
summary(tree_one$height_m)
```

```
R>   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
R>  19.20  21.95   23.16   23.16  24.38   26.52
```

```
sd(tree_one$height_m)
```

```
R> [1] 1.942129
```

```
shapiro.test(tree_one$height_m)
```

```
R>
R> Shapiro-Wilk normality test
R>
R> data:  tree_one$height_m
R> W = 0.96545, p-value = 0.4034
```

I apply the assumption check to the sample values themselves because this is a one-sample design. The Shapiro-Wilk test shows no strong evidence against normality for tree height, and the summary statistics do not suggest any extreme outlier dominating the sample. That visual comparison is shown in Figure 7.1. That visual comparison is shown in Figure 7.1.

```
ggplot(tree_one, aes(x = sample, y = height_m)) +
  geom_boxplot(fill = "indianred", alpha = 0.3, colour = "black") +
  geom_hline(yintercept = 23.0, colour = "dodgerblue2", linewidth = 0.9) +
  geom_hline(yintercept = 24.5, colour = "indianred2", linewidth = 0.9) +
  labs(x = NULL, y = "Height (m)") +
  coord_flip()
```

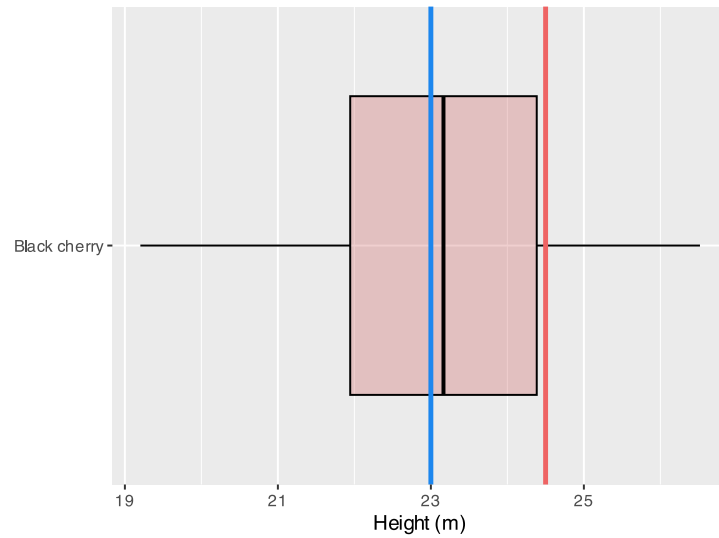


FIGURE 7.1. One-sample tree-height data in metres. The blue line marks a reference mean of 23.0 m and the red line marks 24.5 m.

The sample is centred close to 23.16 m. Visually, 23.0 m looks plausible and 24.5 m looks rather high. The assumption check supports proceeding with a one-sample t -test because the sample is reasonably well behaved and the sample size is moderate.

7.5.2 State the Hypotheses

For the reference value of 23.0 m:

$$H_0 : \mu = 23.0$$

$$H_a : \mu \neq 23.0$$

For the reference value of 24.5 m:

$$H_0 : \mu = 24.5$$

$$H_a : \mu \neq 24.5$$

7.5.3 Apply the Test

```
t_tree_75 <- t.test(tree_one$height_m, mu = 23.0)
t_tree_80 <- t.test(tree_one$height_m, mu = 24.5)

t_tree_75

R>
R> One Sample t-test
R>
R> data: tree_one$height_m
R> t = 0.47245, df = 30, p-value = 0.64
R> alternative hypothesis: true mean is not equal to 23
R> 95 percent confidence interval:
R> 22.45242 23.87718
R> sample estimates:
R> mean of x
R> 23.1648

t_tree_80

R>
R> One Sample t-test
R>
R> data: tree_one$height_m
R> t = -3.8278, df = 30, p-value = 0.0006115
R> alternative hypothesis: true mean is not equal to 24.5
R> 95 percent confidence interval:
R> 22.45242 23.87718
R> sample estimates:
R> mean of x
R> 23.1648
```

7.5.4 Interpret the Results

The mean tree height is 23.16 m. Relative to 23.0 m, that difference is small and the confidence interval includes the reference value. Relative to 24.5 m, the difference is larger and the confidence interval excludes the reference value. I therefore fail to reject H_0 when $\mu_0 = 23.0$, but reject H_0 when $\mu_0 = 24.5$.

This example shows why the null value must have biological meaning. The data have remained constant, but the inferential conclusion changes because the reference value has shifted.

7.5.5 Reporting

WRITE-UP**Methods**

A one-sample t -test was used to compare the mean height of black cherry trees with two possible reference means, 23.0 m and 24.5 m. The original data were recorded in feet and converted to metres before analysis. The distribution of tree heights was inspected before the test was applied.

Results

The sample mean tree height was 23.16 m (SD = 1.94, $n = 31$). When tested against a reference mean of 23.0 m, the one-sample t -test provided no strong evidence for a difference ($t_{30} = 0.47$, $p > 0.05$), and the 95% confidence interval for the mean (22.45 to 23.88 m) included 23.0 m. When tested against 24.5 m, the same sample differed from the reference mean ($t_{30} = -3.83$, $p < 0.001$), and the same confidence interval excluded 24.5 m.

Discussion

The inference depends on the biological relevance of the reference value. These data are compatible with a mean tree height near 23.0 m and they reject a mean as high as 24.5 m.

DO IT NOW!

The built-in `sleep` dataset (type `?sleep`) contains extra hours of sleep gained by patients under two drug treatments. Treat it as a one-sample test: for Drug 1, test whether the mean extra sleep is different from 0 (*i.e.*, whether the drug has any effect on sleep duration).

- Run the EDA: make a histogram and note whether the distribution looks roughly normal.
- State H_0 and H_a .
- Apply `t.test()` with `mu = 0`.
- Report the result in one sentence following the format from the Write-Up example above.

7.6 TWO-SAMPLE T-TESTS

In a two-sample design, the grouping variable has two levels and the observations in one group are independent of those in the other. Now I ask if the two population means are equal.

7.6.1 When to Use a Two-Sample t -Test

Use a two-sample t -test when:

- the response variable is continuous;
- the grouping variable has exactly two independent levels;
- the observations in one group are independent of those in the other;
- the within-group distributions are reasonably well behaved, or the sample sizes are large enough for mild departures from normality to be unimportant;
- the biological question concerns the mean difference between the two groups.

For the standard two-sided form, the hypotheses are:

$$H_0 : \mu_A = \mu_B$$

$$H_a : \mu_A \neq \mu_B$$

or, equivalently,

$$H_0 : \mu_A - \mu_B = 0$$

$$H_a : \mu_A - \mu_B \neq 0$$

7.6.2 *Choosing Between Student and Welch*

Welch's two-sample *t*-test should be the default. Use Student's *t*-test only when the group spreads are similar and the sample sizes are also similar enough that the pooled-variance assumption is defensible.

Use this rule:

1. start with plots and grouped summaries;
2. if the spreads are clearly different, use Welch;
3. if the spreads are similar, Welch is still acceptable and often simplest;
4. use Student only when you want the pooled-variance form for a clear reason.

Welch's test is robust, widely used, and already the default in `t.test()`. In most routine analyses there is no penalty for choosing it.

7.6.3 *Other Tests for Homogeneity*

Formal tests such as Levene's test or Bartlett's test provide supporting evidence about group spread. They should not become the main decision step. Visual comparison of spread comes first, and Welch remains the safer choice when the spreads differ.

If the apparent variance difference is tied to strong skewness or a long upper tail, a transformation may be preferable to a raw-scale *t*-test. In [Chapter 6](#) I developed those diagnostic decisions, and the ozone example below demonstrates this in practice in the final test.

7.7 EXAMPLE 2: PENGUIN FLIPPER LENGTH IN TWO SPECIES

The penguins dataset provides a realistic two-group comparison. Here I compare flipper length between Adelie and Chinstrap penguins. The grouping variable is species, and it has two independent levels in this subset.

```
peng_two <- penguins >
  select(species, flipper_length_mm) >
  drop_na() >
  filter(species %in% c("Adelie", "Chinstrap")) >
  mutate(species = factor(species, levels = c("Adelie", "Chinstrap")))
```

7.7.1 Do an Exploratory Data Analysis (EDA)

I first inspect the grouped summaries, the within-group normality checks, and the spread across groups.

```
peng_two >
  group_by(species) >
  summarise(
    n = n(),
    mean = mean(flipper_length_mm),
    sd = sd(flipper_length_mm),
    shapiro_p = shapiro.test(flipper_length_mm)$p.value,
    .groups = "drop"
  )
```

```
R> # A tibble: 2 x 5
R>   species     n mean    sd shapiro_p
R>   <fct>   <int> <dbl> <dbl>   <dbl>
R> 1 Adelie    151  190.  6.54    0.720
R> 2 Chinstrap  68   196.  7.13    0.811

car::leveneTest(flipper_length_mm ~ species, data = peng_two)
```

```
R> Levene's Test for Homogeneity of Variance (center = median)
R>      Df F value Pr(>F)
R> group  1  0.6238 0.4305
R>      217
```

I apply the assumption checks within each species because the two-sample design compares group means. The Shapiro-Wilk tests show no strong evidence against normality within either species, and Levene's test provides no strong evidence that the group spreads differ. Together with the boxplots in Figure 7.2, these checks support a standard two-group mean comparison.

```
ggplot(peng_two, aes(x = species, y = flipper_length_mm, fill = species)) +
  geom_boxplot(alpha = 0.35, colour = "black", outlier.shape = NA) +
  geom_jitter(width = 0.08, alpha = 0.6, size = 1.5) +
  labs(x = "Species", y = "Flipper length (mm)") +
  theme(legend.position = "none")
```

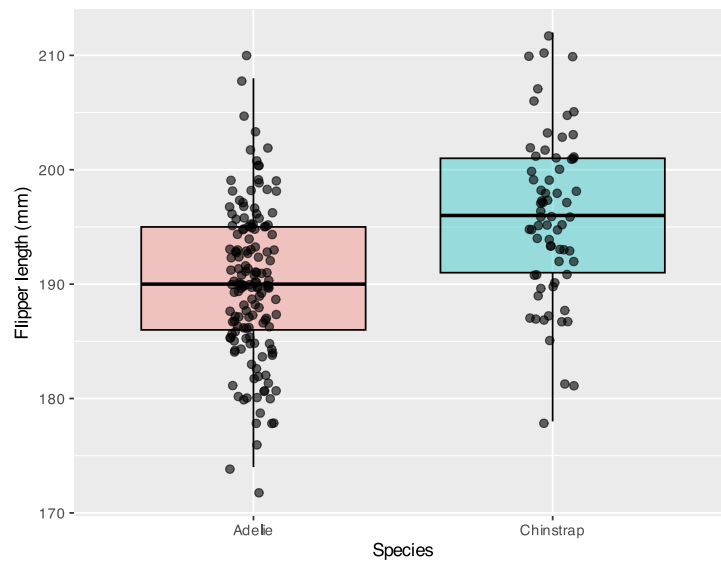


FIGURE 7.2. Flipper length in Adelie and Chinstrap penguins. The group centres differ, while the group spreads are similar.

The boxplots show a clear shift in centre, with Chinstrap penguins tending to have longer flippers. The spreads are similar. The implication is that the assumptions do not force a change of method here, so a Welch two-sample t -test is a defensible default.

7.7.2 State the Hypotheses

$$H_0 : \mu_{\text{Adelie}} = \mu_{\text{Chinstrap}}$$

$$H_a : \mu_{\text{Adelie}} \neq \mu_{\text{Chinstrap}}$$

7.7.3 Apply the Test

```
peng_welch <- t.test(flipper_length_mm ~ species, data = peng_two)
peng_welch
```

```
R>
R> Welch Two Sample t-test
R>
R> data: flipper_length_mm by species
R> t = -5.7804, df = 119.68, p-value = 6.049e-08
```

```

R> alternative hypothesis: true difference in means between group Adelie and group
Chinstrap is not equal to 0
R> 95 percent confidence interval:
R> -7.880530 -3.859244
R> sample estimates:
R> mean in group Adelie mean in group Chinstrap
R> 189.9536 195.8235

```

The formula `flipper_length_mm ~ species` tells R to compare the flipper lengths across the levels of the grouping variable `species`. This same formula notation appears again in ANOVA and regression chapters.

7.7.4 Interpret the Results

The estimated difference in mean flipper length is about 5.9 mm, with Chinstrap penguins having the larger mean. The 95% confidence interval excludes zero, and the p -value is very small. We therefore reject H_0 and conclude that the two species differ in mean flipper length.

The confidence interval shows the plausible range for the size of the difference, which is the biologically useful part of the result.

7.7.5 Reporting

WRITE-UP

Methods

Mean flipper length was compared between Adelie and Chinstrap penguins using a two-sample Welch t -test. Species was treated as a two-level grouping factor, and flipper length was the continuous response variable. Grouped summaries and graphical checks were used to inspect the data before the test was applied.

Results

Adelie penguins had a mean flipper length of 190.0 mm (SD = 6.54, $n = 151$), whereas Chinstrap penguins had a mean flipper length of 195.8 mm (SD = 7.13, $n = 68$). A two-sample Welch t -test indicated that mean flipper length differed between species ($t_{119.68} = -5.78$, $p < 0.001$). The estimated mean difference was 5.87 mm, and the 95% confidence interval ranged from 3.86 to 7.88 mm, with Chinstrap penguins having the larger mean.

Discussion

The estimated difference is several millimetres and the confidence interval remains well away from zero, so the species difference is biologically substantial as well. This is the two-group case that Chapter 8 generalises to more than two groups.

DO THIS NOW!

Create a visualisation that makes the mean difference and the spread difference in this penguin example easy to compare at a glance.

SELF-ASSESSMENT TASK 7-1

Please refer to this two-sided two-sample t -test:

```
# random normal data
set.seed(666)
r_two <- data.frame(dat = c(rnorm(n = 20, mean = 4, sd = 1),
                           rnorm(n = 20, mean = 5, sd = 1)),
                   sample = c(rep("A", 20), rep("B", 20)))

# perform t-test
# note how we set the `var.equal` argument to TRUE because we know
# our data has the same SD (they are simulated as such!)
t.test(dat ~ sample, data = r_two, var.equal = TRUE)

R>
R> Two Sample t-test
R>
R> data: dat by sample
R> t = -1.9544, df = 38, p-value = 0.05805
R> alternative hypothesis: true difference in means between group A and group B is
not equal to 0
R> 95 percent confidence interval:
R> -1.51699175 0.02670136
R> sample estimates:
R> mean in group A mean in group B
R> 4.001438 4.746584
```

- Repeat this analyses using the Welch's `t.test()`. (/5)
- Repeat your analysis, above, using the even more old-fashioned Equation 4 in the lecture. Show the code, talk us through the step you followed to read the p -values off the table of t -statistics. (/10)

7.8 EXAMPLE 3: A TWO-SAMPLE COMPARISON AFTER LOG TRANSFORMATION

The ozone example from Chapter 6 can now be carried through to a t -test for inference. I compare ozone concentrations in **May-June** with those in **August-September**. In the `airquality` dataset, ozone is recorded in parts per billion (ppb). On the raw scale, the data are strongly right-skewed and the group variances differ. I apply a log transformation, which improves both features. So the two-sample comparison is done on the log scale and then back-transformed for reporting.

```
aq <- na.omit(airquality[, c("Ozone", "Month")]) >
as_tibble() >
filter(Month %in% c(5, 6, 8, 9)) >
mutate(
  season = factor(
```

```

    if_else(Month %in% c(5, 6), "May-June", "August-September")
  ),
  log_ozone = log(Ozone)
)

```

7.8.1 Do an Exploratory Data Analysis (EDA)

First, I inspect the grouped distributions on the raw and transformed scales in Figure 7.3.

```

plt_raw <- ggplot(aq, aes(season, Ozone, fill = season)) +
  geom_boxplot(alpha = 0.35, colour = "black", outlier.shape = NA) +
  geom_jitter(width = 0.08, alpha = 0.6, size = 1.5) +
  labs(x = NULL, y = "Ozone (ppb)") +
  theme(legend.position = "none")

plt_log <- ggplot(aq, aes(season, log_ozone, fill = season)) +
  geom_boxplot(alpha = 0.35, colour = "black", outlier.shape = NA) +
  geom_jitter(width = 0.08, alpha = 0.6, size = 1.5) +
  labs(x = NULL, y = "log(Ozone)") +
  theme(legend.position = "none")

ggarrange(plt_raw, plt_log, ncol = 2)

```

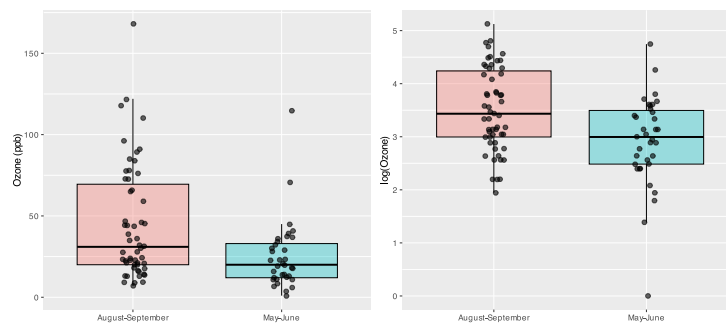


FIGURE 7.3. Ozone concentrations before and after log transformation. A) Raw ozone values (ppb) by season. B) Log-transformed ozone values by season.

My formal checks support the same conclusion:

```

aq >
  group_by(season) >
  summarise(
    raw_w = unname(shapiro.test(Ozone)$statistic),
    raw_p = shapiro.test(Ozone)$p.value,
    log_w = unname(shapiro.test(log_ozone)$statistic),
    log_p = shapiro.test(log_ozone)$p.value,
    raw_mean = mean(Ozone),
    geo_mean = exp(mean(log_ozone)),
    .groups = "drop"
  )

```

```

R> # A tibble: 2 x 7
R>   season      raw_w      raw_p log_w  log_p raw_mean geo_mean
R>   <fct>      <dbl>      <dbl> <dbl> <dbl> <dbl>    <dbl>
R> 1 August-September 0.858 0.0000112 0.973 0.260    44.9    33.6
R> 2 May-June        0.765 0.00000448 0.939 0.0537    25.1    18.6
R> car::leveneTest(Ozone ~ season, data = aq)

R> Levene's Test for Homogeneity of Variance (center = median)
R>      Df F value Pr(>F)
R> group 1  6.3393 0.01362 *
R>      88
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> car::leveneTest(log_ozone ~ season, data = aq)

R> Levene's Test for Homogeneity of Variance (center = median)
R>      Df F value Pr(>F)
R> group 1  0.0964 0.7569
R>      88

```

I apply the assumption checks within season on both the raw and transformed scales because this is still a two-group comparison. On the raw scale, both groups show strong evidence against normality and the variances differ. On the log scale, the group distributions are much closer to normal and the variance difference largely disappears. The implication is that the raw-scale comparison is a poor choice, while the transformed comparison is defensible.

7.8.2 State the Hypotheses

Because the analysis is now on the log scale, the hypotheses concern the mean log ozone concentration:

$$H_0 : \mu_{\log, \text{May-June}} = \mu_{\log, \text{August-September}}$$

$$H_a : \mu_{\log, \text{May-June}} \neq \mu_{\log, \text{August-September}}$$

Back-transformation changes the way the result is expressed. Equality on the log scale corresponds to a ratio of 1 on the original scale, so the null hypothesis becomes:

$$H_0 : \frac{\text{geometric mean ozone in August-September}}{\text{geometric mean ozone in May-June}} = 1$$

The alternative hypothesis is that this ratio is not 1. A ratio of 1 means that the two periods have the same geometric mean ozone concentration in ppb. A ratio greater than 1 means that ozone is higher in August-September. A ratio less than 1 means that ozone is lower in August-September.

7.8.3 Apply the Test

Welch's two-sample *t*-test remains the default choice in R and is appropriate here.

```
ozone_t <- t.test(log_ozone ~ season, data = aq)
ozone_t

R>
R> Welch Two Sample t-test
R>
R> data: log_ozone by season
R> t = 3.3038, df = 67.271, p-value = 0.001531
R> alternative hypothesis: true difference in means between group August-September and
group May-June is not equal to 0
R> 95 percent confidence interval:
R> 0.2350506 0.9523976
R> sample estimates:
R> mean in group August-September      mean in group May-June
R>          3.514982                  2.921258

aq >
  group_by(season) >
    summarise(
      geometric_mean = exp(mean(log_ozone)),
      .groups = "drop"
    )

R> # A tibble: 2 x 2
R>   season      geometric_mean
R>   <fct>          <dbl>
R> 1 August-September      33.6
R> 2 May-June              18.6

ratio_est <- exp(ozone_t$estimate[[1]] - ozone_t$estimate[[2]])
ratio_ci <- exp(ozone_t$conf.int)

tibble(
  comparison = "August-September / May-June",
  ratio = ratio_est,
  lower = ratio_ci[1],
  upper = ratio_ci[2]
)

R> # A tibble: 1 x 4
R>   comparison      ratio lower upper
R>   <chr>          <dbl> <dbl> <dbl>
R> 1 August-September / May-June  1.81  1.26  2.59
```

7.8.4 Interpret the Results

The test on the log scale compares mean log ozone concentrations between the two periods. After back-transformation, the result is easier to interpret as a ratio. August-September ozone concentrations are estimated to be about 1.81 times those in May-June, with a 95% confidence interval from 1.26 to 2.59.

The geometric mean is the back-transformed mean from the log scale. It differs from the arithmetic mean in that it summarises multiplicative variation rather than additive variation. The arithmetic mean is the ordinary average obtained by adding

values and dividing by the sample size. The geometric mean is obtained by averaging the logged values and then back-transforming. For right-skewed data, it is often less influenced by the upper tail than the arithmetic mean.

In this example, I show the complete sequence of decision steps. I start with diagnosing the assumption violation on the raw scale, then I transform the data for a clear reason, I re-assess the assumptions, apply the test, and then I return the result to the original scale for reporting.

7.8.5 Reporting

WRITE-UP

Methods

Ozone concentrations (ppb) were compared between May-June and August-September using a two-sample t -test. Exploratory plots and assumption checks showed strong right-skew and unequal variances on the raw scale, so ozone was log-transformed before analysis. A Welch two-sample t -test was then applied to log-transformed ozone concentrations. Results were back-transformed and reported as geometric means and their ratio.

Results

On the transformed scale, ozone concentrations differed between the two periods (Welch two-sample t -test: $t_{67.27} = 3.30$, $p < 0.01$). The geometric mean ozone concentration was 18.6 ppb in May-June and 33.6 ppb in August-September. After back-transformation, August-September ozone concentrations were estimated to be 1.81 times those in May-June (95% CI 1.26 to 2.59).

Discussion

The transformed analysis supports higher ozone concentrations in August-September than in May-June. Because the model was fitted on the log scale, the clearest interpretation is multiplicative: the later period has ozone concentrations about 1.8 times higher.

7.9 PAIRED T-TESTS

In a paired design, the observations are linked. The pairing may come from repeated measurements on the same individuals, left-right comparisons on the same organism, or deliberately matched experimental units. The analysis is done on the within-pair differences.

7.9.1 When to Use a Paired t-Test

Use a paired t -test when:

- the response variable is continuous;
- the two measurements are meaningfully paired within the same sampling unit;
- the pairs are independent of one another;
- the paired differences are reasonably well behaved.

For the standard two-sided form, the hypotheses are:

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

where μ_d is the population mean of the paired differences.

7.10 EXAMPLE 4: BEFORE-AND-AFTER MEASUREMENTS ON PIGLETS

I will use a small example in which piglets are weighed before and after a three-week feeding period. The inferential question is whether body mass increases over time within the same individuals.

```
set.seed(666)
piglets <- data.frame(
  id = 1:12,
  before = round(rnorm(12, mean = 8.2, sd = 0.7), 2),
  after = round(rnorm(12, mean = 9.0, sd = 0.7), 2)
)

piglets$diff <- piglets$after - piglets$before
piglets
```

```
R>   id before after  diff
R> 1   1  8.73  9.61  0.88
R> 2   2  9.61  7.80 -1.81
R> 3   3  7.95  9.09  1.14
R> 4   4  9.62  8.95 -0.67
R> 5   5  6.65  9.60  2.95
R> 6   6  8.73  9.24  0.51
R> 7   7  7.29  8.59  1.30
R> 8   8  7.64  9.55  1.91
R> 9   9  6.95  8.52  1.57
R> 10 10  8.17  8.17  0.00
R> 11 11  9.71  9.89  0.18
R> 12 12  6.96  8.78  1.82
```

7.10.1 Do an Exploratory Data Analysis (EDA)

Before running the paired t -test, I inspect the paired differences because these are the values used directly by the test.

```
summary(piglets$diff)
```

```
R>   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
R> -1.810  0.135   1.010   0.815  1.633   2.950
```

```
sd(piglets$diff)
```

```
R> [1] 1.277024
```

```
shapiro.test(piglets$diff)
```

```
R>
R> Shapiro-Wilk normality test
R>
R> data:  piglets$diff
R> W = 0.9798, p-value = 0.9828
```

I apply the assumption check to the paired differences not to the raw before and after measurements. The Shapiro-Wilk test shows no strong evidence against normality in those differences, so I can use the paired t -test without changing method. The distribution of differences is shown in Figure 7.4. The distribution of differences is shown in Figure 7.4.

```
ggplot(data = piglets, aes(x = diff)) +
  geom_histogram(binwidth = 0.5, fill = "indianred", colour = "black", alpha = 0.5) +
  geom_vline(xintercept = 0, colour = "dodgerblue2", linewidth = 0.9) +
  labs(x = "Paired difference in mass (after - before)", y = "Frequency")
```

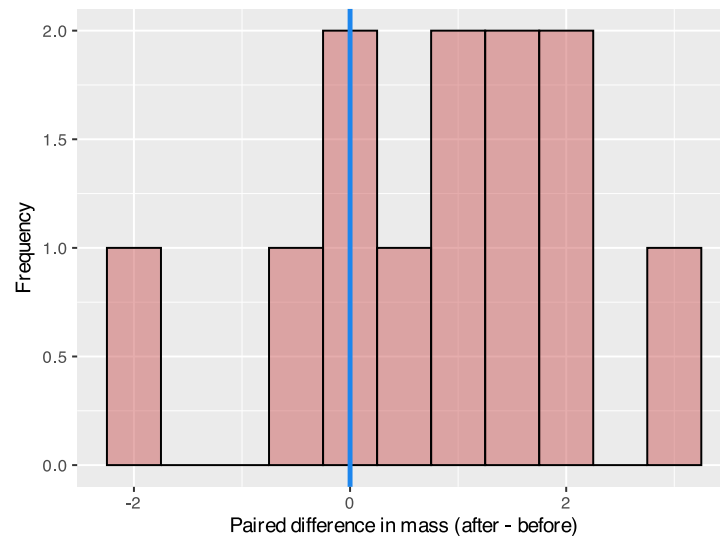


FIGURE 7.4. Distribution of paired differences in piglet body mass after minus before the feeding period.

The paired differences are centred above zero, which already suggests that the piglets tended to gain mass over the feeding period. The distribution is well behaved enough for a paired t -test, so the next step is to test whether the mean difference is different from zero.

7.10.2 State the Hypotheses

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

If the biological question had specified direction in advance, such as whether mass increased after feeding, the same paired design could be analysed with a one-sided alternative.

7.10.3 Apply the Test

```
t.test(piglets$after, piglets$before, paired = TRUE)

R>
R> Paired t-test
R>
R> data: piglets$after and piglets$before
R> t = 2.2108, df = 11, p-value = 0.04915
R> alternative hypothesis: true mean difference is not equal to 0
R> 95 percent confidence interval:
R>  0.003617508 1.626382492
R> sample estimates:
R> mean difference
R>      0.815
```

The paired t -test is a one-sample test applied to the set of paired differences. That is why the raw before and after values are not analysed as though they were independent groups.

7.10.4 Interpret the Results

The mean paired difference is positive, so the piglets tended to be heavier after the feeding period. The confidence interval for the mean difference lies just above zero and the p -value is slightly below 0.05. We therefore reject H_0 and conclude that body mass increased over the feeding period in this example.

7.10.5 Reporting

WRITE-UP**Methods**

Piglet body mass was measured before and after a three-week feeding period, and the resulting paired observations were analysed with a paired t -test. The analysis was based on the within-individual differences in mass.

Results

Piglet body mass increased over the three-week feeding period. A paired t -test showed that the mean within-individual change in body mass differed from zero ($t_{11} = 2.21$, $p < 0.05$). The mean paired difference was 0.815 kg, with a 95% confidence interval from 0.004 to 1.626 kg.

Discussion

The paired design strengthens the inference because each piglet acts as its own control. The biological conclusion therefore concerns change within individuals rather than a difference between independent groups.

7.11 WHEN ASSUMPTIONS ARE DIFFICULT

The response to assumption problems depends on the sampling and experimental design and on the source of the problem.

Situation	Typical response
One sample, small n , strong skew or extreme outliers	Consider transformation or a one-sample Wilcoxon procedure
Two groups, similar shape but different spreads	Use Welch's t -test
Two groups, skewness linked to spread	Transform, re-check, then analyse if the new scale is defensible
Two groups, strong outliers or badly distorted shapes	Consider a Wilcoxon rank-sum procedure
Paired data, non-normal paired differences	Consider a paired Wilcoxon signed-rank procedure

Use Welch when the main problem is unequal spread. Use a transformation when the problem is scale, such as a long right tail that also drives a variance difference. Use a rank-based alternative when the mean-based comparison itself is no longer a good description of the biological question.

Keep the following rules in mind:

- do not switch to Student's t -test merely because Levene's test is non-significant;
- do not treat a formal normality test as the only decision criterion;

- do not transform data automatically; transform for a stated reason and then re-check;
- do not forget that a large sample helps with mild non-normality but does not fix non-independence.

7.12 OPTIONAL: ONE-SIDED T-TESTS

One-sided tests should be used only when the direction of the effect is specified before the data are analysed. If the real question is whether two means differ, use the standard two-sided form.

7.12.1 *One-Sided One-Sample t-Test*

Suppose the forestry question is directional: are the sampled trees shorter than 24.5 m on average?

The hypotheses are:

$$H_0 : \mu \geq 24.5$$

$$H_a : \mu < 24.5$$

```
t.test(tree_one$height_m, mu = 24.5, alternative = "less")
```

```
R>
R> One Sample t-test
R>
R> data: tree_one$height_m
R> t = -3.8278, df = 30, p-value = 0.0003058
R> alternative hypothesis: true mean is less than 24.5
R> 95 percent confidence interval:
R>      -Inf 23.75683
R> sample estimates:
R> mean of x
R>  23.1648
```

```
t.test(tree_one$height_m, mu = 24.5, alternative = "greater")
```

```
R>
R> One Sample t-test
R>
R> data: tree_one$height_m
R> t = -3.8278, df = 30, p-value = 0.9997
R> alternative hypothesis: true mean is greater than 24.5
R> 95 percent confidence interval:
R>  22.57277      Inf
R> sample estimates:
R> mean of x
R>  23.1648
```

The first test matches the direction implied by the data and the question, but the second does not. Direction changes the hypothesis, the tail area, and the interpretation.

7.12.2 One-Sided Two-Sample t -Test

Suppose the biological question is directional: are Adelie penguin flippers shorter than Chinstrap penguin flippers?

The hypotheses are:

$$H_0 : \mu_{\text{Adelie}} \geq \mu_{\text{Chinstrap}}$$

$$H_a : \mu_{\text{Adelie}} < \mu_{\text{Chinstrap}}$$

```
t.test(flipper_length_mm ~ species, data = peng_two, alternative = "less")
R>
R> Welch Two Sample t-test
R>
R> data: flipper_length_mm by species
R> t = -5.7804, df = 119.68, p-value = 3.025e-08
R> alternative hypothesis: true difference in means between group Adelie and group
Chinstrap is less than 0
R> 95 percent confidence interval:
R>      -Inf -4.186534
R> sample estimates:
R> mean in group Adelie mean in group Chinstrap
R>      189.9536      195.8235

t.test(flipper_length_mm ~ species, data = peng_two, alternative = "greater")
R>
R> Welch Two Sample t-test
R>
R> data: flipper_length_mm by species
R> t = -5.7804, df = 119.68, p-value = 1
R> alternative hypothesis: true difference in means between group Adelie and group
Chinstrap is greater than 0
R> 95 percent confidence interval:
R> -7.55324      Inf
R> sample estimates:
R> mean in group Adelie mean in group Chinstrap
R>      189.9536      195.8235
```

Again, the test only makes sense when the direction was specified before the analysis. One-sided tests are not a rescue strategy for a non-significant two-sided result.

7.13 RANK-BASED ALTERNATIVES WITHIN THE T-TEST FAMILY

The parametric t -tests in this chapter are mean-based procedures. When the response is badly non-normal in a small sample, strongly skewed, or dominated by influential

outliers, a rank-based alternative may be more defensible. These methods belong to the same inferential family as the t -tests because they serve the same one-sample, two-sample, or paired designs.

The three main rank-based counterparts are:

- **one-sample Wilcoxon signed-rank test** for a single sample compared with a reference value;
- **Wilcoxon rank-sum test** for two independent groups;
- **paired Wilcoxon signed-rank test** for matched or before-after observations.

All three are handled by `wilcox.test()` in R:

- `wilcox.test(x, mu = ...)` for a one-sample problem;
- `wilcox.test(x, y)` or `wilcox.test(response ~ group, data = ...)` for two independent groups;
- `wilcox.test(x, y, paired = TRUE)` for paired data.

These procedures are often introduced as alternatives when the mean is a poor summary of the data. They should not be described as direct tests of medians in every case. Their interpretation depends on the design and on the structure of the distributions being compared.

USE THE DESIGN, NOT JUST THE FUNCTION NAME

The one-sample and paired Wilcoxon procedures are often confused because they use the same function. The distinction is exactly the same as for t -tests:

- one sample against a reference value → one-sample signed-rank test;
- two related measurements on the same units → paired signed-rank test;
- two independent groups → rank-sum test.

7.14 DECISION STEPS FOR ANY T-TEST

For most problems in this chapter, the workflow is:

1. identify the design: one sample, two independent groups, or paired data;
2. visualise the data;
3. check the relevant assumptions for shape, spread, and independence;
4. choose the test, with Welch as the default for two groups;
5. run the test;
6. interpret the estimated effect, its confidence interval, and the p -value together.

If you follow these steps you'll prevent a large number of common mistakes.

DO IT NOW!

Work through the six decision steps above using the built-in `mtcars` dataset. Compare fuel economy (`mpg`) between cars with automatic (`am = 0`) and manual (`am = 1`) transmissions.

Step 1: What is the design?

Step 2: Make a grouped box plot of `mpg ~ factor(am)`.

Step 3: Apply Shapiro-Wilk within each group and Levene's test across groups.

Step 4: Choose Welch's or Student's *t*-test based on step 3.

Step 5: Run `t.test(mpg ~ factor(am), data = mtcars, var.equal = ...)`.

Step 6: Write a one-sentence results statement with the test statistic, *df*, *p*-value, and 95% CI.

7.15 A T-TEST WORKFLOW

Now that the individual cases are in place, I can walk through a complete analysis in R. For this example we use the `ecklonia` data from [Intro R Workshop: Data Manipulation, Analysis, Graphing](#).

7.15.1 Loading Data

Before I analyse anything, I need to load the data and reshape them into long format.

```
ecklonia <- read_csv(here::here("data", "BCB744", "ecklonia.csv")) >
gather(key = "variable", value = "value", -species, -site, -ID)
```

7.15.2 Visualising Data

I visualise the data first in Figure 7.5.

```
ggplot(data = ecklonia, aes(x = variable, y = value, fill = site)) +
  geom_boxplot(colour = "black", fill = "dodgerblue4", alpha = 0.4) +
  coord_flip()
```

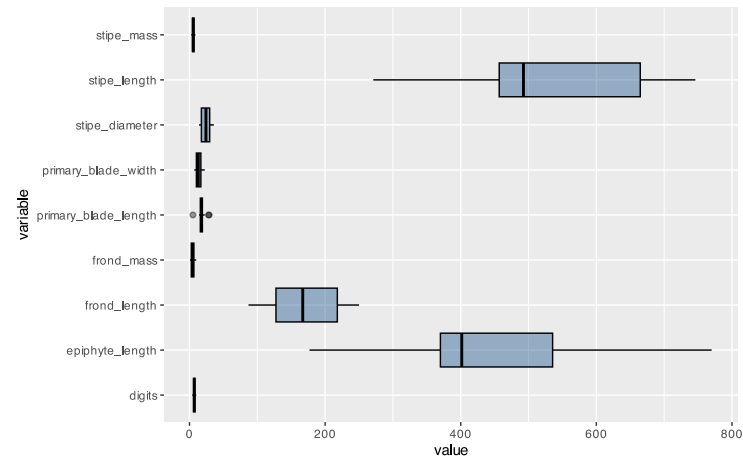


FIGURE 7.5. Boxplots showing differences in morphometric properties of the kelp *Ecklonia maxima* at two sites in False Bay.

The measurements occupy very different scales, which makes the full figure difficult to compare directly between sites. But, nevertheless, the figure's value is orientating me to the variables in the dataset and showing me that a more focused comparison is needed before I can state a hypothesis clearly.

7.15.3 Formulating a Hypothesis

Let me narrow the question to stipe mass and focus only on two sites, Batsata Rock and Boulders Beach, as shown in Figure 7.6.

```
ecklonia_sub <- ecklonia >
  filter(variable == "stipe_mass")

ggplot(data = ecklonia_sub, aes(x = variable, y = value, fill = site)) +
  geom_boxplot(colour = "black", alpha = 0.4) +
  coord_flip() +
  labs(y = "Stipe mass (kg)", x = "") +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```

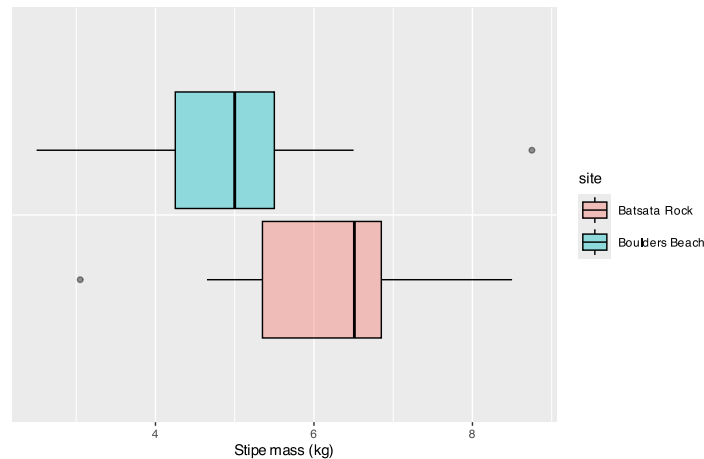


FIGURE 7.6. Boxplots showing the difference in stipe mass (kg) of the kelp *Ecklonia maxima* at two sites in False Bay.

I can state the directional question as: are kelp stipes heavier at Batsata Rock than at Boulders Beach?

The corresponding hypotheses are:

$$H_0 : \mu_{\text{Batsata}} \leq \mu_{\text{Boulders}}$$

$$H_a : \mu_{\text{Batsata}} > \mu_{\text{Boulders}}$$

7.15.4 Choosing a Test

There are two independent groups, one continuous response, and a directional hypothesis specified in advance. That combination points to a one-sided two-sample *t*-test. Because the group spreads appear similar, Student's formulation is acceptable here, although Welch would also be defensible.

7.15.5 Checking Assumptions

The response is continuous and the sites define the two independent groups. I still need to inspect spread and approximate normality within groups.

```
ecklonia_sub >
  group_by(site) >
  summarise(
    stipe_mass_var = var(value),
```

```

    stipe_mass_norm = shapiro.test(value)$p.value,
    .groups = "drop"
)

```

```

R> # A tibble: 2 x 3
R>   site      stipe_mass_var stipe_mass_norm
R>   <chr>      <dbl>      <dbl>
R> 1 Batsata Rock      2.00      0.813
R> 2 Boulders Beach    2.64      0.527

```

The grouped summaries suggest similar spread and no obvious normality problem. If these checks had failed badly, [Chapter 6](#) provides the next steps.

7.15.6 Running an Analysis

```

t.test(value ~ site, data = ecklonia_sub,
       var.equal = TRUE, alternative = "greater")
R>
R> Two Sample t-test
R>
R> data: value by site
R> t = 1.8741, df = 24, p-value = 0.03657
R> alternative hypothesis: true difference in means between group Batsata Rock and group
Boulders Beach is greater than 0
R> 95 percent confidence interval:
R>  0.09752735      Inf
R> sample estimates:
R> mean in group Batsata Rock mean in group Boulders Beach
R>      6.116154      4.996154

```

7.15.7 Interpreting the Results

The output gives the test statistic, degrees of freedom, p -value, and confidence interval for the mean difference. If the one-sided p -value is small enough, I reject H_0 and conclude that mean stipe mass is greater at Batsata Rock. The confidence interval gives the plausible size of that difference.

7.15.8 Drawing Conclusions

The scientific conclusion should consider what local ecological or hydrodynamic conditions could account for the difference in stipe mass between the two sites.

7.15.9 Reporting

WRITE-UP**Methods**

Stipe mass of *Ecklonia maxima* was compared between Batsata Rock and Boulders Beach using a one-sided two-sample Student's t -test. The analysis treated site as the grouping factor and stipe mass as the continuous response variable, after checking that the group spreads were similar and that the data were approximately normal within each site.

Results

Mean stipe mass of *Ecklonia maxima* was higher at Batsata Rock (6.12 ± 1.41 kg SD, $n = 13$) than at Boulders Beach (5.00 ± 1.63 kg SD, $n = 13$). A one-sided two-sample Student's t -test indicated that stipe mass at Batsata Rock was greater than at Boulders Beach ($t_{24} = 1.87$, $p < 0.05$). The one-sided 95% confidence interval for the mean difference ranged from 0.10 kg to ∞ .

Discussion

The biological conclusion is that kelp at Batsata Rock had heavier stipes than kelp at Boulders Beach in this comparison. The next step would be to explain why that difference exists.

7.15.10 *Going Further*

In this example, I isolated two sites so that the comparison could fit a two-sample t -test. The full dataset contains more than two sites, however, and that changes the analysis. Once the question becomes whether mean stipe mass differs across several (i.e., > 2) sites, the correct next step is ANOVA rather than a bunch of pairwise t -tests. In [Chapter 8](#), I develop that extension and explain why repeated two-group testing inflates Type I error.

The biological questions also inflates at that point. If stipe mass differs among several sites, what combination of exposure, nutrient regime, competition, or age structure produced that pattern?

SELF-ASSESSMENT TASK 7-2

For each scenario below, identify the correct t -test design (one-sample, two-sample, or paired) and state which assumption needs to be checked, and where:

- You measure the resting heart rates of 25 athletes and compare them with the published population mean of 72 bpm. (/2)
- You compare the nitrogen content of leaves from trees in two adjacent forest types (20 trees per forest). (/2)
- You weigh 15 juvenile tortoises in March and again in October and ask whether they gained mass. (/2)
- You measure the length of left and right tibiae from 18 rabbits. (/2)

Discuss with a partner whether (b) and (d) are structurally similar or different. What changes between them? (/2)

7.16 CONCLUSION

t -tests are among the first inferential tools used in biostatistics because they connect directly to common biological questions. The main decision is design: one sample, two independent groups, or paired observations. From there, the workflow simply requires that you inspect the data, check the assumptions, choose the appropriate test, and interpret the estimated effect with its uncertainty.

8

ANOVA

Comparing the Means of More than Two Groups

“He uses statistics as a drunken man uses lamp posts — for support rather than for illumination.”

— Marissa Mayer

IN THIS CHAPTER

- one-way ANOVA
- Tukey HSD
- factorial ANOVA
- Welch’s one-way ANOVA
- Kruskal-Wallis rank-sum test

CHEATSHEET

Find here a [Cheatsheet](#) on statistical methods.

TASKS TO COMPLETE IN THIS CHAPTER

- Self-Assessment Task 8-1 (/8)
- Self-Assessment Task 8-2 (/6)
- Self-Assessment Task 8-3 (/16)
- Self-Assessment Task 8-4 (/5)
- Self-Assessment Task 8-5 (/5)
- Self-Assessment Task 8-6 (/5)
- [Self-Assessment instructions and full task overview](#)

```
chicks <- as_tibble(ChickWeight)
chicks_day21 <- filter(chicks, Time == 21)
```

DO IT NOW!

As a reminder of what *t*-tests are about, use your BCB742 Bokbaai data, and devise a research question that requires the application of a *t*-test. Work through the entire *t*-test workflow, and produce the Methods, Results, and Discussion as an output (together with the code that enabled your analysis).

ANOVA extends the comparison-of-means idea from [Chapter 7](#) to more than two groups. The question changes from “do these two means differ?” to “do these group means all come from the same population mean?” The response variable is still continuous and the predictor still categorical, but now we have more than two groups.

Although ANOVA is also a linear model with categorical predictors, this chapter begins with grouped-data displays because they are intuitive and useful for exploratory analyses. Formally, however, the assumptions of ANOVA belong to the model errors and are therefore checked most directly through residuals from the fitted ANOVA model. The later regression chapters, especially [Chapter 11](#), [Chapter 12](#), and [Chapter 14](#), make that linear-model connection clear.

8.1 WHY NOT DO MULTIPLE T-TESTS?

The first impulse with several groups is often to compare every pair with a separate *t*-test. That is the wrong decision. Each additional test creates another opportunity for a false positive, so the overall Type I error rate rises as the number of pairwise comparisons increases.

In the chicks data we have four diets. Pairwise testing would require six null hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 = \mu_3$$

$$H_0 : \mu_1 = \mu_4$$

$$H_0 : \mu_2 = \mu_3$$

$$H_0 : \mu_2 = \mu_4$$

$$H_0 : \mu_3 = \mu_4$$

TABLE 8.1. Probability of at least one Type I error when all pairwise t -tests are run among K groups.

K	0.2	0.1	0.05	0.02	0.01	0.001
2	0.20	0.10	0.05	0.02	0.01	0.00
3	0.49	0.27	0.14	0.06	0.03	0.00
4	0.74	0.47	0.26	0.11	0.06	0.01
5	0.89	0.65	0.40	0.18	0.10	0.01
10	1.00	0.99	0.90	0.60	0.36	0.04
20	1.00	1.00	1.00	0.98	0.85	0.17
100	1.00	1.00	1.00	1.00	1.00	0.99

With each repeated test, the probability of rejecting at least one true null hypothesis increases.

ANOVA makes this a global question:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

$$H_a : \text{not all group means are equal}$$

If the global null hypothesis is rejected, we then ask which groups differ using a post hoc procedure such as Tukey's HSD.

BONFERRONI CORRECTION

Should you insist on doing multiple t -tests, Bonferroni correction divides the significance threshold by the number of tests. It reduces the false-positive rate, but it can become very conservative. The better solution, when the scientific question concerns several group means at once, is to fit the appropriate multi-group model, such as an ANOVA, rather than correct a stack of pairwise t -tests after the fact.

SELF-ASSESSMENT TASK 8-1

- Why should we not just apply t -tests once per each of the pairs of comparisons we want to make? (/3)
- Using the table above, calculate the probability of at least one false positive when comparing five groups ($K = 5$) at $\alpha = 0.05$ using pairwise t -tests. How many comparisons would that involve? Then calculate the same probability at $\alpha = 0.01$. Is a Bonferroni correction or a single ANOVA the better solution? Discuss with a partner and be ready to explain your reasoning. (/5)

8.2 HOW ANOVA PARTITIONS VARIATION

ANOVA gets to the global question by partitioning the total variation in the response into two components:

- variation among the group means;
- variation within the groups.

If the group means differ more than we would expect from within-group variation alone, the ANOVA result becomes large enough to reject H_0 . This partition is illustrated in Figure 8.1. This partition is illustrated in Figure 8.1.

```
grand_mean <- mean(chicks_day21$weight)
ggplot(chicks_day21, aes(Diet, weight)) +
  geom_jitter(width = 0.08, alpha = 0.5, size = 1.5, colour = "grey40") +
  stat_summary(fun = mean, geom = "point", size = 2.8, colour = "red4") +
  geom_hline(yintercept = grand_mean, colour = "dodgerblue3", linetype = 2, linewidth = 0.7) +
  labs(x = "Diet", y = "Chick mass (g)")
```

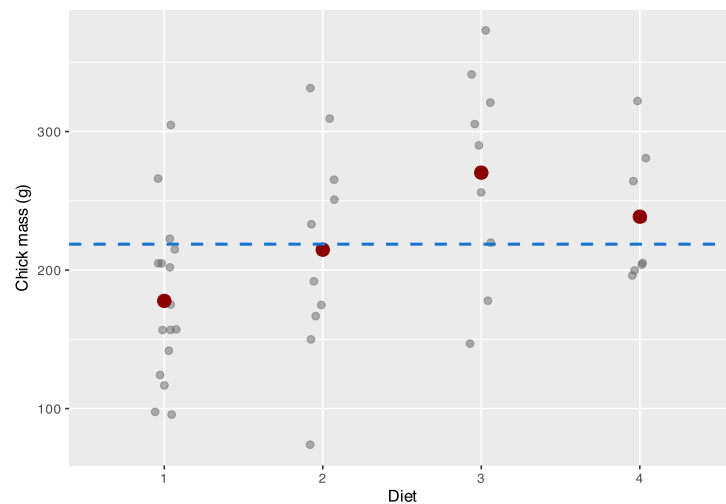


FIGURE 8.1. Visual intuition for ANOVA. The dashed blue line is the grand mean. The red points are the diet means. Distance from the red points to the blue line contributes to between-group variation. Distance from observations to their group mean contributes to within-group variation.

MATHEMATICAL DETAIL

Suppose there are k groups, Y_{ij} is the j -th observation in group i , $\bar{Y}_i$ is the mean of group i , and $\bar{Y}$ is the grand mean.

$$SS_{\text{total}} = SS_{\text{among}} + SS_{\text{within}} \quad (8.1)$$

The total sum of squares is:

$$SS_{\text{total}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 \quad (8.2)$$

The among-groups sum of squares is:

$$SS_{\text{among}} = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 \quad (8.3)$$

The within-groups sum of squares is:

$$SS_{\text{within}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad (8.4)$$

ANOVA converts these to mean squares:

$$MS_{\text{among}} = \frac{SS_{\text{among}}}{k - 1}, \quad MS_{\text{within}} = \frac{SS_{\text{within}}}{N - k} \quad (8.5)$$

and compares them with the F -ratio:

$$F = \frac{MS_{\text{among}}}{MS_{\text{within}}} \quad (8.6)$$

Large F values indicate that the group means differ more than we would expect from within-group variation alone.

8.3 DATA STRUCTURE AND DECISION RULE

Use a one-way ANOVA when:

- the response variable is continuous;
- the predictor is categorical with three or more independent levels;
- the observations are independent;
- the within-group distributions are reasonably well behaved.

Use this decision rule:

- if the group spreads are similar, use classical one-way ANOVA with `aov()`;
- if the data are strongly non-normal and a rank-based comparison is more defensible, use Kruskal-Wallis with `kruskal.test()`.

Boxplots with jitter are the first EDA tool to use here. Look for:

- separation in group centres;
- large differences in spread;
- outliers that dominate one group;
- shapes that suggest strong skewness.

These grouped plots are exploratory checks, not the final diagnostic step. In ANOVA itself, the assumptions are about the residual variation left after fitting the group means. I use the grouped plots here because they make the data structure easy to see, but the full residual-based analysis is developed explicitly in [Chapter 11](#).

ABOUT HOMOGENEITY TESTS

You've heard this before: Levene's test and Bartlett's test are supporting tools, not the main decision pathway. Visual comparison of spread comes first. If the group spreads differ substantially, switch to Welch's ANOVA rather than treating a non-significant homogeneity test as permission to ignore the problem.

8.4 R FUNCTIONS

The main functions used in this chapter are:

- `aov()` for classical ANOVA;
- `summary()` for the ANOVA table;
- `TukeyHSD()` for post hoc pairwise comparisons;
- `oneway.test()` for Welch's one-way ANOVA;
- `kruskal.test()` for the Kruskal-Wallis rank-sum test.

A further practical point is worth stating now, even though I explain it properly only later. Once an ANOVA model has been fitted, the assumptions are checked most directly on the residuals from that model. In this chapter I still use grouped plots and summaries as the first pass because they are easy to read, but I also show the residual diagnostics briefly so that the workflow is already in place. [Chapter 11](#) returns to this carefully and explains why residuals become the main object of diagnosis.

8.5 EXAMPLE 1: CHICK MASSES UNDER FOUR DIETS

At Day 21, do chicks fed four diets differ in mean body mass?

I set my hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : not all group means are equal

8.5.1 The One-Way Design

This is a one-way design because there is one grouping variable, Diet, with four levels. The response variable is chick mass.

8.5.2 Do an Exploratory Data Analysis (EDA)

I begin with a boxplot and jittered points so that the distribution within each diet is visible in Figure 8.2.

```
ggplot(chicks_day21, aes(x = Diet, y = weight)) +  
  geom_boxplot(  
    width = 0.6,  
    fill = "grey85",  
    colour = "black",  
    outlier.shape = NA  
  ) +  
  geom_jitter(width = 0.08, alpha = 0.5, size = 1.5, colour = "grey40") +  
  stat_summary(  
    fun = mean,  
    geom = "point",  
    size = 2.5,  
    colour = "black"  
  ) +  
  stat_summary(  
    fun.data = mean_sdl,  
    fun.args = list(mult = 1),  
    geom = "errorbar",  
    width = 0.12,  
    linewidth = 0.5,  
    colour = "black"  
  ) +  
  labs(x = "Diet", y = "Chick mass (g)")
```

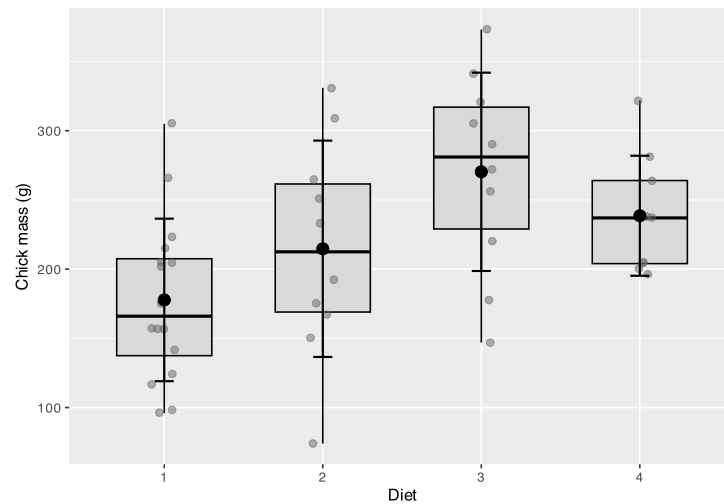


FIGURE 8.2. Boxplot and jittered observations for chick mass at Day 21 under four diets. Black points show the diet means and the vertical bars show one standard deviation above and below each mean.

The plot suggests that Diet 3 has the largest mean mass and Diet 1 the smallest. The spreads are not identical, but they are similar enough that a classical one-way ANOVA is plausible.

I support that visual judgement with grouped summaries and a homogeneity test.

```
chicks_day21 >
  group_by(Diet) >
  dplyr::summarise(
    n = dplyr::n(),
    mean = mean(weight),
    sd = sd(weight),
    shapiro_p = shapiro.test(weight)$p.value,
    .groups = "drop"
  )
```

```
R> # A tibble: 4 x 5
R>   Diet     n mean   sd shapiro_p
R>   <fct> <int> <dbl> <dbl>     <dbl>
R> 1 1      16 178.  58.7  0.591
R> 2 2      10 215.  78.1  0.949
R> 3 3      10 270.  71.6  0.895
R> 4 4       9 239.  43.3  0.186
```

```
car::leveneTest(weight ~ Diet, data = chicks_day21)
```

```
R> Levene's Test for Homogeneity of Variance (center = median)
R>    Df F value Pr(>F)
R> group 3  1.1597 0.3367
R>    41
```

```
chicks_aov1 <- aov(weight ~ Diet, data = chicks_day21)
summary(chicks_aov1)
```

The Shapiro-Wilk tests show no strong evidence against normality within any diet at Day 21, and Levene's test shows no strong evidence that the spreads differ. So, there is no strong reason to not apply an ANOVA.

8.5.3 Apply the Test

```
R>           Df Sum Sq Mean Sq F value Pr(>F)
R> Diet           3   57164    19055   4.655 0.00686 **
R> Residuals    41  167839     4094
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8.5.4 Brief Residual Check

Before interpreting the ANOVA table, I take a quick look at the fitted residuals (Figure 8.3). The residual-versus-fitted panel checks whether the spread is broadly stable across the fitted group means, and the Q-Q plot checks whether the residual distribution is close enough to normal for the usual ANOVA inference.

I do not develop these plots fully here because the full thinking behind residual tests belongs to [Chapter 11](#). For now, the point is simply that ANOVA assumptions are checked on the fitted model as well as through grouped exploratory displays.

```
par(mfrow = c(1, 2))
plot(chicks_aov1, which = 1)
plot(chicks_aov1, which = 2)
par(mfrow = c(1, 1))
```

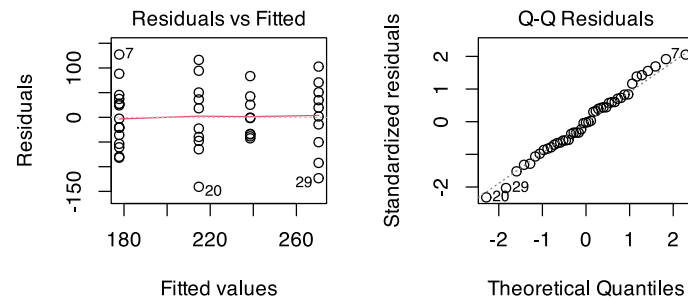


FIGURE 8.3. Brief residual diagnostics for the chick-mass ANOVA. The left panel shows residuals versus fitted values and the right panel shows a normal Q-Q plot of the residuals.

For this example, the residuals do not show an alarming departure from the model assumptions. The grouped summaries and the residual diagnostics therefore tell the same broad story: a classical one-way ANOVA is reasonable here.

SELF-ASSESSMENT TASK 8-2

- What does the outcome say about the chicken masses? Which ones are different from each other? (/2)
- Devise a graphical display of this outcome. (/4)

The ANOVA table answers the global question only. A significant result means that at least one diet mean differs from at least one other diet mean.

8.5.5 Identify Which Groups Differ

Once there is enough evidence to no longer accept the global null hypothesis, I use Tukey's HSD to identify the pairwise differences while controlling the family-wise error rate.

```
TukeyHSD(chicks_aov1)
```

```
R> Tukey multiple comparisons of means
R> 95% family-wise confidence level
R>
R> Fit: aov(formula = weight ~ Diet, data = chicks_day21)
R>
R> $Diet
R>      diff      lwr      upr      p adj
```



```
R> 2-1 36.95000 -32.11064 106.01064 0.4868095
R> 3-1 92.55000 23.48936 161.61064 0.0046959
R> 4-1 60.80556 -10.57710 132.18821 0.1192661
R> 3-2 55.60000 -21.01591 132.21591 0.2263918
R> 4-2 23.85556 -54.85981 102.57092 0.8486781
R> 4-3 -31.74444 -110.45981 46.97092 0.7036249
```

Each row in the Tukey output is one pairwise comparison. The `diff` column is the estimated mean difference between the two diets. The `lwr` and `upr` columns give the confidence interval for that difference. If the confidence interval does not cross zero, the two diets differ at the chosen family-wise confidence level. If the interval includes zero, the data do not support a clear difference for that pair. The `p adj` column gives the multiplicity-adjusted p -value for the same comparison.

```
plot(TukeyHSD(chicks_aov1))
```

95% family-wise confidence level

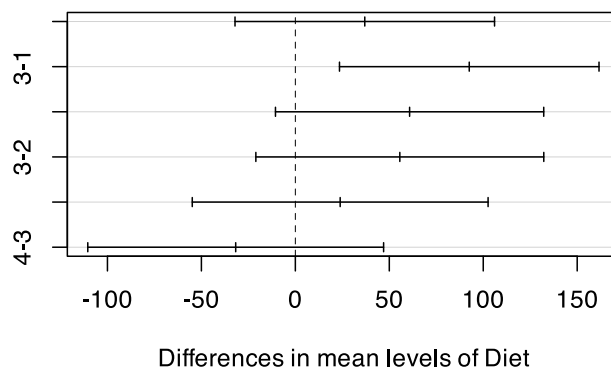


FIGURE 8.4. Tukey HSD comparisons among the four chick diets at Day 21.

The pairwise pattern is easier to scan in Figure 8.4.

DO IT NOW!

Produce a `ggplot2` version of the Tukey result.

SELF-ASSESSMENT TASK 8-3

Look at the help file for the `TukeyHSD()` function to better understand what the output means.

- How does one interpret the results? What does this tell us about the effect that different diets has on the chicken weights at Day 21? **(/3)**
- Figure out a way to plot the Tukey HSD outcomes in **ggplot**. **(/10)**
- Why does the ANOVA return a significant result, but the Tukey test shows that not all of the groups are significantly different from one another? **(/3)**

8.5.6 Interpret the Results

The global ANOVA is significant, so I reject H_0 and conclude that diet explains some of the variation in chick mass at Day 21. The clearest pairwise difference is between Diet 1 and Diet 3. Diet 3 chicks are about 92.6 g heavier on average than Diet 1 chicks.

The observed diet means are 177.8 g, 214.7 g, 270.3 g, and 238.6 g. Those differences are not all the same size, and the Tukey step shows that only some of them are large enough relative to within-diet variation to be declared significant.

The overall effect size is also worth reporting. For this ANOVA, about 25% of the variation in Day 21 chick mass is associated with diet ($\eta^2 \approx 0.25$).

8.5.7 Reporting

WRITE-UP**Methods**

Chick mass at Day 21 was compared among four diet treatments with a one-way ANOVA. The grouped distributions were inspected with boxplots and summary statistics, and group spread was checked before fitting the model. Tukey's HSD procedure was then used for post hoc pairwise comparisons.

Results

At Day 21, mean chick mass differed among diets (one-way ANOVA: $F_{3,41} = 4.66$, $p < 0.01$). Chicks fed Diet 1 had a mean mass of 177.8 g (SD = 58.7), compared with 214.7 g (SD = 78.1) for Diet 2, 270.3 g (SD = 71.6) for Diet 3, and 238.6 g (SD = 43.3) for Diet 4. Tukey's HSD showed that the clearest pairwise difference was between Diet 1 and Diet 3 ($p < 0.01$), with Diet 3 chicks heavier by 92.6 g on average. Diet accounted for about 25% of the variation in chick mass at Day 21 ($\eta^2 \approx 0.25$).

Discussion

The post hoc structure shows selective diet differences rather than universal separation among all four diet means. The global ANOVA detects structure among the means; Tukey's HSD identifies where that structure lies.

DO IT NOW!

Repeat the ANOVA analysis above using the built-in `PlantGrowth` dataset instead of the `chicks` data. This dataset contains dry weight of plants under a control condition and two treatment conditions.

- Make a box plot of `weight ~ group`.
- Check the normality assumption within each group (histograms or Shapiro-Wilk).
- Fit a one-way ANOVA: `aov(weight ~ group, data = PlantGrowth)`.
- If the global test is significant, apply Tukey's HSD and identify which groups differ.
- Write a two-sentence Results statement following the format above.

8.6 FACTORIAL ANOVA

So far, there has been one grouping variable only. I now extend the same idea to more than one factor.

In the `ChickWeight` data, diet is not the only categorical structure. Time also changes chick mass. That gives two distinct questions:

- what is the effect of diet?
- does the effect of diet change through time?

8.6.1 A Detour We Should Avoid

One tempting approach is to run a separate one-way ANOVA at each time point. That repeats the same multiple-testing problem that motivated ANOVA in the first place.

```
times <- c(0, 2, 10, 21)
p_values <- numeric(length(times))

for (i in seq_along(times)) {
  time_data <- filter(chicks, Time == times[i])
  fit <- aov(weight ~ Diet, data = time_data)
  p_values[i] <- summary(fit)[[1]][["Pr(>F)"]][1]
}

time_scan <- data.frame(
  Time = times,
  p_value = p_values
)

time_scan
```

```
R>   Time      p_value
R> 1    0 0.3459145928
R> 2    2 0.0055475424
R> 3   10 0.0009891053
R> 4   21 0.0068579588
```

The results are interpretable, but my method was less than ideal (and not to mention, also repetitive). Each time point gets its own ANOVA, and the family-wise error rate rises again. Separate ANOVAs across time repeat the same multiple-testing problem as separate *t*-tests across diets.

SELF-ASSESSMENT TASK 8-4

- What do you conclude from the above series of ANOVAs? (/3)
- What problem is associated with running multiple tests in the way that we have done here? (/2)

SELF-ASSESSMENT TASK 8-5

- How is time having an effect? (/3)
- What hypotheses can we construct around time? (/2)

8.6.2 Two Factors in One Model

A two-way ANOVA brings diet and time into the same model. Below, Time is converted to a factor because the question is about distinct sampling times. Should I keep

time as a continuous covariate, this would change the model into a linear regression, as I show in [Chapter 12](#) and extend in [Chapter 14](#).

```
chicks_two_way <- filter(chicks, Time %in% c(0, 21)) >
  mutate(Time = factor(Time))

summary(aov(weight ~ Diet + Time, data = chicks_two_way))
```

```
R>           Df Sum Sq Mean Sq F value    Pr(>F)
R> Diet           3  39595    13198    5.987 0.00091 ***
R> Time           1 734353    734353 333.120 < 2e-16 ***
R> Residuals     90 198402     2204
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SELF-ASSESSMENT TASK 8-6

- Write out the hypotheses for this ANOVA. (/2)
- What do you conclude from the above ANOVA? (/3)

With this additive model, I ask whether diet and time each have an effect on chick mass, assuming that the diet effect is the same at both time points.

8.6.3 Interaction in Factorial ANOVA

If the effect of diet changes with time, the additive model is too simple. We then include an interaction term:

```
summary(aov(weight ~ Diet * Time, data = chicks_two_way))
```

```
R>           Df Sum Sq Mean Sq F value    Pr(>F)
R> Diet           3  39595    13198    6.839 0.000343 ***
R> Time           1 734353    734353 380.521 < 2e-16 ***
R> Diet:Time       3  30504    10168    5.269 0.002187 **
R> Residuals     87 167898     1930
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction addresses the question whether diet-related differences are the same at Time 0 and Time 21. Stated differently, does the effect of diet change depending on when we sample the chicks? In this subset, yes, it depends. The interaction term is significant, so the effect of diet depends on time. [Chapter 15](#) develops the interpretation of interactions in much more detail.

DO IT NOW!

- a. What does the interaction term mean biologically in this chick example?
- b. Why do we lose information if we fit only separate one-way ANOVAs across time?

8.7 IF ASSUMPTIONS FAIL

Should some of the assumption tests fail, the one-way alternatives you could try are:

- use classical ANOVA when the group spreads are similar and the grouped distributions are reasonably well behaved;
- use Welch's ANOVA when the main problem is unequal spread across groups;
- use Kruskal-Wallis when a rank-based one-way comparison is more defensible;
- switch to a different model when the real issue is dependence, repeated measurement, or a non-Gaussian response structure.

8.7.1 Welch's One-Way ANOVA

Welch's ANOVA keeps the same one-way comparison-of-means question but does not assume equal variances.

```
oneway.test(weight ~ Diet, data = chicks_day21)
```

```
R>
R> One-way analysis of means (not assuming equal variances)
R>
R> data: weight and Diet
R> F = 4.6618, num df = 3.000, denom df = 20.594, p-value = 0.01219
```

As before, the question is, does at least one group mean differ? The interpretation is also the same as for classical one-way ANOVA. The difference is only that Welch's test does not pool the within-group variances into one common estimate.

8.7.2 Kruskal-Wallis Rank-Sum Test

If the response is too badly behaved for a mean-based one-way ANOVA, Kruskal-Wallis is the standard one-way rank-based alternative.

```
kruskal.test(weight ~ Diet, data = chicks_day21)
```

```
R>
R> Kruskal-Wallis rank sum test
R>
R> data: weight by Diet
R> Kruskal-Wallis chi-squared = 10.585, df = 3, p-value = 0.0142
```

As with one-way ANOVA, a significant Kruskal-Wallis result is global. It tells us that the groups do not all occupy the same part of the rank distribution, but it does not identify which groups differ.

```

pairwise.wilcox.test(
  x = chicks_day21$weight,
  g = chicks_day21$Diet,
  p.adjust.method = "holm"
)

R>
R> Pairwise comparisons using Wilcoxon rank sum test with continuity correction
R>
R> data: chicks_day21$weight and chicks_day21$Diet
R>
R>   1      2      3
R> 2 0.616 -      -
R> 3 0.026 0.573 -
R> 4 0.126 0.616 0.616
R>
R> P value adjustment method: holm

```

Kruskal-Wallis is a one-way method. There is no direct two-way Kruskal-Wallis counterpart that plays the same introductory role for factorial designs. Once the design includes multiple factors, the analysis needs a different modelling strategy rather than a simple rank-based extension of the one-way test.

DO IT NOW!

The `InsectSprays` dataset (built into R) gives counts of insects under six spray treatments. Counts are non-negative integers that may not meet ANOVA's normality assumption.

- Make a box plot: `boxplot(count ~ spray, data = InsectSprays)`. Is there obvious right-skew or unequal variance?
- Run the Shapiro-Wilk test within each spray group.
- Apply the Kruskal-Wallis test: `kruskal.test(count ~ spray, data = InsectSprays)`.
- If significant, apply the pairwise Wilcoxon test with Holm correction.
- Compare this result with a log-transformed ANOVA on the same data. Do the conclusions agree?

8.8 WHAT THIS CHAPTER DOES NOT COVER

- ANCOVA** is a linear model with a factor and a continuous covariate. It belongs in [Chapter 14](#), where factors and covariates are treated within the same model.
- Repeated-measures, blocked, or within-individual designs** are dependence problems. Those belong in [Chapter 19](#).
- Multifactor rank-based procedures** are not a clean introductory counterpart to factorial ANOVA. Once the design reaches that level, it is usually better to rethink the modelling approach directly.

8.9 CONCLUSION

ANOVA is the multi-group extension of the mean-comparison questions introduced with t -tests. The sequence is straightforward: recognise when repeated pairwise testing is inappropriate, ask one global question, inspect the grouped data, fit the model, and then use a post hoc procedure only if the global null hypothesis has been rejected.

The one-way case gives the foundation. Factorial ANOVA adds more than one categorical predictor. Later chapters connect this grouped-data view of ANOVA to the broader linear-model family.

9

Correlation and Association

Quantifying Relationships Without Imposing a Response Model

IN THIS CHAPTER

- correlation as a measure of association
- choosing Pearson, Spearman, or Kendall
- correlation matrices and heatmaps
- correlation and non-independence
- correlation versus regression

CHEATSHEET

Find here a [Cheatsheet](#) on statistical methods.

TASKS TO COMPLETE IN THIS CHAPTER

- Self-Assessment Task 9-1 (/10)
- Self-Assessment Task 9-2 (/20)
- [Self-Assessment instructions and full task overview](#)

In the previous chapters, I asked whether means differ among groups. Correlation answers, do two variables vary together?

Here we concern ourselves with **association** rather than group comparison. It is the last of the staple inferential tools and prepares the transition into regression. Correlation does not impose a response model and it simply measures the strength and direction of association between two variables. As we shall see later, the purpose

of regression is to have additional expectations about the *roles* of the variables: one variable is the response and the other predicts.

Correlation coefficients are effect sizes. Their sign shows direction and their magnitude shows how tightly the variables vary together. Coefficients vary from -1.0 (perfect inverse correlation), to 0 (no association), to 1.0 (perfect positive correlation).

9.1 CHOOSING THE APPROPRIATE CORRELATION

The main decision we will face is the form of the relationship.

- Use **Pearson's correlation** when the relationship is approximately linear and the main question concerns linear co-variation.
- Use **Spearman's rank correlation** when the relationship is monotonic but not especially linear, or when ranked order is more informative than raw spacing.
- Use **Kendall's rank correlation** when the emphasis is on concordance in rank order, especially with ordinal data, many tied ranks, or a direct question about agreement between rankings.

Start with a scatterplot and then decide whether the pattern is:

- roughly linear;
- monotonic but curved or unevenly spaced;
- clustered, outlier-driven, or structured by site, time, or some other grouping.

Association does not imply causation. Correlation can reveal co-variation, but it cannot identify the mechanism behind it.

9.1.1 *How to Read Magnitude*

Rough verbal labels can help, but they are only a starting point:

- around 0.1 to 0.3 : weak association;
- around 0.3 to 0.7 : moderate association;
- above 0.7 : strong association.

Those labels are only heuristics. A coefficient of 0.4 may be unremarkable in one system and biologically substantial in another, so it is highly context specific. A moderate correlation may be biologically useful if the variables are noisy or hard to measure. A strong correlation may still be uninformative if it is driven by site structure, repeated measurements, or a lurking third variable. Sample size also affects stability because correlation estimates are less stable in small datasets.

9.2 DATA STRUCTURE AND DIAGNOSTICS

Three questions should be settled before calculating a coefficient.

9.2.1 *What is the Data Structure?*

- Each observation in one variable must correspond to the same observation in the other variable.
- The sampling units should be independent.
- Correlation is calculated for one variable pair at a time, even if many pairs are later assembled into a matrix.

9.2.2 *What is the Relationship Form?*

- Pearson focuses on linear association.
- Spearman and Kendall focus on ordered association.
- A scatterplot is the main diagnostic because it shows whether the pattern is linear, monotonic, clustered, or broken by outliers.

9.2.3 *What's the Data Quality?*

- Outliers can distort any coefficient, but Pearson is especially sensitive to them.
- Site, time, transect, quadrat, or repeated-measures structure can create correlations that do not reflect the biological question of interest.

9.2.4 *Secondary Considerations*

Distributional shape is less important than relationship form, outliers, and independence. Pearson is most sensitive to non-linearity and influential points. Normality is not the first issue to inspect.

DO IT NOW!

Using the built-in `iris` dataset, examine the relationship between `Sepal.Length` and `Petal.Length`:

- a. Make a scatter plot coloured by `Species`. Is the overall pattern linear?
- b. Make the same scatter plot with only the *Setosa* points. Is the pattern still linear?
- c. Based on your plots, is there a data-structure issue that could affect your correlation estimate if you compute a single pooled Pearson coefficient?

What would you do differently to compute a correlation that correctly reflects the within-species relationship?

9.3 R FUNCTIONS

The main function in this chapter is `cor.test()`. Use:

- `method = "pearson"` for Pearson's product-moment correlation;
- `method = "spearman"` for Spearman's rank correlation;
- `method = "kendall"` for Kendall's rank correlation.

The function `cor()` calculates the coefficient itself and is useful for pairwise matrices. The inferential version, `cor.test()`, is usually more useful in worked examples because it returns the coefficient, test statistic, and *p*-value.

MATHEMATICAL DETAIL

For Pearson correlation:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (9.1)$$

For Spearman rank correlation:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (9.2)$$

For Kendall rank correlation:

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} \quad (9.3)$$

These formulas show the structure of the coefficients. In practice, first decide which coefficient matches the pattern in the data.

9.4 PEARSON CORRELATION

Pearson's *r* measures the strength and direction of linear association. It is appropriate when the relationship is reasonably well described by a straight line and no small set of outliers dominates the pattern.

9.5 EXAMPLE 1: PEARSON CORRELATION IN ECKLONIA MAXIMA

I begin by asking whether the length of *Ecklonia maxima* stipes tends to increase together with frond length.

```
ecklonia <- read_csv(here::here("data", "BCB744", "ecklonia.csv"))
```

9.5.1 Do an Exploratory Data Analysis (EDA)

```
ecklonia > summarise(
  n = n(),
  mean_stipe_length = mean(stipe_length),
  sd_stipe_length = sd(stipe_length),
  mean_fronde_length = mean(fronde_length),
  sd_fronde_length = sd(fronde_length)
)

R> # A tibble: 1 x 5
R>       n mean_stipe_length sd_stipe_length mean_fronde_length sd_fronde_length
R>   <int>         <dbl>         <dbl>         <dbl>         <dbl>
R> 1     26           531.           132.           171.           49.4

r_print <- paste0(
  "r = ",
  round(cor(ecklonia$stipe_length, ecklonia$fronde_length), 2)
)

ggplot(data = ecklonia, aes(x = stipe_length, y = fronde_length)) +
  geom_smooth(method = "lm", colour = "blue3", se = FALSE, linewidth = 1) +
  geom_point(size = 2.7, colour = "red3", shape = 16) +
  geom_label(x = 300, y = 240, label = r_print) +
  labs(x = "Stipe length (cm)", y = "Fronde length (cm)")
```

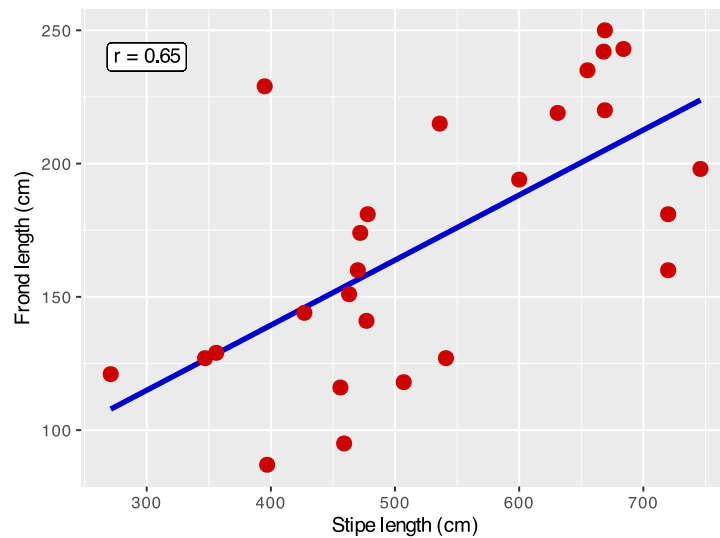


FIGURE 9.1. Scatterplot showing the relationship between *Ecklonia maxima* stipe length and frond length. The fitted line is included only to show the overall linear tendency.

The scatterplot in Figure 9.1 shows a clear positive linear trend. The points are not tightly packed around the fitted line, but the relationship is straight enough that Pearson's coefficient is appropriate. The main diagnostic here is the linear form of the line and univariate normality is secondary.

9.5.2 State the Hypotheses

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

Here ρ is the population Pearson correlation coefficient.

9.5.3 Apply the Test

```
cor.test(
  x = ecklonia$stipe_length,
  y = ecklonia$frond_length,
  use = "everything",
  method = "pearson"
)
```

```

R>
R> Pearson's product-moment correlation
R>
R> data: ecklonia$stipe_length and ecklonia$frond_length
R> t = 4.2182, df = 24, p-value = 0.0003032
R> alternative hypothesis: true correlation is not equal to 0
R> 95 percent confidence interval:
R>  0.3548169 0.8300525
R> sample estimates:
R>      cor
R> 0.6524911

```

9.5.4 Interpret the Results

The correlation is positive and moderately strong. Longer stipes tend to be associated with longer fronds. The coefficient is large enough to be biologically informative, and the p -value is well below 0.001, so we reject H_0 .

The effect size is the coefficient itself. Here $r = 0.65$ indicates a fairly strong linear association for a biological dataset of this kind. That does not imply causation. It shows co-variation.

9.5.5 Reporting

WRITE-UP

Methods

The linear association between stipe length and frond length in *Ecklonia maxima* was assessed with a Pearson product-moment correlation.

Results

Stipe length and frond length in *Ecklonia maxima* were positively correlated (Pearson correlation: $r = 0.65$, $n = 26$, $p < 0.001$), indicating that kelps with longer stipes also tended to have longer fronds.

Discussion

The two morphological variables co-vary strongly in these data. The result shows association only. It does not identify a causal mechanism.

DO IT NOW!

Using the `mtcars` dataset, compute the Pearson correlation between `hp` (horsepower) and `mpg` (fuel economy).

- First make a scatter plot. Does the relationship look linear?
- Compute `cor.test(mtcars$hp, mtcars$mpg)`. Report the correlation coefficient, 95% CI, and p -value.

- c. Write a one-sentence interpretation following the reporting style used in the Write-Up above (include r , df , and p -value).
- d. Does the negative sign make biological sense? Explain.

9.6 SPEARMAN RANK CORRELATION

When the relationship is ordered but not especially linear, Spearman's ρ is often the better choice. It replaces the raw values with ranks and asks whether the two variables tend to increase together in rank order.

9.7 EXAMPLE 2: SPEARMAN'S ρ WHEN THE RANK PATTERN IS CLEARER THAN THE LINEAR ONE

In the *ecklonia* data, the relationship between stipe diameter and primary blade length is not especially well described by a straight line, but there is still a biological suggestion that thicker stipes tend to occur with longer primary blades.

9.7.1 Do an Exploratory Data Analysis (EDA)

As I always do, I start with a scatterplot.

```
ecklonia > summarise(
  n = n(),
  mean_stipe_diameter = mean(stipe_diameter),
  sd_stipe_diameter = sd(stipe_diameter),
  mean_primary_blade_length = mean(primary_blade_length),
  sd_primary_blade_length = sd(primary_blade_length)
)
```

```
R> # A tibble: 1 x 5
R>   n mean_stipe_diameter sd_stipe_diameter mean_primary_blade_length
R>   <int>           <dbl>           <dbl>           <dbl>
R> 1    26             24.2             6.74             17.9
R> # i 1 more variable: sd_primary_blade_length <dbl>

ggplot(ecklonia, aes(x = stipe_diameter, y = primary_blade_length)) +
  geom_point(shape = 1, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = FALSE, colour = "firebrick") +
  labs(x = "Stipe diameter (mm)", y = "Primary blade length (cm)")
```

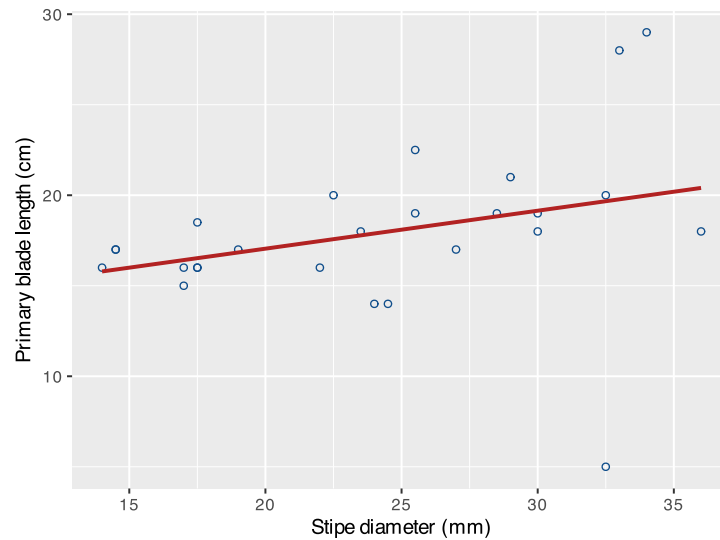



FIGURE 9.2. Relationship between stipe diameter and primary blade length in *Ecklonia maxima*. The fitted line is shown only as a visual guide to the overall trend.

The scatterplot in Figure 9.2 suggests an increasing pattern, but the spacing around a straight line is uneven and a few observations influence the fitted line strongly. The biological question is about ordered increase rather than precise linear scaling. That points to Spearman rather than Pearson.

9.7.2 State the Hypotheses

$$H_0 : \rho_s = 0$$

$$H_a : \rho_s \neq 0$$

Here ρ_s is the population Spearman rank-correlation coefficient.

9.7.3 Apply the Test

```
cor.test(
  ecklonia$stipe_diameter,
  ecklonia$primary_blade_length,
  method = "spearman",
  exact = FALSE
)
```

R>

```
R> Spearman's rank correlation rho
R>
R> data: ecklonia$stipe_diameter and ecklonia$primary_blade_length
R> S = 1444.1, p-value = 0.008311
R> alternative hypothesis: true rho is not equal to 0
R> sample estimates:
R>      rho
R> 0.5062992
```

It is useful to compare this with Pearson's correlation on the same variable pair:

```
cor.test(
  ecklonia$stipe_diameter,
  ecklonia$primary_blade_length,
  method = "pearson"
)

R>
R> Pearson's product-moment correlation
R>
R> data: ecklonia$stipe_diameter and ecklonia$primary_blade_length
R> t = 1.6413, df = 24, p-value = 0.1138
R> alternative hypothesis: true correlation is not equal to 0
R> 95 percent confidence interval:
R> -0.07946077 0.62777344
R> sample estimates:
R>      cor
R> 0.3176688
```

9.7.4 Interpret the Results

Spearman's ρ shows a moderate positive rank association. In these data, the rank-based signal is clearer than the strictly linear one. Pearson understates the pattern because it is asking a different question.

This is the main reason to use Spearman: the biological pattern is ordered, but the raw spacing around a line is not especially stable.

9.7.5 Reporting

WRITE-UP

Methods

A Spearman rank correlation was used to assess the monotonic association between stipe diameter and primary blade length in *Ecklonia maxima*. A rank-based method was preferred because the visual pattern suggested ordered increase without a clean linear form.

Results

Stipe diameter and primary blade length showed a moderate positive rank association (Spearman correlation: $\rho = 0.51$, $n = 26$, $p < 0.01$), indicating that kelps with thicker stipes also tended to have longer primary blades. On the same variable pair, the Pearson

correlation was weaker and not statistically convincing at the 5% level ($r = 0.32$, $p > 0.05$).

Discussion

Spearman's ρ is more informative here because the biological conclusion concerns ordered increase rather than exact linear scaling.

9.8 KENDALL RANK CORRELATION

Kendall's τ measures concordance in rank ordering. It asks whether pairs of observations tend to agree in how they are ordered on the two variables.

9.9 EXAMPLE 3: KENDALL'S τ FOR CONCORDANCE IN RANK ORDER

Suppose my question is whether longer primary blades also tend to be wider in the same rank order. This is a direct concordance question, which is the natural setting for Kendall's τ .

9.9.1 *Do an Exploratory Data Analysis (EDA)*

```
ggplot(ecklonia, aes(x = primary_blade_length, y = primary_blade_width)) +  
  geom_point(shape = 1, colour = "seagreen4") +  
  labs(x = "Primary blade length (cm)", y = "Primary blade width (cm)")
```

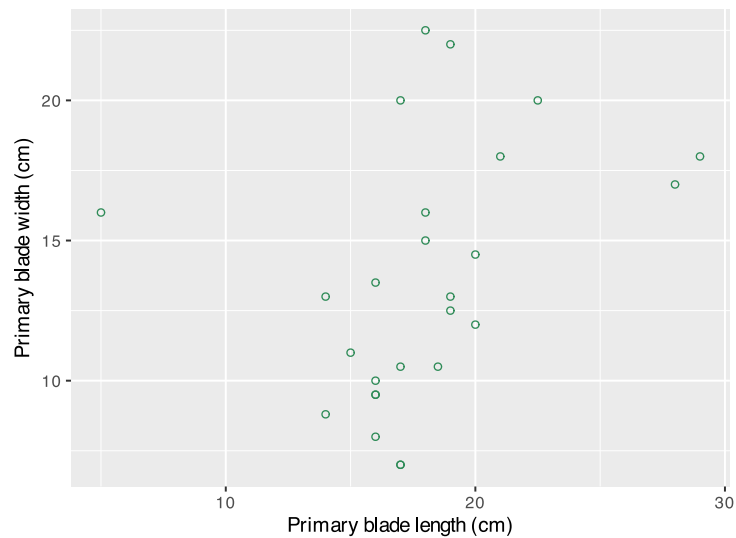


FIGURE 9.3. Relationship between primary blade length and primary blade width in *Ecklonia maxima*. The question here is whether longer blades also tend to rank as wider blades.

The scatterplot in Figure 9.3 shows a positive tendency, but the main point is not exact linear spacing. The question is whether larger values in one variable tend to be matched by larger values in the other. That is a concordance question, so Kendall's τ is appropriate.

9.9.2 State the Hypotheses

$$H_0 : \tau = 0$$

$$H_a : \tau \neq 0$$

Here τ is the population Kendall rank-correlation coefficient.

9.9.3 Apply the Test

```
cor.test(
  ecklonia$primary_blade_length,
  ecklonia$primary_blade_width,
  method = "kendall"
)
```

R>

```
R> Kendall's rank correlation tau
R>
R> data: ecklonia$primary_blade_length and ecklonia$primary_blade_width
R> z = 2.3601, p-value = 0.01827
R> alternative hypothesis: true tau is not equal to 0
R> sample estimates:
R>      tau
R> 0.3426171
```

9.9.4 Interpret the Results

Kendall's τ is positive, so the rank ordering is broadly consistent: longer blades also tend to be wider. The effect is moderate rather than strong, and the p -value is below 0.05, so we reject H_0 .

Kendall is often less familiar than Spearman, but its interpretation is very direct when the scientific question is about agreement in rank order.

9.9.5 Reporting

WRITE-UP

Methods

Kendall's τ was used to assess concordance between primary blade length and primary blade width in *Ecklonia maxima*.

Results

Primary blade length and primary blade width showed a positive association in rank order (Kendall correlation: $\tau = 0.34$, $n = 26$, $p < 0.05$), indicating that longer blades also tended to be wider.

Discussion

Kendall's τ is useful when the scientific message concerns agreement in rank order rather than precise linear scaling.

9.10 CORRELATION MATRICES AND HEATMAPS

Once a single pairwise relationship is understood, the same idea can be scaled to many continuous variables at once.

```
ecklonia_sub <- ecklonia %>%
  select(-species, -site, -ID)

ecklonia_sub <- ecklonia_sub[, order(colnames(ecklonia_sub))]
```

TABLE 9.1. A pairwise Pearson correlation matrix of the *Ecklonia* dataset.

variable	digits	epiphyte_length	frond_length	frond_mass	primary_blade_length	primary_blade_width
digits	1.00	0.05	0.36	0.28	0.10	0.14
epiphyte_length	0.05	1.00	0.61	0.44	0.26	0.41
frond_length	0.36	0.61	1.00	0.57	-0.02	0.28
frond_mass	0.28	0.44	0.57	1.00	0.15	0.36
primary_blade_length	0.10	0.26	-0.02	0.15	1.00	0.34
primary_blade_width	0.14	0.41	0.28	0.36	0.34	1.00
stipe_diameter	0.24	0.54	0.39	0.51	0.32	0.32
stipe_length	0.24	0.61	0.65	0.51	0.13	0.13
stipe_mass	0.07	0.51	0.39	0.47	0.16	0.16

```
ecklonia_pearson <- round(cor(ecklonia_sub), 2)

ecklonia_pearson >
as.data.frame() >
tibble::rownames_to_column("variable") >
gt() >
fmt_number(columns = -variable, decimals = 2) >
tab_style(
  style = cell_text(weight = "bold"),
  locations = cells_column_labels(everything())
)
```

By producing many pairwise comparisons at once, these matrices offer a useful exploratory tool. They identify strong positive and negative relationships amongst many variables. Some of these correlations will appear by chance and others will be meaningful. `corrplot(ecklonia_pearson, method = "circle", na.label.col = "white")` can be used to guide further inspection and not be used as a finished inferential result.

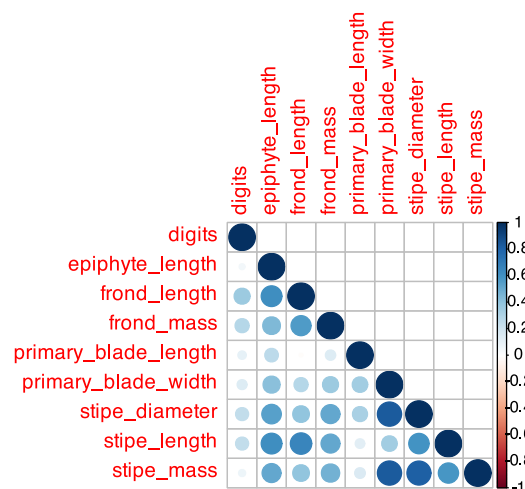


FIGURE 9.4. Pairwise correlations showing the strength of all Pearson correlations between variables as a scale from red (negative) to blue (positive).

```
ecklonia_pearson >
  as.data.frame() >
  mutate(x = rownames(ecklonia_pearson)) >
  pivot_longer(
    cols = -x,
    names_to = "y",
    values_to = "r"
  ) >
  ggplot(aes(x, y, fill = r)) +
  geom_tile(colour = "white") +
  scale_fill_gradient2(
    low = "blue",
    high = "red",
    mid = "white",
    midpoint = 0,
    limit = c(-1, 1),
    na.value = "grey95",
    space = "Lab",
    name = "r"
  ) +
  labs(x = NULL, y = NULL) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  coord_fixed()
```

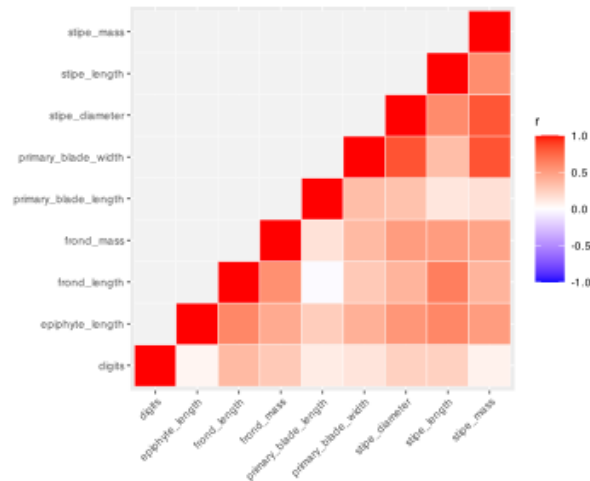


FIGURE 9.5. Pairwise correlations of the *Ecklonia* dataset visualised as a heatmap in **ggplot2**.

Together, Figure 9.4 and Figure 9.5 show the same correlation structure in two different visual styles. The matrix and heatmap are useful for scanning the dataset quickly, but they should be treated with caution. They do not control the false-positive rate across all displayed comparisons, and they do not explain why a correlation exists.

DO IT NOW!

Using the penguins dataset from **palmerpenguins**, compute a pairwise Pearson correlation matrix for the four continuous morphological variables (`bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`).

- Filter to complete cases first (`drop_na()`), then compute the matrix using `cor()`.
- Identify the strongest positive and strongest negative correlation.
- Does any of these correlations surprise you biologically? Why might `bill_depth_mm` and `bill_length_mm` have a negative pooled correlation across species?

9.11 CORRELATION AND NON-INDEPENDENT DATA

Even the correct coefficient can mislead when the observations are not independent. This is common in ecology because measurements are often grouped within sites, transects, quadrats, times, or individuals.

9.12 EXAMPLE 4: A CORRELATION CAN BE MODIFIED BY SITE STRUCTURE

I again use the kelp data.

```
ecklonia >
  group_by(site) >
  summarise(
    n = n(),
    r_site = cor(stipe_length, epiphyte_length)
  )

R> # A tibble: 2 x 3
R>   site      n r_site
R>   <chr>   <int> <dbl>
R> 1 Batsata Rock    13  0.223
R> 2 Boulders Beach  13  0.920

ggplot(ecklonia, aes(x = stipe_length, y = epiphyte_length, colour = site)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 0.7) +
  geom_smooth(aes(group = 1),
    method = "lm",
    se = FALSE,
    colour = "black",
    linetype = "dashed",
    linewidth = 0.6) +
  labs(x = "Stipe length (cm)", y = "Epiphyte length (cm)", colour = "Site")
```

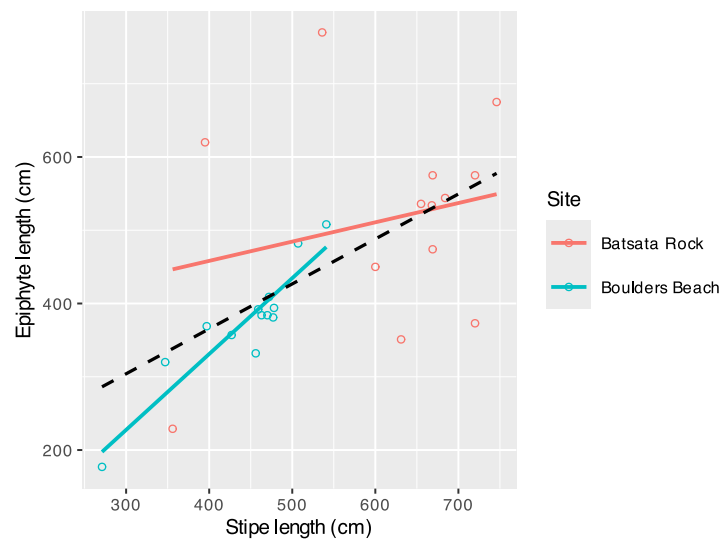


FIGURE 9.6. Relationship between stipe length and epiphyte length in *Ecklonia maxima*, coloured by site. The black dashed line is the naive overall fit, while the coloured lines show the within-site trends.

The contrast between the pooled fit and the within-site fits is clear in Figure 9.6.

The pooled correlation across all observations is clearly positive, but the plot also shows clustering by site. Part of the overall association reflects site differences rather than individual-level co-variation.

This is why independence must be taken seriously. A naive pooled coefficient mixes at least two sources of structure:

- variation among individuals within sites;
- variation between sites.

If the biological question is about site-level differences, then site should be modelled explicitly. If the question is about individual-level association, a grouped or mixed-effects model is more defensible than a pooled correlation.

9.13 CORRELATION VERSUS REGRESSION

Correlation and regression are related, but they do different jobs.

- **Correlation** quantifies how strongly two variables vary together.
- **Regression** estimates how the expected value of a response changes with a predictor.

Both are rooted in covariance, but Pearson's correlation standardises covariance into a unitless coefficient while simple linear regression uses the same covariance structure to estimate a slope. See the next chapter.

If the question is “do these variables co-vary?”, correlation is appropriate. If the question is “how much does the response change when the predictor changes?”, the analysis has already become a regression.

DO IT NOW!

For each of the following research questions, decide whether correlation or regression is more appropriate, and explain why:

- a. A botanist measures leaf area and stomatal density in 50 plants. She wants to know whether larger leaves have more stomata per unit area.
- b. A conservation biologist wants to predict the number of species in a forest patch from the patch area.
- c. Two ecologists each independently count the number of invasive plants in 20 quadrats. They want to know whether their counts agree.
- d. A physiologist wants to know how oxygen consumption changes for each 1°C increase in body temperature.

Which of these require a slope estimate? Which only require a measure of co-variation? Discuss with a partner.

9.14 IF ASSUMPTIONS FAIL

If the relationship is not linear, the variables are ordinal, or the data contain influential outliers, Pearson correlation may not be appropriate. In such cases:

- use Spearman or Kendall for rank-based association;
- inspect the scatterplot before trusting any coefficient;
- ask whether site, time, or repeated-measures structure is inflating the coefficient;
- treat matrices as exploratory summaries;
- move to regression if the real question is one of response and predictor, or if confounding and grouped structure need to be handled directly.

Assumption checking itself is discussed in [Chapter 6](#).

SELF-ASSESSMENT TASK 9-1

For each of the following pairs of variables, decide which correlation method (Pearson, Spearman, or Kendall) is most appropriate and briefly justify your choice:

- a. Body mass (g) and wing span (mm) in a sample of 120 birds — the scatterplot looks roughly linear. (/2)
- b. Ecologist ranks 15 sites from least to most disturbed (1–15); another ecologist ranks the same 15 sites by species richness. The question is whether more disturbed sites tend to rank lower in diversity. (/2)
- c. Sea surface temperature and chlorophyll-a concentration in a large oceanographic dataset — the relationship is clearly curvilinear (cooling increases phytoplankton). (/2)
- d. For (c), does Spearman also have an advantage over Pearson when data are skewed? (/2)
- e. A nutritionist scores diet quality on a 5-point ordinal scale and records BMI for 40 participants. (/2)

SELF-ASSESSMENT TASK 9-2

Find **two datasets** of your own and do a full correlation analysis on each. Briefly describe the data and why they exist. State the hypotheses, do an EDA, make exploratory figures, choose and justify the appropriate correlation method, assess assumptions, and write up the results in publication style.

Rubric

Criterion	Excellent (Full Marks)	Partial Credit	Absent / Poor	Marks
1. Dataset Choice and Justification	Two variables (from one or more datasets) are clearly described and justified as candidates for correlation analysis; rationale is thoughtful and contextually informed.	Variables are chosen and described but the rationale is vague or unconvincing.	Variable selection appears arbitrary or trivial; little or no justification is given.	/2
2. Hypothesis Framing	Null and alternative hypotheses are explicitly stated and aligned with the correlation analysis (e.g., $H_0 : \rho = 0$). Contextual meaning is clearly explained.	Hypotheses are present but poorly articulated or lacking contextual relevance.	Hypotheses are missing, incorrect, or misaligned with the analysis.	/2

3. Exploratory Data Analysis	EDA includes summary statistics, variable distribution inspection, and consideration of linearity or monotonicity. Potential issues (<i>e.g.</i> , outliers) are noted.	EDA is attempted but lacks depth or overlooks important features such as skewness or relationship form.	No meaningful EDA is performed before conducting the correlation.	/3
4. Exploratory Figures	Appropriate visualisation (<i>e.g.</i> , scatterplot with smoothing line, marginal histograms) is clear, labelled, and supports interpretation.	A plot is included but is unclear, poorly formatted, or not well interpreted.	No plot is provided, or the plot is irrelevant or uninformative.	/2

5. Correlation Method and Calculation	The correlation method is appropriate to the data characteristics, with Pearson, Spearman, or Kendall chosen and justified. Code and output are correct and clearly reported.	The method is used correctly but without justification, or there are some reporting issues.	Correlation is applied mechanically or incorrectly; code or output is missing.	/3
6. Significance and Effect Size	The p -value and correlation coefficient (r , ρ , or τ) are reported, with interpretation of both statistical and practical significance.	Results are reported but not clearly interpreted or contextualised.	The p -value or coefficient is misinterpreted, or key output is missing.	/2

7. Assumption Checking and Discussion	Relevant assumptions are addressed according to the chosen method (e.g., relationship form, outliers, independence), supported by appropriate plots and discussion.	Some assumptions are discussed or partially checked, but the reasoning is unclear or incomplete.	There is no discussion or evidence of assumption checking.	/3
8. Written Results Section	Results are presented in a clear, concise, publication-ready format, with technical correctness and a logical flow from EDA to conclusion.	Results are readable but disorganised, imprecise, or not fully connected to the evidence.	Results are unclear, incorrect, or unstructured.	/3
Total: /20				

9.15 SUMMARY

Correlation quantifies association without fitting a response model. The working sequence is:

1. plot the data;
2. decide whether the question is really about association;
3. choose Pearson, Spearman, or Kendall according to the pattern in the data;
4. check independence before trusting the coefficient;
5. interpret sign, magnitude, and uncertainty together;
6. treat pairwise matrices as exploratory, not definitive.

10

Finding Your Way on the Statistical Landscape

A Practical Decision Guide for Core Inferential Questions

IN THIS CHAPTER

- how two questions about your data narrow the choice to one method family
- a visual reference grid linking question structure to method
- when to leave the core tests behind and move into modelling

CHEATSHEET

- Here is a [course synthesis](#) for BCB744 Biostatistics.
- Find here a [cheatsheet](#) on statistical methods.
- Here is a [study design cheatsheet](#).

TASKS TO COMPLETE IN THIS CHAPTER

- None

By this point, you have already met the main inferential tools one by one. What is often still missing is a reliable way to navigate among them without treating statistics as a software menu. This chapter closes the inferential block by bringing those separate procedures back under one decision structure.

The important shift is that method choice should follow from the biological question, the response type, and the design structure that produced the data. When those are clear, the correct family of methods narrows quickly. When they are vague, no decision grid can rescue the analysis.

10.1 TWO IMPORTANT QUESTIONS

By now you have met the main tools: *t*-tests, ANOVA, and correlation. Choosing among them, and knowing when none of them is quite right, comes down to two questions about your data.

Question 1: What kind of response variable do you have?

Most biological responses are continuous measurements (length, mass, concentration). Some are binary or proportional (success/failure, proportion infected). Others are counts (number of individuals, number of events). The response type determines which model family is even sensible.

Question 2: What does the explanatory structure look like?

You are either comparing groups or describing a relationship. If you are comparing groups, how many, and are the observations in each group independent of one another or paired? If you are describing a relationship, is the predictor categorical or continuous?

That is really all you need to start. The method families I introduce in this course each answer one combination of those two questions. Figure 10.1 shows the connections at a glance.

10.2 THE CORE TESTS

10.2.1 *Comparing a Single Group to a Reference*

If you have one sample and want to know whether the mean is consistent with some reference value, use a **one-sample *t*-test**. The reference might be a regulatory threshold, a historical baseline, or a theoretical expectation. When the data are too skewed for a mean-based summary to be sensible, the **Wilcoxon signed-rank test** provides a rank-based alternative. Both are in [Chapter 7](#).

10.2.2 *Comparing Two Groups*

Two independent groups call for a **two-sample *t*-test**. When the same individuals or plots are measured under two conditions, or before and after a treatment, use a **paired *t*-test** instead. Pairing removes background noise and often sharpens the

One-sample t-test or proportion test	Wilcoxon signed-rank test or binomial / GLM extension	One mean or one proportion
Two-sample t-test	Wilcoxon rank-sum test or regression / GLM	Two independent groups
Paired t-test	Wilcoxon signed-rank test or mixed model	Paired or before-after data
ANOVA	Kruskal-Wallis or factorial / mixed-model extension	Three or more groups
Pearson / Spearman / Kendall correlation	Regression if one variable is a response	Two variables vary together
Linear regression	Polynomial, multiple regression, interactions	Continuous response with predictors
Proportion tests (simple cases)	GLMs: logistic or beta-style approaches	Binary or proportional response
Not a core introductory test	GLMs: Poisson / negative binomial	Count response
Not handled by core independent-sample tests	Mixed models / hierarchical models	Repeated or nested observations
Not handled well by straight-line models	GAMs	Curved but smooth relationship
Not handled well by straight-line models	Nonlinear regression / GNLm-style extensions	Mechanistic nonlinear response
Core Choice	Next Extension	Question Structure

FIGURE 10.1. Links from common biological question types to the main method and the next extension when the problem becomes more complex.

comparison. When group spreads differ noticeably, prefer Welch’s version, which makes no assumption of equal variance. Rank-based alternatives (`wilcox.test()`) are available when the distributional assumptions fail clearly. All of this is in [Chapter 7](#).

10.2.3 Comparing Three or More Groups

Once there are three or more groups, **ANOVA** replaces the *t*-test. Although the reasoning is the same (we still compare means), ANOVA keeps the false-positive rate in check across all comparisons at once. Factorial ANOVA extends this to two or more categorical factors and their interaction. When group spreads are unequal, Welch’s ANOVA applies. **Kruskal-Wallis** is the rank-based fallback. These are in [Chapter 8](#).

10.2.4 Asking Whether Two Variables Move Together

When the aim is association between two continuous variables and neither is designated the response, use **correlation**. Pearson’s *r* captures linear association; Spearman’s ρ and Kendall’s τ work on ranks and are appropriate when the relationship is monotone but not linear, or when the data contain outliers. Correlation says noth-

ing about prediction or cause — if you want a response model, move to regression instead. [Chapter 9](#) covers this.

10.2.5 *Modelling How One Variable Responds to Another*

When one variable is clearly the response and the other is used to explain it, move from correlation to **regression**. The simplest case is **simple linear regression**, where the response changes approximately linearly with the predictor. When the relationship is bendy, extend the model rather than forcing a straight line. **Poly-nomial regression** describe different ways the mean response can vary with the predictor within the same simple linear regression framework, but other ways to deal with non-linearity sit outside of the “core” group of tests and models. Regression is developed in [Chapter 12](#) and extended in later chapters.

DO IT NOW!

For each research question below, identify the correct method from the five core-test subsections above. Be specific: state whether you would use a one-sample *t*-test, two-sample *t*-test, paired *t*-test, ANOVA, or correlation, and give a one-sentence justification.

- A marine biologist measures the shell width of 25 mussels and wants to know if it exceeds the minimum harvestable size of 45 mm.
- A nutritionist checks whether dietary fat intake (g/day) is related to blood cholesterol (mmol/L) in 60 adults.
- An ecologist compares seed germination rates (%) across four light treatments.
- A physiologist measures blood glucose in 20 rats before and after an insulin injection.
- A forester compares the diameter at breast height in a treated forest stand versus an untreated control stand.

10.3 BEYOND THE CORE TESTS

The core tests cover a lot of ground, but they rest on assumptions that do not always hold.

A NAMING SHORTCUT

As you move into more advanced methods, some naming patterns start to help:

- **Generalised ...** usually means that the basic linear-model idea has been extended to allow response variables that are not continuous Gaussian measurements. For example, a **generalised linear model** lets the response follow distributions such as binomial or Poisson instead of the ordinary normal-error model used in simple linear regression.
- **Mixed ...** usually means that the model contains both **fixed effects** (the main predictors you want to estimate and interpret) and **random effects** (terms that account for grouping, repeated measurements, nesting, or other non-independence in the design). A **linear mixed model** is the standard example.

This terminology is useful because it lets you recognise what kind of complication a method is designed to handle. If you later encounter names such as **generalised additive model**, **generalised linear mixed model**, or **nonlinear mixed model**, the name already tells you something important about the model: whether it is extending the response distribution, the dependence structure, the shape of the relationship, or some combination of these. Learning to read the names of methods in this way will make the more advanced chapters much easier to navigate.

Use GLMs when the response is not continuous and Gaussian. Counts, proportions, and binary outcomes each have their own distributional family, and forcing them into a Gaussian model produces unreliable inference. [Chapter 20](#) introduces GLMs.

Use mixed models when observations are not independent, such as repeated measurements on the same individual, samples nested within sites, or any design where the experimental unit is not the same as the observational unit. The problem is one of design, not of distribution, and it cannot be fixed by transforming the response. [Chapter 19](#) addresses this.

Apply GAMs when the relationship is clearly curved in ways a straight line cannot describe, or use **nonlinear regression** when the shape of the relationship is tied in a mechanistic way to the biological phenomenon. [Chapter 21](#) and [Chapter 22](#) cover these extensions.

DO IT NOW!

Each scenario below clearly breaks the boundary of the core tests. For each one, identify which extension applies (GLM, mixed model, GAM, or nonlinear regression) and explain in one sentence why the standard core test is not adequate:

- a. You monitor the same 30 fish at four time points and want to compare their growth trajectories.

- b. You record the presence/absence of a disease across 200 animals and want to relate it to three continuous predictors.
 - c. You measure how predation rate changes with prey density and the pattern follows a saturating (Holling Type II) curve.
 - d. Phytoplankton chlorophyll concentration follows a clearly non-monotonic, hump-shaped relationship with depth.
- Maybe revisit these questions once you've familiarised yourself with the advanced methods.

10.4 ASSUMPTIONS: CHECK BEFORE YOU PROCEED

Every method in the table rests on assumptions. The most consequential one is **independence**: if your observations are not independent of one another (because they come from the same individual, the same site, or a before-after design) no distributional fix will rescue the analysis. This was cast in stone when you designed your sampling campaign or experiment, and the data properties have nothing to do with it.

After independence, check whether the distributional assumptions are roughly plausible. For the core *t*-tests and ANOVA, the concern is strong skew or extreme outliers. Use the checks introduced in [Chapter 6](#). If the assumptions fail clearly, there is usually a rank-based alternative or a transformation that gets you back on solid ground.

Do not switch methods without being able to say which assumption failed and why. Choosing a non-parametric test because you have heard it is “safer” is not a reason.

DO IT NOW!

Answer the two questions above for each of the following biological scenarios and decide which method family applies. Write your answers in the format: “Response type: [X]. Explanatory structure: [Y]. Method: [Z].”

- a. You compare the running speed (m/s) of 40 lizards from a hot habitat versus 40 from a cool habitat.
 - b. You record whether 200 fish (binary: alive or dead) survived a hypoxia event at three oxygen levels.
 - c. You measure the growth rate of 30 bacteria colonies (mm/day) and ask whether it is correlated with ambient temperature.
 - d. You track the mass of 15 individual frogs before and after a cold season.
 - e. You compare plant height across five fertiliser treatments (12 plants per treatment).
- After classifying each, check your answers against the grid above.

DO IT NOW!

Look at the five scenarios listed in the previous. For each one, identify the most likely assumption-checking step you would perform before running the test you selected. Then, for scenario (b) (fish survival, binary response), explain why no amount of checking for normality within groups will help, and what you should do instead.

As a final synthesis exercise, take any dataset you have been working with this semester and classify it using the two questions at the start of this chapter. Map it to the grid, identify the appropriate method, and briefly note which assumption you would check first.

10.5 CONCLUSION

The purpose of this chapter is to show that a small number of design and data questions do most of the work in narrowing the choice of method. Once you know what the response is, how the explanatory structure is organised, and whether the observations are independent, the inferential landscape becomes much easier to navigate.

This also marks a transition in the course. The chapters so far have focused on named inferential procedures for common question types. From the next chapter onward, I move into models as the main analytical framework. Instead of choosing among isolated tests, I begin to think in terms of fitted mean structure, residuals, diagnostics, and model revision.

Part III

Modelling

In this part, I move from standalone inferential procedures to explicit statistical models. The conceptual step is important: rather than asking only whether groups differ or variables are associated, I begin to describe how a response varies with predictors, what functional form that relationship takes, and whether the fitted model is an adequate representation of the data.

This is where residuals, regression equations, model specification, interaction terms, confounding, and model evaluation become central. The methods are more flexible than the earlier tests, but they also demand more discipline in how the model is formulated and checked.

Read this part as a continuation of inference rather than a departure from it. The same statistical reasoning remains in place; the model simply becomes the main unit of thought.

11

Residuals and Model-Based Diagnostics

Checking Fitted Models Before Interpreting Them

IN THIS CHAPTER

- the shift from raw-data checks to model-based checks;
- what a fitted model, a predicted value, an error, and a residual are;
- why regression assumptions apply to residuals;
- how to read residual-versus-fitted and residual Q-Q plots;
- how diagnostic patterns guide model revision.

TASKS TO COMPLETE IN THIS CHAPTER

- None

In earlier chapters, I often began with raw-data displays and simple checks on the quantities being analysed. In [Chapter 6](#), for example, I asked whether values within groups, paired differences, or observed associations were well behaved enough for the planned test. That remains a sensible starting point for simple procedures such as *t*-tests.

ANOVA already sits partly inside the linear-model framework, even if I introduced it earlier through grouped means and exploratory plots. For ANOVA, as for regression more generally, the formal assumptions concern the model errors and are therefore checked most directly through residuals from the fitted model.

Regression changes the object being checked. A regression model splits the response into a systematic component that the model tries to explain and an error component that remains unexplained. After fitting the model, I inspect the residuals because they are the observed estimates of that remaining error.

In this chapter, I first define the model structure, fitted values, errors, and residuals. I then map the main regression assumptions to the plots or design checks that address them. I end by showing how those diagnostics guide model revision in later chapters.

11.1 MODELS, ERRORS, AND FITTED VALUES

A **model** is a mathematical description of how a response changes with one or more predictors.

A **response variable** is the outcome being explained or predicted. Sometimes it is called the outcome or dependent variable.

A **predictor variable** is a measured quantity used to explain or predict the response. It is also known as the independent variable.

A **fitted model** is the version of the model after its unknown quantities have been estimated from the data.

A **predicted value** is the value the fitted model expects for the response at a given predictor value.

For a simple linear regression, I can write the model for observation i as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The term $\beta_0 + \beta_1 X_i$ is the **systematic component** or **mean structure**. It gives the conditional mean $E[Y_i | X_i]$ when the model is appropriate. The term ε_i is the **error**. It is the unobserved departure of the observed response from that conditional mean.

This decomposition is important because two different parts of the model are being checked. Linearity concerns the mean structure, while constant variance concerns the spread of the errors around that mean structure.

Consider a biological question: how does plant height change with light availability? Here, plant height is the response and light availability is the predictor. A fitted model would describe the expected plant height at each light level. The predicted value is therefore the height that the fitted model gives for a plant growing under a specific amount of light.

The purpose of this chapter is explain what is being checked once a model has been fitted, and in the next chapter I fit the straight-line model and assess its fit.

11.2 RESIDUALS

A **residual** is the difference between an observed value and its predicted value:

$$\text{residual} = \text{observed value} - \text{predicted value}$$

Residuals are observed quantities. Errors are theoretical quantities in the model and are not observed directly. Residuals estimate those unobserved errors and are used for diagnostic checks.

In practical terms, residuals carry information about that remaining variation, which may arise from omitted predictors, measurement error, natural biological variation, or a model form that is too simple for the pattern in the data.

For the real-data sections of this chapter, I will use the penguins dataset from the palmerpenguins package. After removing missing values, I will treat body mass (body_mass_g) as the response and flipper length (flipper_length_mm) as the predictor.

```
data("penguins", package = "palmerpenguins")

penguins_tbl <- as_tibble(penguins) >
  drop_na(flipper_length_mm, body_mass_g) >
  mutate(obs_id = row_number())

model <- lm(body_mass_g ~ flipper_length_mm, data = penguins_tbl)

penguin_resids <- penguins_tbl >
  mutate(
    predicted_mass = fitted(model),
    residual = residuals(model)
  )

penguin_resids >
  select(obs_id, species, flipper_length_mm, body_mass_g, predicted_mass, residual) >
  slice_head(n = 8) >
  knitr::kable(
    digits = 2,
    col.names = c("Observation", "Species", "Flipper length", "Observed body mass",
                  "Predicted body mass", "Residual")
  )
```

Observation	Species	Flipper length	Observed body mass	Predicted body mass	Residual
1	Adelie	181	3750	3212.26	537.74
2	Adelie	186	3800	3460.68	339.32
3	Adelie	195	3250	3907.85	-657.85
4	Adelie	193	3450	3808.48	-358.48
5	Adelie	190	3650	3659.43	-9.43
6	Adelie	181	3625	3212.26	412.74

Observation	Species	Flipper length	Observed body mass	Predicted body mass	Residual
7	Adelie	195	4675	3907.85	767.15
8	Adelie	193	3475	3808.48	-333.48

Each row in the table shows the same calculation. The fitted model predicts body mass from flipper length and the residual records the remaining difference between the observed and predicted body mass.

The steps to extract the predicted values and residuals are:

```
model <- lm(body_mass_g ~ flipper_length_mm, data = penguins_tbl)
fitted(model) # predicted values
residuals(model) # raw residuals
```

Throughout this chapter, I use raw residuals because the goal is to establish the basic thinking behind the diagnostics. Later chapters introduce standardised and studentised residuals when I discuss unusual and influential observations.

DO IT NOW!

Look at the table printed above. For observations 1 and 4, verify by hand that $\text{residual} = \text{observed body mass} - \text{predicted body mass}$. Then answer: which of the eight penguins shown is most strongly over-predicted by the model, and which is most strongly under-predicted?

11.3 WHY ASSUMPTIONS APPLY TO RESIDUALS

A fitted model aims to describe the systematic part of the relationship between a response and its predictors. Fitting a linear model removes the conditional mean $E[Y | X]$, leaving residuals centred at zero.

I can show that visually with a small simulated example. In panel A of Figure 11.1, the response increases with the predictor and the fitted line has a clear positive slope. In panel B, I plot the residuals from that fitted model against the same predictor. The slope has been removed, so the fitted line is now horizontal around zero. The residuals are therefore the remainder after fitting, which is why they become the object of diagnosis.

Fitting a linear model therefore amounts to subtracting the fitted mean trend from the data and then asking whether any structure remains.

```
set.seed(74411)
sim_dat <- tibble(
  light = seq(1, 10, length.out = 100),
```



```

height = 5 + 1.8 * light + rnorm(100, sd = 2)
)

sim_mod <- lm(height ~ light, data = sim_dat)

sim_plot_dat <- bind_rows(
  sim_dat >
    transmute(light, value = height, panel = "A. Observed Data"),
  sim_dat >
    transmute(light, value = residuals(sim_mod), panel = "B. Residuals After Fitting")
)

zero_line <- tibble(
  panel = "B. Residuals After Fitting",
  yint = 0
)

ggplot(sim_plot_dat, aes(x = light, y = value)) +
  geom_hline(
    data = zero_line,
    aes(yintercept = yint),
    inherit.aes = FALSE,
    colour = "grey40",
    linewidth = 0.5
  ) +
  geom_point(size = 1.9, colour = "steelblue4") +
  geom_smooth(method = "lm", se = FALSE, colour = "firebrick", linewidth = 0.8) +
  facet_wrap(~panel, scales = "free_y") +
  labs(
    x = "Predictor value",
    y = "Response value"
  )

```

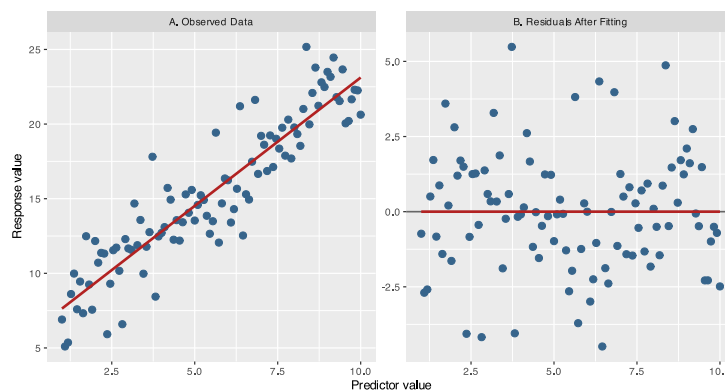


FIGURE 11.1. The systematic part of a relationship is removed when residuals are extracted. Panel A shows sloped scattered data with a fitted line. Panel B shows the residuals from that same fitted model against the predictor. The fitted line is now horizontal because the linear trend has been removed.

Assumptions therefore apply to the residuals because the residuals represent the unexplained variation. If the residuals still show a pattern, the model has left structure behind, and if they have unstable spread, the uncertainty estimates become less reliable.

DO IT NOW!

Reproduce Figure 11.1 for a different simulated dataset. Use a steeper slope and more noise:

```
set.seed(99)
my_dat <- tibble(
  x = seq(1, 10, length.out = 80),
  y = 2 + 3.5 * x + rnorm(80, sd = 5)
)
my_mod <- lm(y ~ x, data = my_dat)
```

Plot the raw data with the fitted line (panel A) and then plot the residuals against x (panel B). Confirm that the residuals are centred on zero and that no clear slope remains. Discuss with a partner why the residual plot looks the way it does even though the original relationship was strongly positive.

Context	What is being tested or checked
t -test	whether the relevant raw values within groups, or the paired differences, are well behaved enough for a mean comparison
paired design	whether the within-pair differences are well behaved enough for the paired analysis
correlation	whether the observed association between the two measured variables is compatible with the intended correlation analysis
ANOVA	whether the <i>residuals</i> from the fitted ANOVA model are well behaved enough for the ANOVA inference, with grouped plots used as preliminary checks
regression	whether the <i>residuals</i> from the fitted model are well behaved enough for the regression inference

11.4 MAIN REGRESSION ASSUMPTIONS

11.4.1 *Linearity*

Linearity means that the fitted model has captured the mean relationship between the predictor and the response with an appropriate straight-line form. In a simple linear regression, the residuals should not retain a sloped or curved pattern after the line has been fitted.

11.4.2 *Constant Variance*

Constant variance means that the spread of the residuals around zero stays broadly similar across the fitted range. If the spread expands or contracts systematically, the variance structure is not being described well by the model.

DO IT NOW!

The table below maps assumptions to the plots that check them. Without looking at the rest of the chapter, try to fill in the “Visual cue” column from memory or first principles. Then read on and compare your answers with the table in the next section.

Assumption	What it concerns	Main check	Your predicted visual cue
Linearity	mean structure	residual vs fitted	?
Constant variance	spread	residual vs fitted	?
Normality	residual distribution	Q-Q plot	?

Discuss: which of these three would you check first, and why?

11.4.3 *Independence*

Independence means that one residual does not carry information about another. This requirement comes mainly from the sampling or experimental design. Repeated measures on the same individual, temporal correlation, spatial clustering, and nested sampling can all violate independence.

When dependence is plausible, I do not rely on the usual residual-versus-fitted or Q-Q plots alone. I also inspect residuals in the order in which the data were collected. For time-series or repeated-measures data, that often means plotting residuals against time and looking for runs, cycles, or long stretches of positive or negative re-

iduals. For spatial data, it may mean plotting residuals against location and looking for clusters of similar residuals.

11.4.4 Normality of Residuals

Approximate normality of residuals means that the residual distribution is reasonably close to a normal distribution. This is most important for standard errors, confidence intervals, and tests in smaller samples. Mean structure, variance structure, and independence usually deserve priority because large departures there often explain what later appears in the residual distribution.

The main checks are thus performed in this order:

Assumption	What it concerns	Main check	Visual cue
Linearity	mean structure $E[Y X]$	residual-versus-fitted plot	slope or curvature in residuals
Constant variance	spread around the mean structure	residual-versus-fitted plot	funnel shape or changing vertical spread
Normality	distribution of residuals	residual Q-Q plot	systematic bends away from the line
Independence	relationship among residuals across sampling units	design plus residuals versus order, time, or space	runs, cycles, or spatial clusters

11.5 DIAGNOSTIC PLOTS

A **diagnostic plot** is a graph used to check whether the fitted model has left behind problematic structure in the residuals.

DIAGNOSTIC WORKFLOW

1. Fit the model.
2. Plot residuals against fitted values to check mean structure and spread.
3. Plot a residual Q-Q plot to check whether the residual distribution is close enough to normal for the intended inference.
4. Revisit the sampling or experimental design, and where relevant plot residuals against time, space, or sampling order to check independence.
5. Revise the model if the diagnostics show leftover structure.

These plots are not decoration. They show where the model is still wrong and suggest what to revise next.

11.5.1 *Residual versus Fitted Plot*

The **residual-versus-fitted plot** places fitted values on the x-axis and residuals on the y-axis.

I read this plot by asking two questions. Do the residuals scatter around zero without a systematic pattern? Does their spread stay broadly similar across the fitted range? Random scatter around zero supports the fitted mean structure. Curvature suggests that the straight-line form is missing part of the relationship. A funnel shape suggests changing variance.

11.5.2 *Q-Q Plot of Residuals*

The **Q-Q plot of residuals** compares the ordered residuals with the values expected from a normal distribution.

Points that follow the reference line support approximate normality. In practice, the centre of the plot is usually more informative than one slightly wayward tail point. If most points track the line and only the most extreme one or two points drift modestly away, I usually treat that as acceptable, especially in a moderate sample. I become more concerned when many points form a clear S-shape, when both tails peel away strongly from the line, or when one or two extreme residuals dominate the ends of the plot. At that stage, the next step is to inspect the raw data, check for influential observations, and ask whether the response scale or response distribution is appropriate.

The standard base-R diagnostic commands are:

```
plot(model, which = 1) # residual versus fitted
plot(model, which = 2) # Q-Q plot
```

DO IT NOW!

Simulate a dataset that deliberately violates constant variance (heteroscedasticity) and plot the residual-versus-fitted plot to see what that violation looks like:

```
set.seed(42)
n <- 100
x <- seq(1, 10, length.out = n)
y <- 3 + 1.5 * x + rnorm(n, sd = 0.4 * x) # SD grows with x
het_mod <- lm(y ~ x)
```

Describe the shape you see. How does it differ from the ideal plot shown next in Figure 11.2?

11.5.3 Example 1: Idealised Diagnostic Plots

I continue with the same simulated dataset used in Figure 11.1. Here that same fitted model is viewed through the two standard diagnostic plots. In Figure 11.2, the residual spread is even and the smooth line stays close to horizontal. In Figure 11.3, the residuals are also approximately normal.

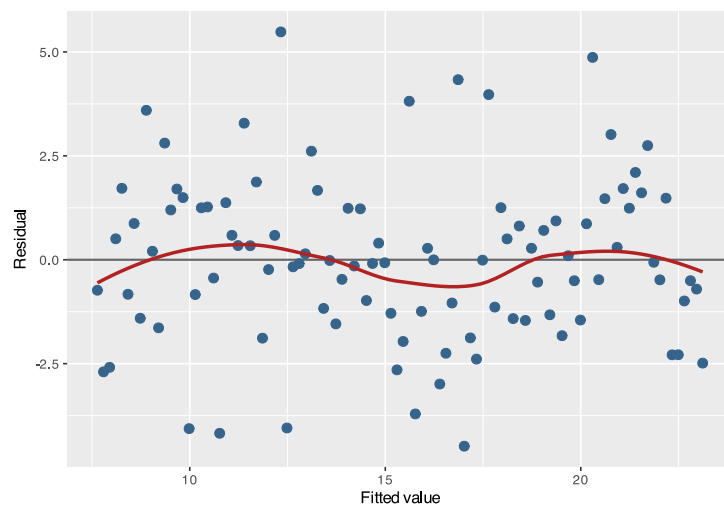


FIGURE 11.2. Idealised residual-versus-fitted plot for a well-behaved straight-line model.

In Figure 11.2, the points form a roughly even cloud around zero. The smooth line stays close to horizontal, so there is no clear leftover mean structure. The vertical spread is also fairly similar across the fitted range, so there is no strong sign of changing variance.

In Figure 11.3, most points lie close to the reference line. That is the pattern expected when the residuals are close enough to normal for standard linear-model inference.

Together, these plots show what I want to see before I interpret a fitted model. Real data rarely look this tidy, which is why the next example is more revealing.

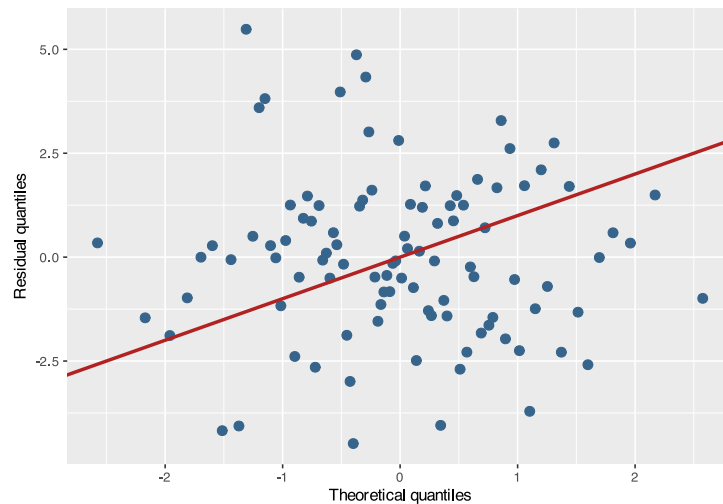


FIGURE 11.3. Idealised Q-Q plot of residuals for a well-behaved straight-line model.

11.6 EXAMPLE 2: WORKED EXAMPLE WITH REAL DATA

I now return to the penguins data. I fit body mass as a function of flipper length using all penguins together. That is a plausible first attempt because larger penguins usually have longer flippers, but it also ignores the fact that the dataset contains different species.

```
model <- lm(body_mass_g ~ flipper_length_mm, data = penguins_tbl)

penguin_aug <- penguins_tbl >
  mutate(
    .fitted = fitted(model),
    .resid = residuals(model)
  )

penguin_aug >
  select(species, flipper_length_mm, body_mass_g, .fitted, .resid) >
  slice_head(n = 6)
```

```
R> # A tibble: 6 x 5
R>   species flipper_length_mm body_mass_g .fitted .resid
R>   <fct>      <int>          <int>    <dbl>  <dbl>
R> 1 Adelie         181          3750   3212.   538.
R> 2 Adelie         186          3800   3461.   339.
R> 3 Adelie         195          3250   3908.  -658.
R> 4 Adelie         193          3450   3808. -358.
R> 5 Adelie         190          3650   3659.   -9.43
R> 6 Adelie         181          3625   3212.   413.
```

The fitted values are the model's expected body masses. The residuals show how far the observed body masses depart from those fitted values.

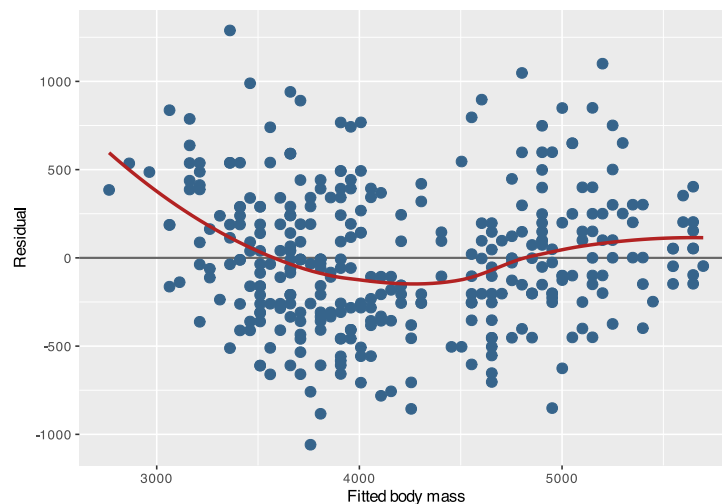


FIGURE 11.4. Residuals versus fitted values for a pooled penguin body-mass model.

Something seems “off” in Figure 11.4. The residuals do not form a roughly random cloud around zero. Instead, the smooth line bends from positive residuals at low fitted masses to negative residuals in the middle and back toward positive residuals at the upper end. That is evidence that one straight line is not describing the mean structure adequately.

Here the likely cause is not an exotic response distribution but pooled biological groups. Adelie, Chinstrap, and Gentoo penguins differ in body size, so one line fitted to all species averages over distinct structures. I have ignored the grouping structure. The next step would therefore be that I revise the model specification, for example by including species as a predictor or by allowing different lines for different species.

DO THIS NOW!

Plot body mass against flipper length for the penguins dataset and colour the points by species. Use that graph to check whether the residual pattern in Figure 11.4 could be explained by pooled group structure rather than by a single curved relationship.

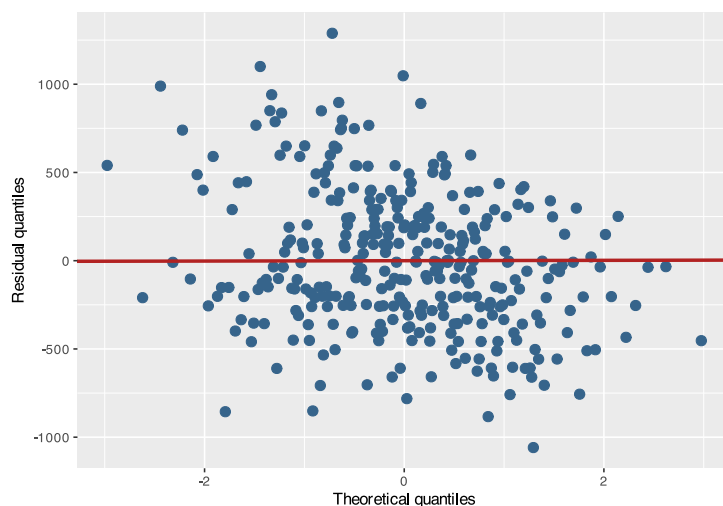


FIGURE 11.5. Q-Q plot of residuals for a pooled penguin body-mass model.

DO IT NOW!

Refit the penguin model including `species` as an additional predictor and re-examine both diagnostic plots:

```
model_species <- lm(body_mass_g ~ flipper_length_mm + species,
  data = penguins_tbl)

aug_species <- penguins_tbl >
  mutate(.fitted = fitted(model_species),
         .resid = residuals(model_species))
```

Does the curved pattern in Figure 11.4 disappear once `species` is included? What does this tell you about the source of the earlier diagnostic problem?

In addition to your residual-vs-fitted plot, also produce a Q-Q plot of the residuals for the new model and compare it with Figure 11.5.

Above, I am preempting some later work. See Chapter 14 before returning here, if necessary.

In Figure 11.5, the residuals are close enough to normal that I would not treat distributional shape as the first problem. Approximate normality does not rescue a poor mean model due to ignoring the grouping structure, so my primary warning comes from Figure 11.4, not from the Q-Q plot.

11.7 LINKING DIAGNOSTICS TO DECISIONS

The penguins example shows why diagnostics should lead directly to model revision. The residual-versus-fitted plot does not suggest a small tweak to an otherwise adequate straight line. It suggests that I have fitted the wrong model because I ignored a biologically relevant grouping variable. In that case, the most sensible correction is to change the model specification by including species, by allowing species-specific slopes, or by fitting separate lines.

That same logic extends beyond this example. When residuals show curvature, the fitted mean structure is incomplete. The correction may be to add an omitted predictor, include an interaction, transform a predictor or the response, add a polynomial term, use a smoother, or move to a nonlinear model whose form matches the biology more closely.

When the residual spread changes across the fitted range, the variance structure is being described poorly. Possible corrections include transforming the response, modelling the variance explicitly, using weighted regression, or choosing a model family whose mean-variance relationship is more appropriate for the data.

When the Q-Q plot shows strong tail departures, the next step is usually to inspect the data rather than to panic about normality in the abstract. Extreme residuals may reflect outliers, influential observations, an awkward response scale, or a response distribution that is not well handled by a Gaussian model.

When residuals are dependent, the solution is rarely a cosmetic adjustment to the same model. Dependence points back to the design and often requires a different modelling framework altogether, such as repeated-measures methods, time-series models, spatial models, or mixed models.

A diagnostic plot therefore points to the aspect of the model that needs revision and helps decide what kind of revision is defensible.

DO IT NOW!

Work through the four diagnostic scenarios described below. For each one, decide which assumption is most likely violated and what the preferred next step is. Discuss your reasoning with a partner before reading further.

Scenario	What the diagnostic shows	Which assumption?	Next step?
A	Residuals fan out strongly from left to right in the residual-vs-fitted plot	?	?
B	Q-Q plot shows a pronounced S-bend, with both tails lifting away from the line	?	?
C	Residuals form a clear arc (positive, then negative, then positive) across the fitted range	?	?
D	Residuals plotted against collection order show long runs of positive values followed by long runs of negative values	?	?

After completing the table, check your answers against the “Linking Diagnostics to Decisions” section above.

11.8 CONNECTION TO EARLIER CHAPTERS

This chapter extends the checking habits developed in [Chapter 6](#). Grouped comparisons checked variation within groups. Paired analyses checked variation in the within-pair differences. Correlation inspected the observed joint pattern directly. Regression keeps the same questions about structure, spread, and dependence, but it asks them after the model has already been fitted.

11.9 FORWARD LINKS

Diagnostics will emerge again in:

- [Chapter 12: Simple Linear Regression](#) fits and interprets the first full regression models.
- [Chapter 14: Multiple Regression and Model Specification](#) applies the same diagnostic ideas when several predictors appear in one model.
- [Chapter 17: Model Checking and Evaluation](#) develops residual analysis, leverage, influence, and model comparison in more detail.
- [Chapter 19: Dependence and Mixed Models](#) takes up the cases where independence fails because observations are clustered or repeated.

11.10 SUMMARY

Regression starts from the decomposition

$$\text{response} = \text{systematic component} + \text{error}$$

After fitting the model, I inspect residuals because they estimate the unobserved errors.

Residual-versus-fitted plots check the mean structure and the variance structure. Q-Q plots check whether the residual distribution is close enough to normal for the intended inference. Independence comes mainly from the design and, where relevant, from residual patterns across time, space, or sampling order.

The purpose of these diagnostics is not description for its own sake. They tell me whether the current model is adequate or whether I need to revise it before interpreting coefficients or testing hypotheses.

12

Simple Linear Regression

The Entry Point to Model-Based Biostatistics

IN THIS CHAPTER

- what a simple linear regression model is;
- when regression is more appropriate than correlation;
- the assumptions behind a straight-line model;
- how to fit a model with `lm()`;
- how to diagnose normality, homoscedasticity, linearity, and outliers;
- how to interpret slopes, fitted values, confidence intervals, and prediction intervals;
- how explanatory and predictive uses of the same regression differ;
- how to report a regression in the style of a Results section.

TASKS TO COMPLETE IN THIS CHAPTER

- Self-Assessment Task 12-1 (/10)
- Self-Assessment Task 12-2 (/30)
- [Self-Assessment instructions and full task overview](#)

Linear models are among the most useful statistical tools available to biologists because they describe and quantify how a **response** variable, Y , changes as one or more **predictor** variables, X , change. In statistics, a model is a mathematical representation of a real process. It is not reality itself, but an idealised description of the part of reality we want to understand. Linear models are especially valuable because they are simple enough to interpret, yet flexible enough to support a large part of modern statistical practice.

In the broadest sense, a linear model is one in which the unknown parameters enter linearly, even if the variables themselves are transformed or combined in more elaborate ways. The most basic member of that family is the **simple linear model**, which is the focus of this chapter. It has one continuous predictor and one continuous response, and it is fitted by **simple linear regression**. The goal may be *explanatory*, where the predictor is hypothesised to influence the response, or *predictive*, where I simply want a formula that estimates likely values of the response from observed values of the predictor. A causal interpretation is therefore common, but it is not required.

Regression analysis is the procedure by which I estimate the model parameters from data. The aim is to fit the model that best captures the observed response-predictor relationship and then interpret the strength, direction, and uncertainty of that relationship. This differs from **correlation**, which quantifies association without imposing a response-predictor structure. When one variable changes systematically with another but neither prediction nor a defensible response-predictor distinction is of interest, correlation is usually the more appropriate tool.

As established in **Chapter 11**, every observed response can be written as $Y_i = \hat{Y}_i + e_i$, where $\hat{Y}_i$ is the fitted value and e_i is the residual. In the previous chapter I focused on residuals, fitted values, and the diagnostic thinking used to assess whether a model is behaving adequately. Here I carry that thinking into the first full regression model in the course and show how the equation, the slope, the fitted line, and the diagnostics all belong to one workflow.

Simple linear regression is therefore the entry point to the wider model-based framework. From here I move to **polynomial regression**, where curvature is handled within the linear-model framework; to **multiple regression**, where several predictors act simultaneously; to **interaction terms**, where the effect of one predictor depends on another; and later to **generalised linear models**, where the same modelling thinking is extended to non-normal responses by introducing a link function and a different error distribution. If the response-predictor structure is defensible and a straight-line mean relationship is biologically plausible, simple linear regression is usually the correct place to begin.

12.1 MAIN CONCEPTS

These ideas organise the chapter.

- **Simple linear regression** models one continuous response as a function of one continuous predictor.
- **The response-predictor distinction** is essential because it is not simply putting a line onto a correlation scatter plot.

- **The slope** is usually the main inferential quantity because it describes the expected change in the response for a one-unit change in the predictor.
- **The intercept** is often less biologically interesting, but it is still part of the fitted model.
- **Residuals** are central to assumption checking because they reveal structure that the model has failed to capture.
- **Confidence intervals** and **prediction intervals** answer different questions and should not be confused.

12.2 NATURE OF THE DATA AND ASSUMPTIONS

12.2.1 *Requirements Before Fitting*

As the experimenter, you must ensure the following requirements *before* a simple linear regression is fitted:

1. **A defensible response-predictor structure:** There should be a theoretical or philosophical basis for treating one variable as the predictor and the other as the response. This may be explicitly causal, but it can also be predictive if that distinction is still biologically sensible.
2. **Independence of observations:** Each measured value of the response must be independent of the others. If repeated measurements, clustered sampling, or temporal dependence are present, a different modelling framework may be required.
When temporal or spatial order is plausible, plot residuals against the order in which observations were collected. Runs of positive or negative residuals, or systematic cycles, indicate dependence that the model has not captured. For the sparrow and penguin examples you will encounter below, measurements are treated as independent, so this check is not performed here.
3. **Continuous predictor:** The predictor variable should be continuous.
4. **Continuous response:** The response variable should also be continuous.

12.2.2 *Assumptions to Check After Fitting*

After the model has been fitted, the following assumptions must be checked:

1. **Normality:** The residuals e_i should be approximately normally distributed.
2. **Homoscedasticity:** The variance of the residuals e_i should be roughly constant across the fitted values.
3. **Linearity:** The mean relationship between the predictor and the response should be approximately linear.

4. **Measurement error in the predictor:** Standard linear regression assumes that the predictor is measured without serious error. In practice this is only approximately true, and we return to this issue in [Chapter 16](#).

As in the earlier inferential chapters, you must pay attention to the workflow. We first inspect the data (EDA and graphically), then fit the model, then examine the residuals (graphically first, then assumption tests if desired), and *only then* interpret the coefficients with confidence.

12.2.3 Assumption and Diagnostics

The table below summarises which diagnostic tool targets which assumption and what pattern to watch for.

Assumption	What it concerns	Main diagnostic	Visual signal to watch for
Linearity	mean structure	residual vs fitted plot	slope or curvature in residuals
Constant variance	spread	residual vs fitted / scale-location	funnel or changing vertical spread
Normality	residual distribution	residual Q-Q plot	systematic bends away from reference line
Independence	relationship among residuals	residuals vs order/time/space	runs, cycles, clusters

12.3 THE MODEL

Simple linear regression is the first statistical method in the work we have done thus far in which we write an explicit equation (the model) for the mean response and then estimate its parameters from data.

The model is:

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad (12.1)$$

In Equation 12.1, Y_i is the response for observation i , X_i is the predictor, α is the intercept, β is the slope, and ϵ_i is the error term. Errors ϵ_i are unobserved theoretical quantities. Residuals e_i are their observed estimates, computed after fitting, and are the objects used for diagnostic checks.

The line is fitted by minimising the sum of squared residuals. This is why ordinary linear regression is often called an **ordinary least squares** method.

The animation below shows the fitted line rotating through the data as the error sum of squares is minimised.

ERRORS AND RESIDUALS

The model contains errors ϵ_i , which are the unobserved theoretical deviations of each true response from the straight-line mean. In most regression models we assume that the errors are independent and identically distributed (i.i.d.). When the errors are approximately normal this can be written as $\epsilon_i \sim N(0, \sigma^2)$. The requirement of mean zero implies that, on average, the model does not systematically over- or under-predict. Constant variance implies that the spread of errors is roughly similar across the predictor range. Independence implies that errors do not carry systematic structure from one observation to the next.

The residuals $e_i = Y_i - \hat{Y}_i$ are the observed estimates of these unobserved errors. It is the residuals that are available after fitting and that we inspect in diagnostic plots.

Violation of these assumptions can lead to biased or inefficient parameter estimates, poor uncertainty estimates, and misleading inference.

The next section formalises how the line is chosen from the data.

DO IT NOW!

Before fitting any model, take a moment to check whether simple linear regression is the right tool. For each scenario below, decide whether the method is appropriate and identify any requirement that is not met:

Scenario	Appropriate?	Which requirement, if any, fails?
Plant height (continuous) modelled as a function of rainfall (continuous); one measurement per plant	?	?
Blood pressure measured five times on each of 20 patients; modelled as a function of time	?	?
Species presence/absence modelled as a function of temperature	?	?

Tree diameter ? ?
 (continuous) modelled as
 a function of stand density
 (continuous); trees within
 the same plot share
 resources

Discuss your answers with a partner and then read on to check them against the requirements listed above.

12.4 THE FITTING RULE

The least-squares criterion used by `lm()` is to choose α and β so that the residual sum of squares is as small as possible:

$$\text{RSS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (12.2)$$

Equation 12.2 is not a regression model. It is the fitting rule that tells us how the software decides which of all possible straight lines is the best-fitting one.

12.5 R FUNCTION

The main function used in this chapter is `lm()`:

```
lm(response ~ predictor, data = df)
```

You can read the formula `response ~ predictor` as “the response is modelled as a function of the predictor.”

The fitted model can then be explored with functions such as:

- `summary()` for the coefficients and overall fit;
- `confint()` for confidence intervals around the coefficients;
- `augment()` from **broom** for fitted values and residuals;
- `predict()` for confidence and prediction intervals;
- `plot()` for standard diagnostic plots;
- `bptest()` from **lmtest** for a formal test of heteroscedasticity.

DO IT NOW!

Fit the sparrow model by hand before running `lm()`. The least-squares slope is $\hat{\beta} = r \cdot (s_Y/s_X)$ and the intercept is $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$.

```
sparrows <- tibble(
  age = c(3, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17),
  wing = c(1.4, 1.5, 2.2, 2.4, 3.1, 3.2, 3.2, 3.9, 4.1, 4.7, 4.5, 5.2, 5.0)
)
```

Then verify your hand-calculated values against `lm(wing ~ age, data = sparrows)`. Do they match? Explain in one sentence why the slope formula involves both the correlation and the ratio of standard deviations.

12.6 OUTLIERS AND THEIR IMPACT ON SIMPLE LINEAR REGRESSION

Outliers are data points that deviate substantially from the overall pattern or trend observed in the data. They can have disproportionate effects on a simple linear regression because the fitted line is estimated by minimising squared residuals, as in Equation 12.2. Extreme observations may therefore influence the slope, the intercept, the standard errors, the confidence intervals, and the diagnostic patterns.

This does not mean that unusual observations must automatically be removed. Some are recording errors and should be corrected or excluded. Others are rare but real biological events and may carry important information. The correct response is therefore to identify potentially influential points, inspect them carefully, and decide whether they reveal error, unusual but valid biology, or a more fundamental model problem.

12.7 EXAMPLE 1: SPARROW WING LENGTH AND AGE**DIAGNOSTIC WORKFLOW FOR SIMPLE LINEAR REGRESSION**

1. Fit the model with `lm()`.
2. Plot residuals versus fitted values then check for curvature (linearity) and changing spread (constant variance).
3. Plot the residual Q-Q plot and check whether the residual distribution is close enough to normal.
4. If temporal or spatial order is plausible, plot residuals against order and check for runs or cycles (independence).
5. Inspect Cook's distance and leverage and identify observations with disproportionate influence.
6. Revise the model if any diagnostic reveals leftover structure.

TABLE 12.1. Sparrow wing-length data used in the introductory simple linear regression example.

Age (days)	Wing length (cm)
3	1.4
4	1.5
5	2.2
6	2.4
8	3.1
9	3.2
10	3.2
11	3.9
12	4.1
14	4.7
15	4.5
16	5.2
17	5.0

I begin with a very small sparrow dataset because it makes the general approach clear. I then go to a fuller worked example using the Adelie penguin data from the `palmerpenguins` package, which is much closer to the style and level of complexity encountered in real biological analyses.

DO IT NOW!

Before looking at the sparrow scatter plot, sketch what you expect the relationship between age and wing length to look like based on your biological knowledge alone. Then run the code below and compare your sketch with the actual data:

```
ggplot(sparrows, aes(x = age, y = wing)) +
  geom_point(shape = 1, colour = "indianred") +
  labs(x = "Age (days)", y = "Wing length (cm)")
```

Does a straight line look like a reasonable summary of this relationship? Are there any obvious outliers or influential-looking points at the extremes of the age range?

12.7.1 Do an Exploratory Data Analysis (EDA)

The sparrow data show the basic form of a simple linear model.

```
summary(sparrows)
```

```
R>      age      wing
```

```

R> Min.   : 3   Min.   :1.400
R> 1st Qu.: 6   1st Qu.:2.400
R> Median :10   Median :3.200
R> Mean   :10   Mean   :3.415
R> 3rd Qu.:14   3rd Qu.:4.500
R> Max.   :17   Max.   :5.200

```

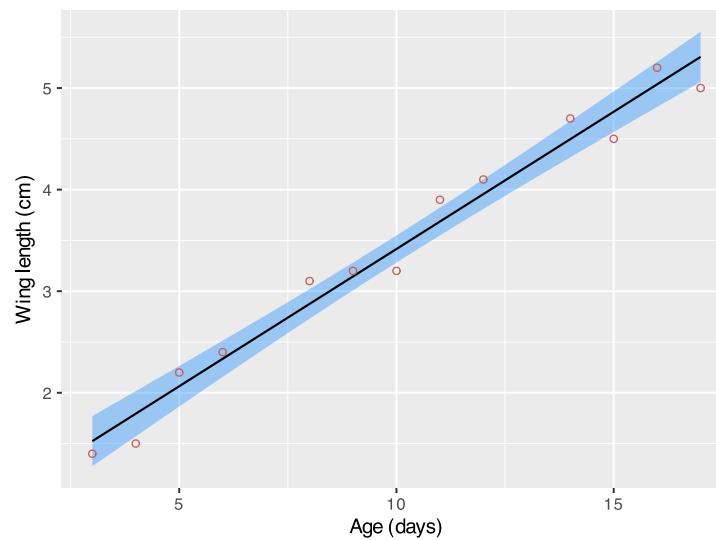


FIGURE 12.1. Wing length of sparrows at different ages.

In Figure 12.1, the scatter plot suggests a clear positive linear relationship: older sparrows tend to have longer wings, and the relationship appears close to linear over the range of the data. This example makes the fitted line and the slope easy to understand before we look at a noisier dataset.

12.7.2 State the Model Question and Hypothesis

With the sparrow example, I ask whether wing length changes systematically with age.

The statistic of interest in a simple linear regression is usually the slope in Equation 12.1, which quantifies the magnitude and direction of dependence of the response on the predictor:

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

If the slope is zero (a more-or-less horizontal line), there is no linear relationship between the predictor and the expected value of the response. If the slope differs from zero, then the predictor helps explain variation in the response.

12.7.3 Fit the Model

```
sparrow_mod <- lm(wing ~ age, data = sparrows)
summary(sparrow_mod)
```

```
R>
R> Call:
R> lm(formula = wing ~ age, data = sparrows)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.30699 -0.21538  0.06553  0.16324  0.22507
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.71309    0.14790   4.821 0.000535 ***
R> age          0.27023    0.01349  20.027 5.27e-10 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.2184 on 11 degrees of freedom
R> Multiple R-squared:  0.9733, Adjusted R-squared:  0.9709
R> F-statistic: 401.1 on 1 and 11 DF,  p-value: 5.267e-10
```

The output provides the intercept, the slope, their standard errors, a test of whether the coefficients differ from zero, the residual standard error, and the model R^2 .

SELF-ASSESSMENT TASK 12-1

- Examine the contents of the regression model object `sparrow_mod`. Explain the main components and how they relate to `summary(sparrow_mod)`. (/3)
- Using values inside the model object, show how to reconstruct the observed response values from the fitted values and residuals. (/3)
- Fit a linear regression through the model residuals and explain the result. (/2)
- Fit a linear regression through the fitted values and explain the result. (/2)

DO IT NOW!

Read the `summary()` output for `sparrow_mod` carefully. Without using a calculator, answer the following questions from the printed output:

- What is the estimated slope, and what does it tell you about the expected change in wing length per additional day of age?
- What is the p -value for the slope, and what conclusion do you draw about $H_0 : \beta = 0$?

3. The residual standard error is printed near the bottom of the summary. What does this number represent in the units of the response variable?
4. What does $R^2 \approx 0.97$ tell you about how much of the variation in wing length the model accounts for?

Discuss your answers with a partner before continuing.

12.7.4 Test the Assumptions

Assumptions in regression are checked after fitting the model.

```
sparrow_aug <- augment(sparrow_mod)
sparrow_aug >
  select(age, wing, .fitted, .resid) >
  head()
```

```
R> # A tibble: 6 x 4
R>   age  wing .fitted .resid
R>   <dbl> <dbl>   <dbl>   <dbl>
R> 1     3   1.4     1.52 -0.124
R> 2     4   1.5     1.79 -0.294
R> 3     5   2.2     2.06  0.136
R> 4     6   2.4     2.33  0.0655
R> 5     8   3.1     2.87  0.225
R> 6     9   3.2     3.15  0.0548
```

a) Normality. The Q-Q panel in Figure 12.2 shows the sample quantiles of the residuals plotted against the theoretical quantiles of a normal distribution. For these data the points track close to the reference line, indicating that the residuals do not depart strongly from normality.

b) Homoscedasticity. The residuals-versus-fitted panel in Figure 12.2 does not reveal a systematic funnel or wedge pattern. The spread of residuals appears reasonably even across the range of fitted values, consistent with the constant-variance assumption.

c) Influential observations. Because the dataset contains only 13 observations, formal influence diagnostics have limited statistical power. No single observation stands out dramatically in the scale-location or Cook's-distance panels, but conclusions based on such a small sample should be interpreted cautiously.

In Figure 12.2, the diagnostic plots suggest that the model is broadly adequate for these data. The residuals do not show severe curvature, the spread is reasonably even, and the Q-Q plot does not suggest a dramatic departure from normality. Because the dataset is small, these plots should be interpreted cautiously, but there is no obvious reason to abandon the linear model.

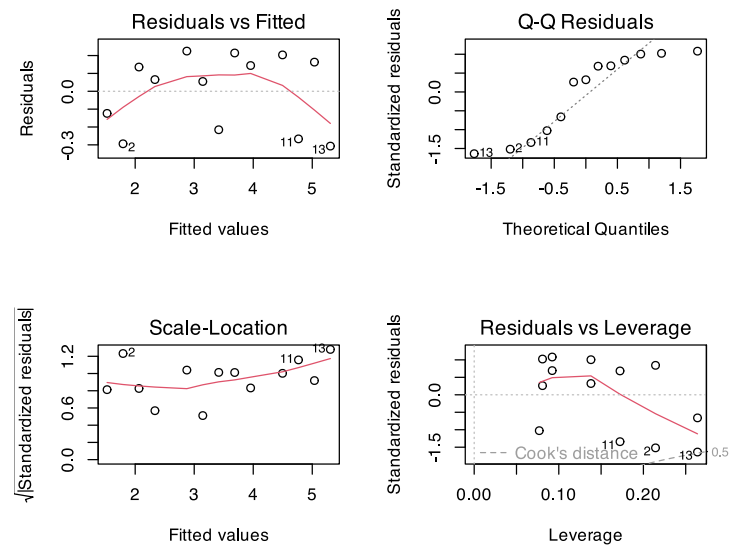


FIGURE 12.2. Standard diagnostic plots for the sparrow simple linear regression. Clockwise from top left: residuals versus fitted values (linearity and constant variance), normal Q-Q plot (residual normality), residuals versus leverage (influential observations), and scale-location plot (homoscedasticity).

12.7.5 Interpret the Results

Inference applies within the observed range of the predictor. Extrapolation beyond the data is possible but requires explicit justification, because the straight-line form may not hold outside the observed range.

I construct the final model fit as a figure, which I will use in my reporting.

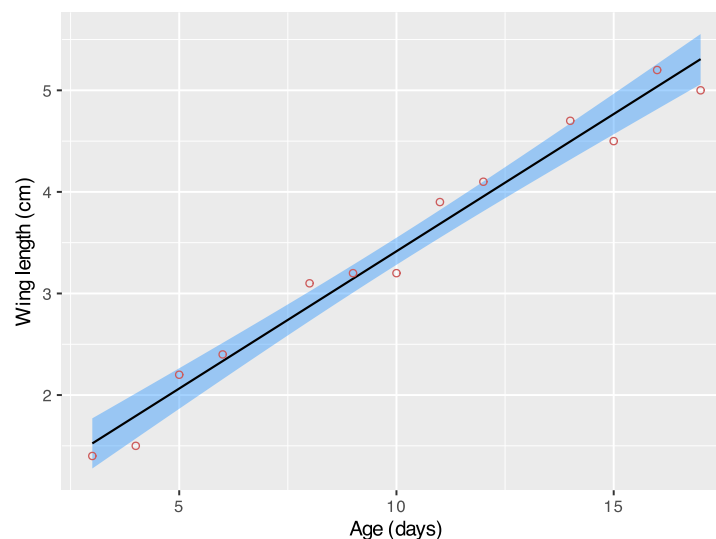


FIGURE 12.3. Wing length as a function of age in the sparrow example. The straight line is the fitted simple linear regression and the blue shading is the 95% confidence interval.

In Figure 12.3, the fitted slope is positive, which means that wing length increases with age. In this example, the slope estimate is about 0.27 cm per day, so the model implies that the expected wing length increases by roughly 0.27 cm for each additional day of age across the range of these observations.

The intercept is the expected wing length when age is zero. Here that value is not biologically the main point of interest. It is simply the point where the fitted line crosses the vertical axis.

The model explains a large proportion of the variation in the observed wing lengths ($R^2 \approx 0.97$), and the test of the slope provides very strong evidence that the linear relationship is not zero ($p < 0.001$).

12.7.6 Reporting

WRITE-UP**Methods**

The relationship between sparrow wing length and age was assessed with a simple linear regression, with wing length as the response variable and age as the continuous predictor. Model adequacy was evaluated from standard residual diagnostics.

Results

Sparrow wing length increased strongly with age in the fitted simple linear regression ($\beta = 0.270$, 95% CI: 0.241 to 0.300; $R^2 = 0.97$; $p < 0.001$) (Figure 12.3). Across the observed age range, older birds therefore had consistently longer wings, with the expected wing length increasing by about 0.27 cm for each additional day of age.

Discussion

This example is useful because it makes the biological interpretation of the slope very clear: age is associated with a strong increase in wing length over the observed range, and the fitted line captures most of the variation in these simple demonstration data.

DO IT NOW!

Use the sparrow model to generate predictions for two new ages and compare the confidence and prediction intervals:

```
new_ages <- tibble(age = c(6, 14))
predict(sparrow_mod, newdata = new_ages, interval = "confidence")
predict(sparrow_mod, newdata = new_ages, interval = "prediction")
```

Answer the following:

1. Which interval is wider at each age, and why?
2. Both intervals are wider at age 14 than at age 6 relative to the interval at the mean age. Explain why interval width varies across the predictor range.
3. Would you trust a prediction for age 25? Why or why not?

12.8 EXAMPLE 2: ADELIE PENGUIN BILL LENGTH AND BODY MASS

The following example uses the penguins dataset from the palmerpenguins package to model bill length as a function of body mass in Adelie penguins.

Although I could also calculate a correlation, I will use a simple linear regression because I want a predictive model that estimates bill length from body mass. This is a defensible use of simple linear regression even though I am not claiming that body mass directly causes bill length.

```
Adelie <- penguins[penguins$species == "Adelie", ]
Adelie <- Adelie[-4, ]
```

TABLE 12.2. Size measurements for adult foraging Adelie penguins near Palmer Station, Antarctica.

Bill length (mm)	Body mass (g)
39.1	3750
39.5	3800
40.3	3250
36.7	3450
39.3	3650
38.9	3625

12.8.1 Do an Exploratory Data Analysis (EDA)

```
dim(Adelie)
```

```
R> [1] 151 8
```

```
summary(Adelie)
```

```
R>      species      island bill_length_mm bill_depth_mm
R> Adelie   :151  Biscoe   :44    Min.   :32.10    Min.   :15.50
R> Chinstrap: 0  Dream    :56    1st Qu.:36.75    1st Qu.:17.50
R> Gentoo   : 0  Torgersen:51    Median :38.80    Median :18.40
R>                                     Mean   :38.79    Mean   :18.35
R>                                     3rd Qu.:40.75    3rd Qu.:19.00
R>                                     Max.   :46.00    Max.   :21.50
R> flipper_length_mm body_mass_g      sex      year
R> Min.   :172      Min.   :2850  female:73    Min.   :2007
R> 1st Qu.:186      1st Qu.:3350   male :73    1st Qu.:2007
R> Median :190      Median :3700   NA's : 5    Median :2008
R> Mean   :190      Mean   :3701                Mean   :2008
R> 3rd Qu.:195      3rd Qu.:4000                3rd Qu.:2009
R> Max.   :210      Max.   :4775                Max.   :2009
```

We see that the dataset contains many more observations than the sparrow example. I focus here on `body_mass_g` and `bill_length_mm`. Both are continuous, and restricting the analysis to Adelie penguins gives me a relatively coherent biological subset for the example.

12.8.2 Create a Plot

```
ggplot(Adelie,
  aes(x = body_mass_g, y = bill_length_mm)) +
  geom_smooth(method = "lm", linewidth = 0.4, alpha = 0.4,
    colour = "black", se = TRUE, fill = "dodgerblue") +
  geom_point(shape = 1, colour = "indianred") +
  labs(x = "Body mass (g)", y = "Bill length (mm)")
```

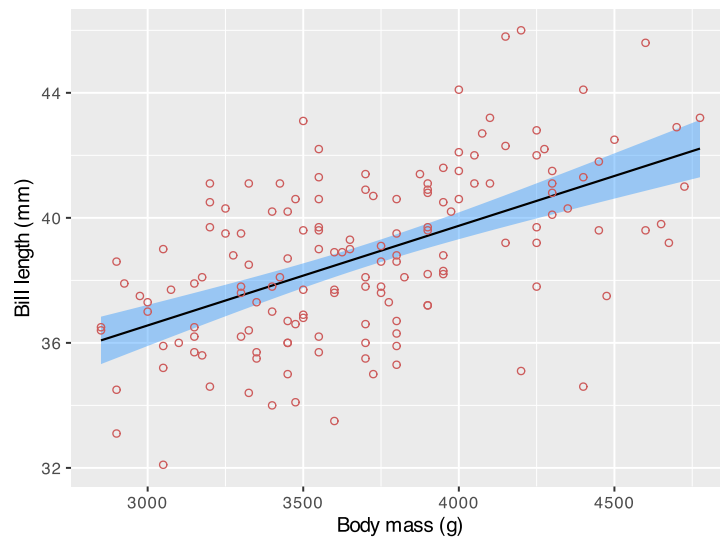


FIGURE 12.4. Scatter plot of the Palmer Station Adelie penguin data with a best fit line.

In Figure 12.4, there is also a clear positive relationship between body mass and bill length despite considerable scatter. This relationship appears linear enough to justify a simple linear model as a first approximation. Creating a publication-quality plot complete with the regression line in place preempts the model fitting, but I can use it later for reporting should it turn out that the model fit is defensible.

12.8.3 State the Hypothesis

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

The null hypothesis is that body mass has no linear association with bill length, while the alternative is that the slope differs from zero.

If the slope is zero, then the predictor does not explain systematic change in the expected response.

12.8.4 Fit the Model

```
mod1 <- lm(bill_length_mm ~ body_mass_g,
           data = Adelie)
summary(mod1)

R>
R> Call:
R> lm(formula = bill_length_mm ~ body_mass_g, data = Adelie)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -6.4208 -1.3690  0.1874  1.4825  5.6168
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept) 2.699e+01  1.483e+00  18.201 < 2e-16 ***
R> body_mass_g 3.188e-03  3.977e-04   8.015 2.95e-13 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 2.234 on 149 degrees of freedom
R> Multiple R-squared:  0.3013, Adjusted R-squared:  0.2966
R> F-statistic: 64.24 on 1 and 149 DF,  p-value: 2.955e-13
```

12.8.5 Test the Assumptions

To facilitate assumption checking I use `augment()` from **broom** to add fitted values, residuals, leverage, and related diagnostics to the data.

```
mod1_data <- augment(mod1)
```

a) *Normality.* I use the Shapiro-Wilk test as one formal check of the residual distribution.

```
shapiro.test(residuals(mod1))

R>
R> Shapiro-Wilk normality test
R>
R> data: residuals(mod1)
R> W = 0.99613, p-value = 0.9637
```

The formal test does not flag a serious departure from normality here, but graphical diagnostics are usually more informative than the test alone, especially in moderate samples where the test has limited power to detect mild departures.

```
p1 <- ggplot(mod1_data, aes(sample = .resid)) +
  stat_qq(shape = 1, colour = "pink") +
  stat_qq_line(colour = "steelblue4") +
  labs(title = "Normal Q-Q", x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_grey()

p2 <- ggplot(mod1_data, aes(x = .resid)) +
  geom_histogram(binwidth = 1, fill = "pink", color = "pink") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
```

```

theme_grey()

p3 <- ggplot(mod1_data, aes(x = .fitted, y = .resid)) +
  geom_point(shape = 1, colour = "pink") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted values", y = "Residuals") +
  theme_grey()

p4 <- ggplot(mod1_data, aes(x = .fitted, y = sqrt(abs(.resid)))) +
  geom_point(shape = 1, colour = "pink") +
  geom_smooth(se = FALSE, colour = "steelblue4") +
  labs(title = "Scale-Location", x = "Fitted values", y = "Sq. Root |Resid.|") +
  theme_grey()

ggarrange(p1, p2, p3, p4, nrow = 2, ncol = 2, labels = "AUTO")

```

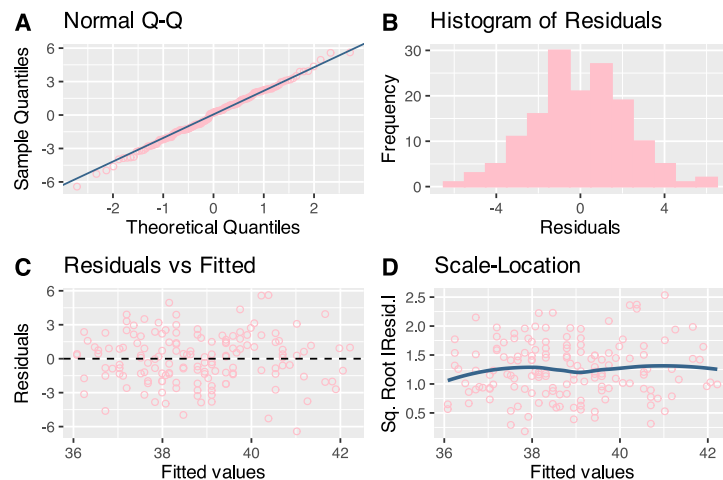


FIGURE 12.5. Diagnostic plots for the simple linear regression of Adelie penguin bill length on body mass. A) Normal Q-Q plot. B) Histogram of residuals. C) Residuals versus fitted values. D) Scale-location plot.

In Figure 12.5, the Q-Q plot and histogram suggest that the residuals are approximately normally distributed. There is no obvious extreme departure that would make the model immediately unusable.

b) Homoscedasticity. The Breusch-Pagan test is one formal check of constant variance.

```

bptest(mod1)

R>
R> studentized Breusch-Pagan test
R>

```

```
R> data: mod1
R> BP = 1.6677, df = 1, p-value = 0.1966
```

The formal test is consistent with the visual impression from the residual plots. Graphical inspection should take precedence: if the scale-location plot shows a clear trend, the spread is not constant regardless of the test p-value. The test does not suggest strong heteroscedasticity, and the residuals-versus-fitted and scale-location panels in Figure 12.5 also indicate that the spread of residuals is reasonably even across the fitted range.

c) Check for Outliers. Four complementary diagnostics reveal different facets of how individual observations affect the fitted model (Figure 12.6, Figure 12.7): DFFITS, Cook's distance, residuals versus leverage, and Cook's distance versus leverage.

```
cooksds_thresh <- 4 / nrow(mod1_data)
dffits_thresh <- 2 * sqrt(2 / nrow(Adelie))

mod1_data <- mod1_data %>%
  mutate(index = row_number(),
         leverage = hatvalues(mod1),
         dffits = dffits(mod1),
         colour = ifelse(.cooksds > cooksds_thresh, "black", "pink"))

plt1 <- ggplot(mod1_data, aes(x = index, y = dffits)) +
  geom_col(fill = ifelse(abs(mod1_data$dffits) > dffits_thresh, "black", "pink")) +
  geom_hline(yintercept = dffits_thresh, linetype = "dashed", colour = "red") +
  geom_hline(yintercept = -dffits_thresh, linetype = "dashed", colour = "red") +
  geom_text(aes(label = ifelse(abs(dffits) > dffits_thresh, as.character(index), "")),
           hjust = 1.0, vjust = 1.0, colour = "darkred") +
  labs(x = "Observation Index", y = "DFFITS") +
  theme_grey()

plt2 <- ggplot(mod1_data, aes(x = index, y = .cooksds, fill = colour)) +
  geom_col() +
  geom_hline(yintercept = cooksds_thresh, linetype = "dashed", colour = "red") +
  geom_text(aes(label = ifelse(.cooksds > cooksds_thresh, as.character(index), "")),
           hjust = 1.2, vjust = 1.0, colour = "darkred") +
  labs(x = "Observation Index", y = "Cook's Dist.") +
  scale_fill_identity() +
  theme_grey()

plt3 <- ggplot(mod1_data, aes(x = .hat, y = .std.resid)) +
  geom_point(aes(size = .cooksds, colour = colour), shape = 1) +
  geom_hline(yintercept = 0, linetype = "dashed", colour = "black") +
  geom_vline(xintercept = 2 * mean(mod1_data$.hat), linetype = "dashed", colour = "blue") +
  geom_smooth(method = "loess", se = FALSE, colour = "steelblue4") +
  geom_text(aes(label = ifelse(.cooksds > cooksds_thresh, as.character(index), "")),
           hjust = 1.0, vjust = 1.0, colour = "darkred") +
  labs(x = "Leverage",
       y = "Std. Resid.",
       size = "Cook's Distance") +
  scale_colour_identity() +
  theme_grey() +
  theme(legend.position = "bottom") +
  guides(size = guide_legend("Cook's Dist.))

plt4 <- ggplot(mod1_data, aes(x = .hat, y = .cooksds)) +
```

```

geom_point(aes(size = .cooks, colour = colour), shape = 1) +
geom_hline(yintercept = cooks_thresh, linetype = "dashed", colour = "red") +
geom_vline(xintercept = 2 * mean(mod1_data$.hat), linetype = "dashed", colour = "blue") +
geom_text(aes(label = ifelse(.cooks > cooks_thresh, as.character(index), "")),
          hjust = 1.0, vjust = 1.0, colour = "darkred") +
labs(x = "Leverage",
     y = "Cook's Dist.",
     size = "Cook's Distance") +
scale_colour_identity() +
theme_grey() +
guides(size = guide_legend("Cook's Distance"))

ggarrange(plt1, plt2, plt3, plt4, nrow = 2, ncol = 2, labels = "AUTO",
          common.legend = TRUE)

```

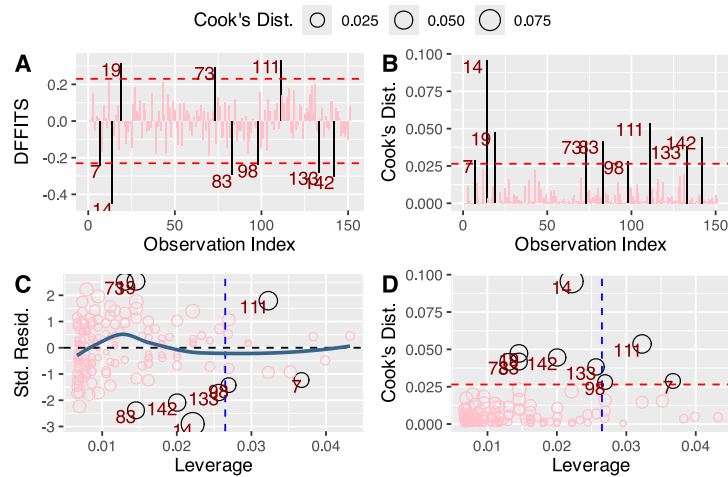


FIGURE 12.6. Diagnostic plots for visual inspection of outliers in the Adelie penguin regression. A) DFFITS. B) Cook's distance. C) Residuals versus leverage. D) Cook's distance versus leverage. Observations beyond the Cook's distance threshold are shown in black and labelled by row number.

```

ggplot(mod1_data, aes(x = body_mass_g, y = bill_length_mm)) +
geom_point(aes(size = .cooks, colour = colour), shape = 1) +
geom_smooth(method = "lm", se = FALSE, colour = "steelblue4") +
geom_text(aes(label = ifelse(abs(.cooks) > cooks_thresh, rownames(mod1_data), "")),
          hjust = 1.0, vjust = 1.0, colour = "darkred") +
labs(x = "Body mass (g)", y = "Bill length (mm)") +
scale_colour_identity() +
theme_grey() +
theme(legend.position = "bottom") +
guides(size = guide_legend("Cook's Distance"))

```

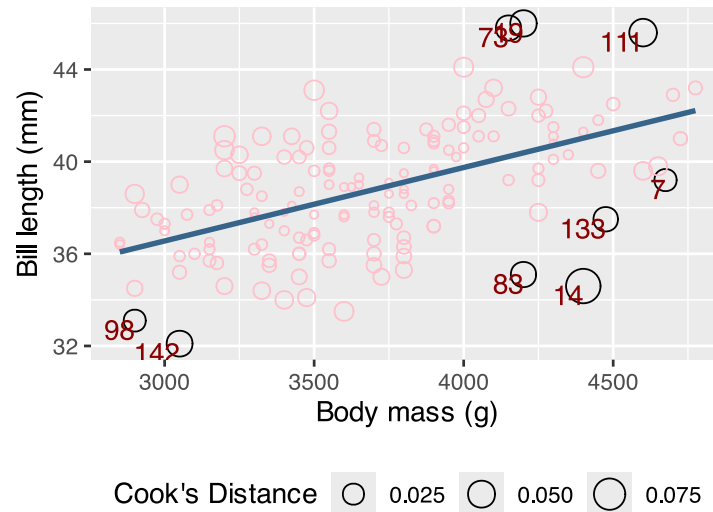



FIGURE 12.7. Scatter plot of Adelie penguin bill length against body mass, with observations exceeding the Cook's distance threshold highlighted.

DFFITs (Difference in Fits) shows how much the fitted value *at each observation* changes when that observation is removed, expressed in estimated standard-error units (Figure 12.6 A). The dashed lines at $\pm 2\sqrt{p/n}$ mark the threshold beyond which removal would noticeably shift the local prediction. Most bars sit well inside the bounds, and none extends dramatically beyond, so no individual penguin strongly controls its own fitted value.

Cook's distance measures *overall* influence by showing the total shift in all fitted values when one observation is excluded. The rough threshold $4/n$ (dashed red line; Figure 12.6 B) identifies candidates for inspection, but they do not have to be automatically deleted without good justification. A handful of observations approach or just cross the line (shown in black and labelled by row number) but no bar is very much taller than its neighbours and the exceedances are marginal.

Figure 12.6 C combines the two sources of concern. Leverage, on the x-axis, measures how far an observation's predictor value sits from the centre of the predictor distribution; a high-leverage point has an unusual X value. The y-axis shows standardised residuals, and point size encodes Cook's distance. The danger zone is upper and lower right: high leverage combined with a large residual means the observation pulls the fitted line toward itself with no counterweight from the rest of the data. The vertical dashed line marks twice the mean leverage. Most points cluster

toward the left and within ± 2 standardised residuals; a few approach the leverage threshold, but none combines unusual predictor position with a large residual.

In Figure 12.6 D, leverage is placed on the x -axis and Cook's distance on the y -axis, making both dimensions of concern visible at once. Observations in the upper-right region (beyond the horizontal Cook's threshold *and* the vertical leverage threshold) would be the highest priority to examine. No point occupies that region here.

Figure 12.7 locates the flagged observations in the original scatter. The highlighted penguins are not at the extremes of either variable and are not pulling the line in any obvious direction. Some observations contribute more to the fitted model than others, as is normal, but no single point dominates the result. If a flagged observation turned out to be a data-entry error, I would correct or exclude it; if it represents a genuine biological extreme, I would keep it and document its influence.

12.8.6 Interpret the Results

Now that the assumptions appear broadly acceptable, I can interpret the fitted model. The slope of the regression line is positive, so bill length increases with body mass. The coefficient is about 3.2×10^{-3} mm/g, meaning that the expected bill length increases by about 0.0032 mm for every additional gram of body mass.

Note that the units of the slope depend directly on the units of both variables. Here the slope is expressed in mm per gram, which produces a small numerical value. Rescaling body mass to kilograms (dividing by 1000) would yield a slope of about 3.2 mm/kg, which is easier to describe verbally. The biological meaning is unchanged.

The multiple R^2 is about 0.30, so the model explains roughly 30% of the observed variation in bill length. A biologically informative regression does not *need* to explain nearly all the variation in the response to be worthwhile.

The test of the slope provides strong evidence that the relationship is not zero ($p < 0.001$). I apply an ANOVA on the fitted model and this leads me to the same practical conclusion, which is that the straight-line model explains a meaningful amount of variation in bill length.

Among the diagnostics, the residuals-versus-fitted plot is the most informative here. The spread is reasonably even and there is no clear curvature, which supports the straight-line mean structure for this species subset.

Had the Q-Q plot shown strong tail departures or the residual spread changed markedly across the fitted range, the next step would be to consider a response transformation, a different model family, or the inclusion of additional predictors.

12.8.7 Reporting

WRITE-UP

Methods

The data analysed in this example were drawn from the Palmer Penguins dataset, which contains measurements on penguins sampled in the Palmer Archipelago, Antarctica. For this worked example, only Adelie penguins were retained. Bill length was treated as the response variable and body mass as the continuous predictor.

A simple linear regression model was fitted using `lm()` in R, with bill length modelled as a function of body mass. Model adequacy was assessed by inspecting residual plots, by applying the Shapiro-Wilk test to the residuals, and by using the Breusch-Pagan test to assess homoscedasticity. Influential observations were explored using Cook's distance, DFFITS, and leverage-based diagnostics.

Results

Bill length increased with body mass in the fitted simple linear regression ($\beta = 0.00319$, $SE = 0.00040$, $t = 8.02$, $p < 0.001$) (Figure 12.4). The model explained about 30% of the variation in bill length ($R^2 = 0.30$), indicating that body mass was an informative but incomplete predictor of bill length. The overall model was also strongly supported by the ANOVA ($F = 64.25$, $p < 0.001$, d.f. = 1, 149).

Discussion

The worked example supports a positive relationship between body mass and bill length in Adelie penguins, but it also shows the limits of a one-predictor model. Body mass explains part of the variation in bill length, not all of it. A fuller biological account would need additional predictors such as sex, age, or ecological context.

DO IT NOW!

The penguin model explained only about 30% of the variation in bill length. Explore whether restricting the data to females improves the fit.

Compare the R^2 and residual standard error from the female-only model with the values from `mod1` (all Adelie penguins combined). Does splitting by sex change the picture? Run the same model for males and compare. What does this tell you about the usefulness of one-predictor models in biology?

12.9 CONFIDENCE AND PREDICTION INTERVALS

The fitted line gives the expected mean response for a given value of the predictor, but two different kinds of interval are commonly needed. A **confidence interval** describes uncertainty in the estimated mean response, whereas a **prediction interval** describes uncertainty for an individual future observation. The prediction interval is

always wider because it must include the scatter of individual observations around the fitted mean.

```
new_x <- tibble(age = c(7, 13))
predict(sparrow_mod, newdata = new_x, interval = "confidence")
```

```
R>      fit      lwr      upr
R> 1 2.604698 2.444344 2.765051
R> 2 4.226072 4.065719 4.386425

predict(sparrow_mod, newdata = new_x, interval = "prediction")
```

```
R>      fit      lwr      upr
R> 1 2.604698 2.097951 3.111444
R> 2 4.226072 3.719325 4.732818
```

I can also visualise the same distinction in the penguin example.

```
pred_conf <- as.data.frame(predict(mod1,
                                   newdata = Adelie,
                                   interval = "confidence"))

pred_pred <- as.data.frame(predict(mod1,
                                   newdata = Adelie,
                                   interval = "prediction"))

results <- cbind(Adelie, pred_conf, pred_pred[, 2:3])
names(results)[c(9:13)] <- c("fit", "lwr_conf", "upr_conf",
                             "lwr_pred", "upr_pred")

ggplot(data = results, aes(x = body_mass_g, y = fit)) +
  geom_line(linewidth = 0.4, colour = "red") +
  geom_ribbon(aes(ymin = lwr_pred, ymax = upr_pred),
            alpha = 0.2, fill = "red") +
  geom_ribbon(aes(ymin = lwr_conf, ymax = upr_conf),
            alpha = 0.2, fill = "blue") +
  geom_point(aes(y = bill_length_mm), shape = 1) +
  labs(x = "Body mass (g)", y = "Bill length (mm)") +
  theme_grey()
```

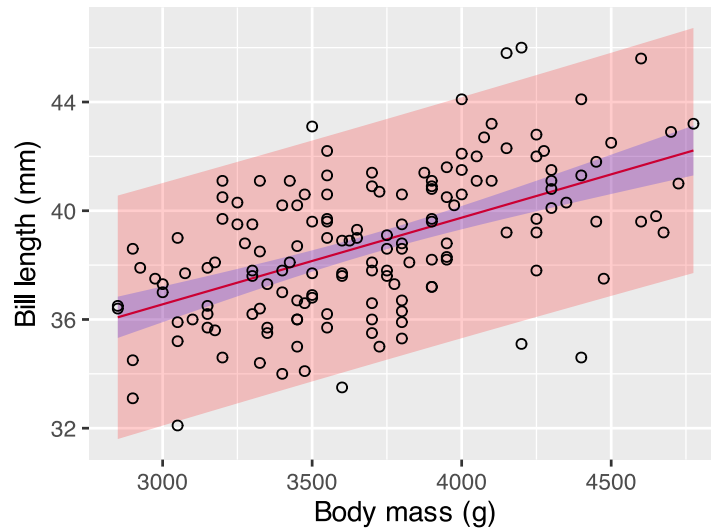


FIGURE 12.8. Adelie penguin bill length model with confidence interval (blue) and prediction interval (pink) around the fitted values.

In Figure 12.8, the blue confidence band is narrower because it describes uncertainty around the fitted mean response, whereas the pink prediction band is wider because it must also accommodate the scatter of individual penguins around that mean. Confidence intervals are therefore useful when the primary interest lies in the mean expected response at a given predictor value. Prediction intervals are more relevant when the goal is to anticipate the range in which an individual future observation may fall.

12.10 PREDICTION VERSUS EXPLANATION

The same fitted straight-line model can be used for at least two different scientific purposes. In an **explanatory** analysis, the main interest is usually the slope itself and what it says about the biological relationship between the predictor and the response. A **predictive** analysis, on the other hand, places the emphasis on the fitted values, prediction intervals, and how accurately the model can anticipate new observations.

The sparrow example is mostly explanatory. We care primarily that wing length increases with age and that the slope is clearly positive. The Adelie penguin example

is closer to a predictive framing because we have treated body mass as a variable from which bill length might be estimated for new individuals.

The distinction is important because it changes what should be emphasised in a Results section. Explanatory work usually emphasises the slope, its uncertainty, and the biological interpretation of the effect. Predictive work still needs the model coefficients, but it should pay much more attention to fitted values, prediction intervals, and the amount of unexplained variation.

This is also why a model with a highly significant slope is not automatically a good predictive model. A relationship can be biologically real and still leave substantial scatter around the fitted line, as in Figure 12.4. Conversely, a model that predicts well is not automatically evidence for a causal mechanism. In Chapter 24, I return to this distinction.

12.11 WHAT TO DO WHEN ASSUMPTIONS FAIL

When diagnostic patterns reveals that the assumptions of the linear model are not adequate, the first step is to ask what those patterns are telling you. Residuals that suggest non-linearity indicate that the model has not absorbed all the systematic structure in the data; the remedy is either to transform the response or predictor variables, or to fit a more flexible model. Polynomial regression, mechanistic non-linear models, and Generalised Additive Models (GAMs) can all accommodate curvature. If residual variance changes strongly with fitted values, consider a transformation or a different modelling framework altogether, such as a Generalised Linear Model (GLM).

Outliers require careful judgement. If they arise from data-entry errors or procedural failures, removal is defensible; however, it should be documented and justified, because outliers may be functionally important as they can reveal rare but real extreme events. When you are confident that an outlier is a genuine observation, robust regression techniques such as M-estimation or least trimmed squares (which downweight influential points rather than discarding them) are worth considering. A final option is to apply an appropriate transformation, such as a logarithm or square root, which can compress the scale sufficiently to reduce the leverage of extreme values without removing any data.

12.12 COMMON MISTAKES

Common mistakes in simple linear regression include:

- using regression when the relationship is only associative and poorly justified as a response-predictor model;

- ignoring non-independence among observations;
- fitting a straight line to a clearly curved relationship;
- treating statistical significance as biological importance;
- reporting R^2 without discussing effect size, uncertainty, or assumptions.

SELF-ASSESSMENT TASK 12-2

Complete a full simple linear regression analysis using life expectancy as the response and schooling as the predictor. Your job is not only to fit the model, but also to identify and justify the treatment of problematic observations.

Use the dataset [kaggle_life_expectancy_data.csv](#).

You should do the following:

1. Prepare the data by selecting the variables needed for the analysis and removing rows with missing values.
2. Fit the initial simple linear regression model.
3. Plot the data and show the fitted regression line.
4. Check the initial model diagnostics graphically.
5. Identify the issue in the dataset.
6. Provide evidence for this issue as a table.
7. Explain why these cases appear problematic for this analysis.
8. Remove the problematic cases and refit the model.
9. Recheck the diagnostics graphically for the revised model.
10. Interpret the final model clearly and state whether it is an improvement over the initial fit. Tabulate the important model-fit statistics in support of your conclusion.
11. End with a short scientific write-up containing Methods, Results, and Discussion sections.

12.12.1 Marking Rubric

Component	Marks
Data preparation	3
Initial model fitting and figure	3
Initial model diagnostics	3
Identifying and explaining the issue	3
Evidence for the issue presented as a table	4
Removing the problematic cases appropriately	2
Refitting the analysis	3

Rechecking the diagnostics	3
Interpreting the final model and comparing it with the initial model	3
Scientific write-up (Methods, Results, Discussion)	3
Total: 30 marks	

12.13 SUMMARY

- Simple linear regression models one continuous response as a function of one continuous predictor.
- The slope is usually the main inferential quantity because it describes how the expected response changes with the predictor.
- Regression differs from correlation because it imposes a response-predictor structure.
- Residual diagnostics are essential because they tell us whether the model is adequate.
- Outlier diagnostics help us decide whether unusual observations are errors, influential extremes, or signs of model misspecification.
- Confidence intervals and prediction intervals answer different questions.

I established the workflow for modelling, above. In the next chapter, I extend the same workflow to curved relationships, and in [Chapter 14](#) I then move to several predictors at once.

13

Polynomial Regression

Curvature Within the Linear Model Framework

IN THIS CHAPTER

- why polynomial regression remains linear in its parameters while capturing curved responses;
- how to read residual curvature as a diagnostic signal that motivates polynomial terms;
- how to fit and compare polynomial models with `lm()`;
- how `poly()` differs from raw powers, and when each is preferable;
- how to interpret the fitted curve biologically, including direction, turning points, and extrapolation limits;
- how to report a polynomial regression consistently with earlier chapters.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Residual (left-over) curvature is the diagnostic signal that motivates polynomial regression. When a straight-line model leaves a curved pattern in the residual-versus-fitted plot, the mean structure is incomplete. Adding polynomial terms corrects that structure and keeps the same least-squares fitting approach, but you'll think about the analysis in the same way as a simple linear regression. The approach to model diagnostics also stays as in Chapters 11 and 12.

As established in [Chapter 11](#), every observed response can be decomposed as $Y_i = \hat{Y}_i + e_i$, where $\hat{Y}_i$ is what the current model predicts and e_i is what it fails to

explain. If those residuals retain a systematic curved pattern, the model is leaving structure behind. Polynomial regression addresses that by adding powers of the predictor to the mean structure.

The fitted curve may look nonlinear on a scatter plot, but the model remains linear in its coefficients. So, this means ordinary least squares, standard errors, confidence intervals, and the residual diagnostic plots from Chapter 11 all remain valid without modification.

Below, I work through when and why to add polynomial terms, how to choose the lowest-order model that removes residual structure, and how to interpret the resulting curve in biological terms.

13.1 IMPORTANT CONCEPTS

Polynomial regression modifies the mean structure of the linear model without changing the fitting procedure. A **quadratic term** (X^2) allows one bend in the response. A **cubic term** (X^3) allows a second bend, but interpretation becomes harder and sensitivity to individual observations increases.

Two constraints apply throughout. **Hierarchy** requires that if X^2 is in the model then X must also be included; if X^3 is included then X and X^2 must also be included. Removing lower-order terms distorts the model geometry and is rarely defensible biologically. **Parsimony** requires that the polynomial degree should be as low as possible while still accounting for the residual structure observed in the simpler model.

DO IT NOW!

Before reading the “When This Method Is Appropriate” section, classify each relationship below as likely requiring (a) a straight-line model, (b) a polynomial model, or (c) a mechanistic nonlinear model. Be prepared to justify your choice.

Biological relationship	Most likely model
Body length increases roughly proportionally with age in juvenile fish over a short growth window	?
CO ₂ uptake in a leaf rises steeply at low light and saturates at high light	?
Enzyme activity peaks at an optimum temperature and declines on both sides	?

Daily running distance in mice increases ?
approximately linearly with days of
training over 3 weeks

Discuss: what feature of the data or underlying biology pushes you toward a polynomial rather than a straight line?

13.2 WHEN THIS METHOD IS APPROPRIATE

Polynomial regression is appropriate when:

- a straight-line model leaves clear curvature in the residual-versus-fitted plot;
- the biological relationship is smooth and continuous rather than step-like or threshold-driven;
- no mechanistic equation is available, but a curved descriptive model is needed;
- the analysis should remain within the `lm()` workflow.

If the biology implies a specific process, such as a Michaelis-Menten uptake curve or an exponential growth model, a [mechanistic nonlinear model](#) fits that expectation more directly and extrapolates more sensibly. If the curvature is complex and cannot be well described by low-order powers, a [generalised additive model](#) is preferable. Low-order polynomials occupy a useful middle ground because they are more flexible than a straight line and less assumption-laden than a mechanistic model. I have seldom needed to apply one in real-life biological applications, however.

13.3 NATURE OF THE DATA AND ASSUMPTIONS

13.3.1 *Requirements Before Fitting*

Polynomial regression uses ordinary least squares and inherits the same data requirements as simple linear regression:

1. **A defensible response-predictor structure.** There should be a theoretical or biological reason to treat one variable as the response and the other as the predictor.
2. **Independence of observations.** Each measured value of the response must be independent of the others. Repeated measurements, clustered sampling, or temporal dependence require a different modelling approach.
3. **Continuous predictor.** The predictor must take values along a continuum over which the polynomial curve can be defined.

4. **Continuous response.** The response should be continuous. Counts or proportions require a different model family.

13.3.2 Assumptions to Check After Fitting

After the model is fitted, residuals are used to check whether the assumptions hold:

1. **Linearity of residuals.** Adding polynomial terms should reduce or remove curvature in the residual-versus-fitted plot. If curvature persists, a higher-order term, a different transformation, or a more flexible model may be needed.
2. **Constant variance.** The spread of residuals should be roughly uniform across the fitted range. High-order polynomial terms can amplify variance at the extremes of the predictor, making heteroscedasticity more likely there.
3. **Approximate normality of residuals.** The Q-Q plot should not show systematic departures. High-degree polynomial terms increase sensitivity to extreme observations, so outliers warrant closer attention as model complexity increases.
4. **Independence.** Residuals should not retain temporal, spatial, or other structured patterns. The check is the same as in Chapter 12: inspect residuals against order, time, or space if those structures are plausible.

13.3.3 Assumption–Diagnostic Mapping

The same assumption-to-diagnostic mapping used in Chapter 12 applies here:

Assumption	Main diagnostic	Visual signal to watch for
Linearity	residual vs fitted plot	slope or curvature in residuals
Constant variance	residual vs fitted / scale-location	funnel or changing vertical spread
Normality	residual Q-Q plot	systematic bends away from reference line
Independence	residuals vs order/time/space	runs, cycles, clusters

DO IT NOW!

The `I()` wrapper in R formulas is easy to forget. Run the two lines below and compare what each one actually fits:

```
data("CO2")
co2_df <- subset(CO2, Treatment == "chilled" & Type == "Quebec")

# Version A – wrong way to try to fit a quadratic
fit_a <- lm(uptake ~ conc^2, data = co2_df)

# Version B – correct quadratic
fit_b <- lm(uptake ~ conc + I(conc^2), data = co2_df)

summary(fit_a)
summary(fit_b)
```

How many terms appear in the coefficient table for each model? Why does `conc^2` without `I()` not add a squared term? What does `conc^2` mean to R's formula parser?

13.4 THE CORE EQUATIONS

What polynomial regression changes is the mean structure. The errors ε_i and the least-squares fitting rule remain unchanged. A cubic polynomial model is:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i \quad (13.1)$$

In Equation 13.1, curvature results from the powers of X , but the model is still linear in the coefficients $\alpha, \beta_1, \beta_2, \beta_3$. A quadratic model omits the cubic term. A straight-line model omits both higher-order terms. The hierarchy of models is nested, which means nested-model ANOVA can be used to compare them.

The observed residuals $e_i = Y_i - \hat{Y}_i$ estimate the unobserved errors ε_i and remain the objects used for all diagnostic checks, exactly as in Chapter 11.

13.5 THE FITTING RULE

Ordinary least squares minimises the residual sum of squares:

$$\text{RSS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

This criterion is identical to what `lm()` applies in simple linear regression. Adding polynomial terms to the model formula simply changes which $\hat{Y}_i$ is computed.

13.6 IMPLEMENTATION

The model is still fitted with `lm()`:

```
lm(y ~ x + I(x^2), data = df)           # quadratic
lm(y ~ x + I(x^2) + I(x^3), data = df)  # cubic
lm(y ~ poly(x, degree = 3), data = df)  # orthogonal polynomial basis
```

The `I()` wrapper. R’s formula language assigns special meaning to $^$. So, x^2 inside a formula means the interaction of x with itself, not the arithmetic square of x . The `I()` function (“as-is”) suppresses that special interpretation and forces R to evaluate the expression inside as ordinary arithmetic. Writing `I(x^2)` therefore passes the squared values of x as a new predictor column, which is what the polynomial model requires.

Raw powers versus `poly()`. When `I(x^2)` and `I(x^3)` are used directly, the model terms are highly correlated with each other and with x , especially when x is not centred near zero. This collinearity inflates coefficient standard errors and can make the individual estimates hard to interpret or numerically unstable. The `poly()` function constructs orthogonal polynomial contrasts that are uncorrelated by design. Coefficient estimates from `poly()` are therefore more numerically stable, and model-comparison statistics are unaffected by the correlation structure. For teaching and biological interpretation, raw powers are usually easier to explain; for numerical reliability, especially at higher degrees, `poly()` is preferable.

Centring. Centring X before computing powers (*i.e.*, replacing X with $(X - \bar{X})$) reduces the correlation between X and X^2 when using raw powers, and makes the intercept interpretable as the predicted response at the mean predictor value. When the predictor is far from zero and spans a wide range, centring is worth the small additional step.

13.7 EXAMPLE: CO₂ UPTAKE AND CURVED RESPONSE

13.7.1 Example Dataset

I use the `CO2` dataset from base R. These are measurements of carbon dioxide uptake by chilled Quebec grass plants (*Echinochloa crus-galli*) across increasing ambient CO₂ concentrations. Uptake is expected to rise steeply at low concentrations and then decelerate as the carboxylation system approaches saturation, so a straight-line model is likely to underfit the relationship.

A researcher seriously studying this system would fit a rectangular hyperbola (the Michaelis-Menten model) because the kinetics of Rubisco carboxylation are well characterised and that functional form applies sensibly. Here I use a polynomial

TABLE 13.1. A subset of the CO₂ dataset showing chilled Quebec plants used in the polynomial-regression example.

Plant	Type	Treatment	conc	uptake
Qc1	Quebec	chilled	95	14.2
Qc1	Quebec	chilled	175	24.1
Qc1	Quebec	chilled	250	30.3
Qc1	Quebec	chilled	350	34.6
Qc1	Quebec	chilled	500	32.5
Qc1	Quebec	chilled	675	35.4
Qc1	Quebec	chilled	1000	38.7
Qc2	Quebec	chilled	95	9.3
Qc2	Quebec	chilled	175	27.3
Qc2	Quebec	chilled	250	35.0

instead, purely to demonstrate the method on a dataset where curvature is visible and unambiguous. The polynomial model is adequate as a descriptive approximation within the observed concentration range but should not be used for extrapolation or for inferring kinetic parameters. The first ten rows of the filtered dataset are shown in Table 13.1.

```
co2_df <- as_tibble(CO2) >
  filter(Treatment = "chilled", Type = "Quebec")
gt(head(co2_df, 10))
```

DO IT NOW!

Before fitting any model, spend a few minutes on the CO₂ dataset directly. Run this summary code and answer the questions below:

```
co2_df <- as_tibble(CO2) >
  filter(Treatment = "chilled", Type = "Quebec")

# How many observations?
nrow(co2_df)

# How many distinct concentration levels?
co2_df > count(conc)

# Quick scatter plot
ggplot(co2_df, aes(x = conc, y = uptake)) +
  geom_point() +
  labs(x = expression(CO[2]~concentration~(mL~L^{-1})),
       y = expression(CO[2]~uptake~(μmol~m^{-2}~s^{-1})))
```

1. At the lowest concentration level, how variable is uptake across observations?

2. Does the relationship look linear? If not, describe the shape.
3. Based on the scatter plot alone, do you expect a straight-line model to over-predict or under-predict uptake in the middle of the concentration range?

13.7.2 Do an Exploratory Data Analysis (EDA)

```
co2_df > summarise(
  n = n(),
  min_conc = min(conc),
  max_conc = max(conc),
  mean_uptake = mean(uptake),
  sd_uptake = sd(uptake)
)

R> # A tibble: 1 x 5
R>   n min_conc max_conc mean_uptake sd_uptake
R>   <int>   <dbl>   <dbl>     <dbl>   <dbl>
R> 1    21     95    1000       31.8     9.64

ggplot(co2_df, aes(x = conc, y = uptake)) +
  geom_point(alpha = 0.75) +
  labs(x = expression(CO[2]~concentration~(mL~L^{-1})),
       y = expression(CO[2]~uptake~(\mu mol~m^{-2}~s^{-1})))
```

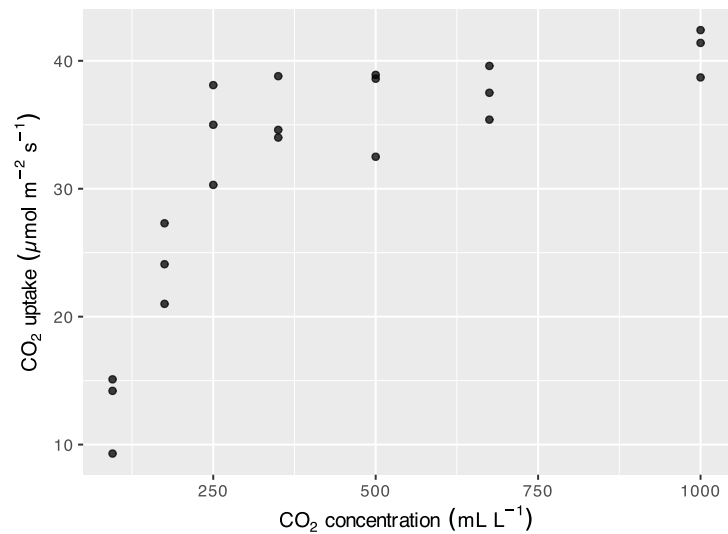


FIGURE 13.1. CO₂ uptake in chilled Quebec plants across concentration levels. The pattern is positive but clearly decelerating.

Figure 13.1 shows a positive relationship with a clear bend: uptake increases steeply at low concentrations and then levels off. Before fitting polynomial terms, I fit a straight-line model and inspect its residuals to confirm the diagnostic signal.

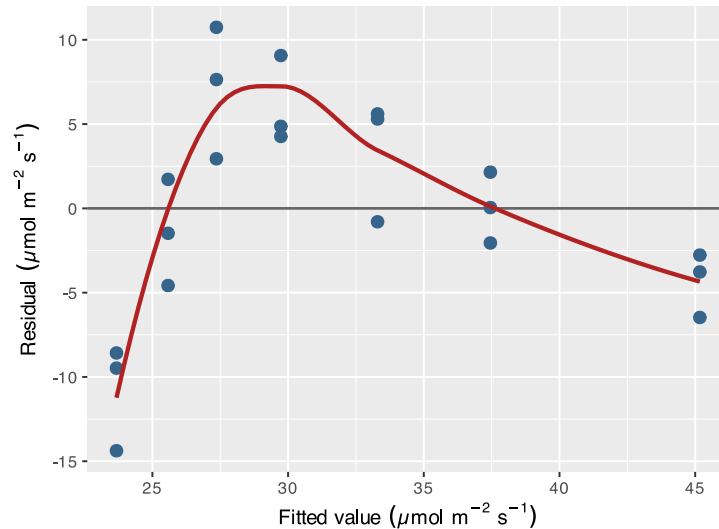


FIGURE 13.2. Residual-versus-fitted plot for the straight-line model. The curved smooth line confirms that the linear mean structure is insufficient.

Figure 13.2 shows an obvious arc. The residuals are negative at low and high fitted values and positive in the middle. The straight-line model is systematically wrong across the concentration range, with the curvature motivating adding a quadratic term.

13.7.3 State the Model Question

The central question is whether adding polynomial terms to the mean structure removes the residual curvature observed above. Polynomial terms are tools for improving model specification, not primarily for testing isolated coefficients.

One useful formal check is whether the quadratic coefficient differs from zero:

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

If $\beta_2 = 0$ cannot be rejected, the data offer no evidence for the quadratic shape. That test should be read alongside the model comparison statistics and the diagnostic plots, not in isolation.

13.7.4 Fit the Model

```
mod_lin <- lm(uptake ~ conc, data = co2_df)
mod_quad <- lm(uptake ~ conc + I(conc^2), data = co2_df)
mod_cub <- lm(uptake ~ conc + I(conc^2) + I(conc^3), data = co2_df)

summary(mod_lin)

R>
R> Call:
R> lm(formula = uptake ~ conc, data = co2_df)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -14.3773  -3.7712   0.0476   4.8664  10.7414
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept) 21.421041   2.583221   8.292 9.79e-08 ***
R> conc         0.023750   0.004919   4.828 0.000117 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 6.631 on 19 degrees of freedom
R> Multiple R-squared:  0.5509, Adjusted R-squared:  0.5273
R> F-statistic: 23.31 on 1 and 19 DF, p-value: 0.0001169

summary(mod_quad)

R>
R> Call:
R> lm(formula = uptake ~ conc + I(conc^2), data = co2_df)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -8.7023 -2.9023   0.2657   2.3714  10.0540
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept) 1.053e+01  3.269e+00   3.221 0.004740 **
R> conc         8.392e-02  1.511e-02   5.554 2.85e-05 ***
R> I(conc^2)    -5.542e-05  1.351e-05  -4.102 0.000669 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 4.898 on 18 degrees of freedom
R> Multiple R-squared:  0.7679, Adjusted R-squared:  0.7421
R> F-statistic: 29.78 on 2 and 18 DF, p-value: 1.954e-06

summary(mod_cub)

R>
R> Call:
R> lm(formula = uptake ~ conc + I(conc^2) + I(conc^3), data = co2_df)
R>
```

```

R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -6.0662 -2.2950  0.3338  1.6063  6.4848
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept) -5.325e+00  3.792e+00  -1.404    0.178
R> conc         2.339e-01  3.128e-02   7.478 9.03e-07 ***
R> I(conc^2)    -3.970e-04  6.818e-05  -5.822 2.04e-05 ***
R> I(conc^3)     2.094e-07  4.145e-08   5.051 9.84e-05 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 3.187 on 17 degrees of freedom
R> Multiple R-squared:  0.9072, Adjusted R-squared:  0.8908
R> F-statistic: 55.39 on 3 and 17 DF,  p-value: 5.509e-09

anova(mod_lin, mod_quad, mod_cub)

R> Analysis of Variance Table
R>
R> Model 1: uptake ~ conc
R> Model 2: uptake ~ conc + I(conc^2)
R> Model 3: uptake ~ conc + I(conc^2) + I(conc^3)
R>   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
R> 1      19 835.48
R> 2      18 431.80  1    403.68 39.746 7.919e-06 ***
R> 3      17 172.66  1    259.14 25.516 9.844e-05 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC(mod_lin, mod_quad, mod_cub)

R>      df      AIC
R> mod_lin  3 142.9485
R> mod_quad  4 131.0877
R> mod_cub   5 113.8378

```

The nested-model ANOVA tests whether each additional term reduces the RSS enough to justify its inclusion. It answers, is the improvement statistically supported? The AIC penalises model complexity and asks, across these models, which offers the best trade-off between fit and parsimony? The two tools are complementary. When both point to the same model, deciding on the best model is easy. When they disagree, for example, the ANOVA supports adding a cubic term but AIC favours the quadratic, I prefer the simpler model (the quadratic) unless the biological context provides a strong reason to expect the additional curvature.

The working rule is to choose the lowest-order model that removes residual structure. Start with the linear model, inspect its residuals, add the quadratic term if curvature remains, then check whether the cubic term still improves the residual pattern. Stop as soon as the diagnostic plots show no further systematic structure.

DO IT NOW!

Look at the ANOVA and AIC output from the three-model comparison:

```
anova(mod_lin, mod_quad, mod_cub)
AIC(mod_lin, mod_quad, mod_cub)
```

Answer the following without running additional code:

1. Which model has the lowest AIC? What does that mean in terms of the trade-off between fit and parsimony?
2. The nested-model ANOVA tests whether adding the quadratic term significantly

```
quad_aug <- augment(mod_quad)
```

```
shapiro.test(residuals(mod_quad))
```

3. Does adding the cubic term beyond the quadratic produce a significant improvement according to the ANOVA? Does AIC agree?
4. Based on these two pieces of evidence together, which model would you choose and why?

13.7.5 Test Assumptions

```
R>
R> Shapiro-Wilk normality test
R>
R> data: residuals(mod_quad)
R> W = 0.9902, p-value = 0.998
bptest(mod_quad)
```

```
R>
R> studentized Breusch-Pagan test
R>
R> data: mod_quad
R> BP = 1.3941, df = 2, p-value = 0.498
par(mfrow = c(2, 2))
plot(mod_quad)
par(mfrow = c(1, 1))
```

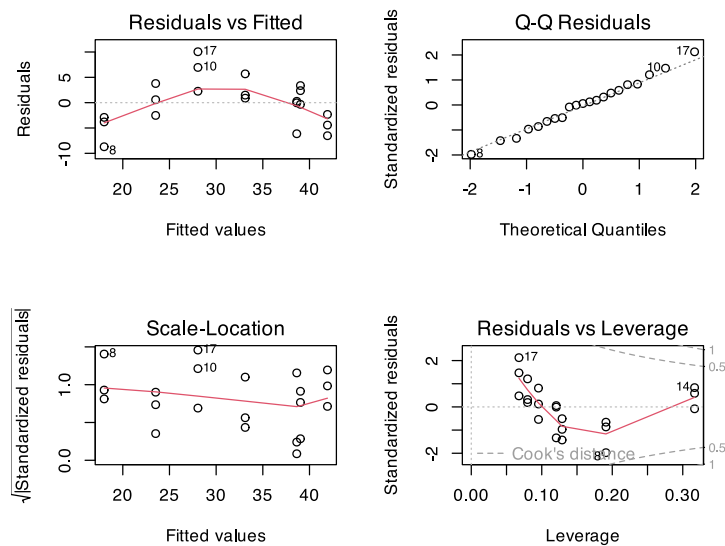


FIGURE 13.3. Standard diagnostics for the quadratic model.

The graphical diagnostics in Figure 13.3 take priority over the formal tests. The Shapiro-Wilk test and the Breusch-Pagan test are supporting evidence, not the primary decision criterion. The residual-versus-fitted panel should show no remaining arc; the Q-Q plot should show residuals tracking the reference line without systematic departures. Both conditions are broadly met here for the quadratic model, which confirms that adding the quadratic term corrected the diagnostic problem identified in Figure 13.2.

DO IT NOW!

Examine the four diagnostic panels in Figure 13.3 for the quadratic model. For each panel, write one sentence describing what you see and whether it raises any concern:

Panel	What you see	Concern?
Residuals vs Fitted	?	?
Normal Q-Q	?	?
Scale-Location	?	?
Residuals vs Leverage	?	?

Then compare the quadratic model diagnostics with what you would expect from the straight-line model (recall Figure 13.2). Has the quadratic term resolved the main diagnostic problem? Is there any remaining issue worth noting?

13.7.6 Interpret the Results

```
co2_pred <- tibble(conc = seq(min(co2_df$conc), max(co2_df$conc), length.out = 200)) >
  mutate(
    linear = predict(mod_lin, newdata = cur_data()),
    quadratic = predict(mod_quad, newdata = cur_data()),
    cubic = predict(mod_cub, newdata = cur_data())
  ) >
  pivot_longer(cols = c(linear, quadratic, cubic),
    names_to = "model", values_to = "fit")

ggplot(co2_df, aes(x = conc, y = uptake)) +
  geom_point(alpha = 0.65) +
  geom_line(data = co2_pred, aes(y = fit, colour = model), linewidth = 0.9) +
  labs(x = expression(CO[2]~concentration~(mL~L^{-1})),
    y = expression(CO[2]~uptake~(\mu mol~m^{-2}~s^{-1})),
    colour = "Model")
```

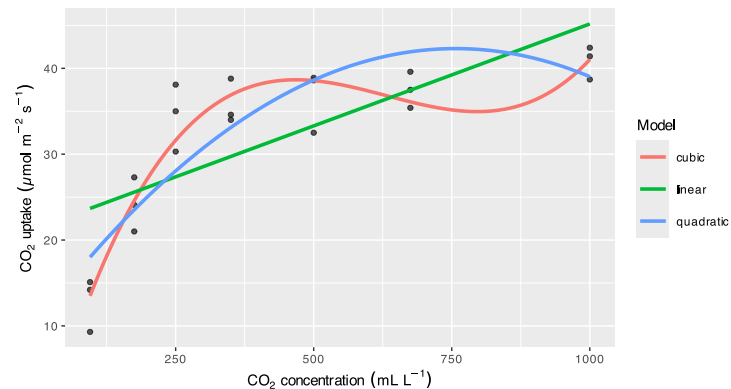


FIGURE 13.4. Linear, quadratic, and cubic polynomial fits to the CO₂ uptake data.

DO IT NOW!

Look at the three fitted curves in Figure 13.4. Without reading the interpretation text below the figure, answer these questions:

1. Where does the straight-line model perform worst relative to the data points?
2. Do the quadratic and cubic curves look noticeably different from each other across the observed concentration range?

3. The quadratic curve does not reverse direction within the observed range. What would it mean biologically if it did reverse?
4. Suppose you wanted to predict CO₂ uptake at a concentration of 2000 mL/L (well beyond the observed range). Which model would you trust the least for this extrapolation, and why?

Discuss your answers with a partner.

Figure 13.4 shows the three fitted curves overlaid on the data. Interpreting a polynomial fit means describing the shape of the response and not simply citing the coefficient values. Individual polynomial coefficients are difficult to interpret in isolation since they depend on the scale of the predictor and the degree of the model. So, they should not normally be the primary result reported.

Be systematic. First, identify the **direction of curvature**: is the response accelerating (concave up) or decelerating (concave down) across the predictor range? For this dataset the response is decelerating because uptake increases with CO₂ concentration but at a diminishing rate. Second, identify any **turning point** in the fitted curve. A quadratic model has at most one turning point. We see in the curve here that it does not reverse within the observed concentration range, so the relationship remains monotonically positive throughout. Third, relate the shape to a **biological process**. You'd mention that the deceleration is consistent with a carboxylation system that saturates as substrate availability increases, matching the biochemistry of C₃ carbon fixation.

Inference applies within the observed predictor range, from the lowest to the highest CO₂ concentration in the dataset. Polynomial curves can behave erratically outside that range. A quadratic function will eventually reverse direction, and a cubic will eventually diverge. So, extrapolation beyond the data requires explicit biological justification and should be treated with caution.

Note also that the scale of the predictor affects the slope interpretation. Concentration here is measured in $\mu\text{L/L}$. A one-unit increase in concentration corresponds to a very small change in the predictor. Rescaling to a biologically natural unit (for example, hundreds of $\mu\text{L/L}$) would produce larger, easier-to-state coefficients without changing the fitted curve.

13.7.7 Reporting

WRITE-UP**Methods**

Carbon dioxide uptake was modelled as a function of ambient CO₂ concentration using ordinary least squares regression on the CO₂ plant dataset (chilled Quebec plants). A straight-line model was fitted first and its residual-versus-fitted plot inspected for curvature. Quadratic and cubic polynomial extensions were then fitted, and model selection used nested-model ANOVA and AIC. The lowest-order model that removed residual curvature was selected. Assumptions were assessed from residual diagnostics.

Results

The straight-line model left clear curvature in the residual-versus-fitted plot, indicating that the linear mean structure was insufficient. Adding a quadratic term substantially improved fit (nested-model comparison: $p < 0.001$) and removed the residual arc. The cubic extension provided no meaningful additional improvement relative to the quadratic model. The selected quadratic model described a positive and decelerating relationship (Figure 13.4): CO₂ uptake increased with concentration across the observed range but at a diminishing rate, consistent with substrate saturation of the carboxylation system. The model should be interpreted within the observed concentration range.

Discussion

A quadratic polynomial provided a parsimonious descriptive model of the concentration-uptake relationship while retaining the `lm()` workflow. The fitted curve is biologically plausible as a first-order approximation of carboxylase saturation, though a mechanistic Michaelis-Menten model would be preferred where the process is to be characterised precisely or extrapolation is required.

13.8 WHAT TO DO WHEN ASSUMPTIONS FAIL

Diagnostic outcomes point to different remedies depending on what they show.

If the quadratic model still leaves residual curvature, fitting the cubic extension is the next step. If curvature persists at degree three (or if the degree required to describe the pattern is becoming ridiculously high) the pattern is probably too complex for polynomial terms and a [generalised additive model](#) is better. A GAM estimates the smooth relationship from the data directly without committing to a polynomial form, which makes it more robust to complex or irregular curvature.

If the biology provides a specific mechanistic expectation (*e.g.*, for example, Michaelis-Menten kinetics, logistic population growth, or a Gompertz growth curve) a [nonlinear model](#) embeds that structure directly and extrapolates in a biologically sensible way. Now, the interpretable, biologically-relevant model parameters also become available, such as V_{max} and K_s for the Michaelis-Menten model. Polynomial

approximations should not substitute for mechanistic models when the process is understood.

If the residual spread increases markedly at the extremes of the predictor range (a pattern that high-degree polynomials can amplify), consider whether a response transformation (log or square root) stabilises variance before adding polynomial terms. If variance is still unstable after transformation, weighted least squares or a generalised linear model with an appropriate variance function may be needed.

13.9 SUMMARY

Polynomial regression extends simple linear regression to curved responses by adding powers of the predictor to the mean structure. The fitting method, inferential tools, and diagnostic workflow remain unchanged. The choice of polynomial degree should be driven by residual diagnostics, in other words, fit the linear model first, inspect the residual-versus-fitted plot, and add polynomial terms only as long as they remove structure from the residuals. Model comparison using nested ANOVA and AIC supports that decision. Interpretation focuses on the shape of the fitted curve, such as the direction, turning points (if present), and biological plausibility, rather than on individual coefficient values. The fitted curve applies within the observed predictor range; extrapolation from polynomial models requires care.

14

Multiple Regression and Model Specification

From Biological Hypotheses to Statistical Models

IN THIS CHAPTER

- why simple regression is often insufficient for biological questions with multiple drivers;
- how multiple regression models a response using several predictors simultaneously;
- why model specification is part of the scientific argument, not just a modelling choice;
- how to distinguish model specification from model selection;
- how to detect and respond to multicollinearity using the variance inflation factor;
- how to read the coefficient table and the ANOVA table, and when each is appropriate;
- how dummy-variable coding handles categorical predictors within the linear model;
- how ANCOVA relates to multiple regression;
- how to fit, diagnose, and report a worked multiple regression example using seaweed biogeographic data.

TASKS TO COMPLETE IN THIS CHAPTER

- None

I modelled the relationship between two variables in **Chapter 12**. Biological systems are rarely that simple. Growth, abundance, species diversity, and dissimilarity in community composition are each driven by several environmental and biological factors acting simultaneously. Multiple linear regression is obvious progression from simple regression to more realistic applications, since it models a continuous response as a linear function of two or more predictors.

Adding predictors creates a more demanding scientific problem alongside the statistical one. Which predictors belong in the model, and why? What does each coefficient mean once others are present? So, I treat multiple regression and model specification together in this chapter, because separating them encourages data-driven assembly of predictors that produces unreliable models. I would rather place emphasis on biology and encourage theory-driven insights as the prime motivation of the model's construction.

The worked examples use seaweed biogeographic data from Smit et al. (2017). The dataset reappears in several places in BCB744 and is well suited here because it contains a continuous response, continuous climate predictors, and a categorical regional predictor.

14.1 IMPORTANT CONCEPTS

- **Multiple regression** models a response using several predictors simultaneously. Each fitted coefficient is a partial effect; this is, the expected change in the response for a one-unit increase in that predictor when all other predictors are held constant.
- **Model specification** is the process of capturing the biological reasoning into a statistical formula. Specifying a model means deciding which variables belong, what functional forms they should take, and whether any interactions are needed.
- **Model selection** is the process of choosing among competing specified models. It follows specification and doesn't replace it.
- **Multicollinearity** arises when predictors are strongly intercorrelated. It inflates coefficient standard errors and makes partial effects difficult to interpret.
- **Categorical predictors** enter the linear model through dummy-variable coding. Numerically, this is no different from including a continuous predictor, but the interpretation changes.

- **ANCOVA** combines a factor and a continuous covariate in one additive linear model. It is not a separate procedure but simply a special case of multiple regression.

14.2 MODEL SPECIFICATION VS MODEL SELECTION

When model building and fitting are approached as a single activity, the workflow isn't optimal and it might result in a sub-optimal outcome. Keep the two processes apart:

Model specification happens before fitting. It means deciding, on biological grounds, which variables could plausibly affect the response, what functional form each relationship might take, and whether any predictors should interact. A good specification reflects the scientific hypothesis, but a poor one produces coefficients that are numerically valid but biologically uninterpretable.

Model selection happens only *after* a plausible set of models has been specified. It asks which of the candidate models is best supported by the data, using criteria such as AIC or nested model comparison. Therefore, selection only makes sense within a well-specified candidate set of models. If the specification is weak, for example because predictors are chosen by convenience rather than theory, selection will just identify the least bad option from a set that may not contain the right model at all.

The practical consequence is that the first question in any multiple regression analysis is not “which variables reduce AIC?” but “which variables represent the hypothesised process, and which are potential confounders?” Model selection tools such as stepwise algorithms or AIC comparison are used in [Example 1](#) below, but they are applied within a candidate variable set that was motivated biologically, not assembled arbitrarily.

The broader problems of omitted variables, confounding, and measurement error are treated in [Chapter 16](#).

14.3 NATURE OF THE DATA

14.3.1 *Requirements Before Fitting*

- **Response variable:** The response should be continuous and measured without systematic error.
- **Predictors:** Predictors may be continuous or categorical. They should represent variables that could plausibly affect the response, not just variables that happen to be available.

- **Independence:** Observations must be independent. Spatial autocorrelation, repeated measurements on the same individual, or clustered sampling structures each require approaches beyond the standard multiple regression presented here.
- **Approximate linearity:** The response should relate approximately linearly to each continuous predictor on the scale used in the model. Scatterplots of the response against each predictor individually are the first check.

DO IT NOW!

Consider this scenario: a researcher wants to know what drives kelp biomass on the South African coast. She measures water temperature, nitrate concentration, wave energy, light availability, and distance from upwelling centres at 80 sites. Before fitting anything, discuss with a partner:

1. Which of these variables would you include in a multiple regression model, and why?
2. Is there any variable you would exclude on biological grounds alone?
3. If two of the predictors are very strongly correlated (say, temperature and distance from upwelling), what does that tell you about what the model will and will not be able to do?

Write down your candidate model as an R formula:

```
lm(kelp_biomass ~ ???, data = kelp_df)
```

There is no single right answer, but every choice must be justifiable.

14.3.2 Assumptions After Fitting

The residual assumptions in multiple regression are the same as in simple regression, and are also applied to the fitted multiple-predictor model.

- **Normality:** Residuals should be approximately normally distributed. A Q-Q plot of the residuals is the standard check.
- **Constant variance:** Residual variance should not trend with fitted values. A residuals-vs-fitted plot is the standard check.
- **No severe multicollinearity:** Predictors should not be so strongly intercorrelated that their partial coefficients become imprecise or uninterpretable. The variance inflation factor (see below) quantifies this.

Even when residual assumptions are satisfied, the model can still be scientifically weak if the variables are chosen without biological justification. Statistical adequacy and scientific defensibility are separate criteria but both are indispensable to model building.

DO IT NOW!

Load the seaweed dataset and produce scatter + line plots of the five candidate predictor variables and the response Y for the ECTZ subset. Look for approximate linearity and for strong correlations among predictors.

```
sw <- read.csv(here::here("data", "BCB743", "seaweed", "spp_df2.csv"))
sw_ectz <- sw[sw$bio == "ECTZ", ]
sw_sub1 <- sw_ectz[, c("Y", "augMean", "febRange", "febSD", "augSD", "annMean")]
pairs(sw_sub1, pch = 16, col = "dodgerblue4", cex = 0.7)
```

From the plot, answer:

1. Do any of the predictors appear to violate the approximate linearity requirement with Y ?
2. Which pairs of predictors look strongly correlated with each other? What problem would that create in a multiple regression?
3. Does Y look roughly symmetric, or would a transformation help?

14.4 THE CORE EQUATION

The multiple regression model is:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i \quad (14.1)$$

Each β_j in Equation 14.1 is a partial slope: the expected change in Y for a one-unit increase in X_j while all other predictors remain constant. This conditional interpretation distinguishes multiple regression from a set of simple regressions run independently, and this is also where misinterpretation most commonly occurs.

A partial coefficient is not a causal effect. It is an adjusted association. If an important driver of the response is omitted from the model and that omitted variable correlates with an included predictor, the included predictor's coefficient will absorb part of the omitted variable's effect. The coefficient then reflects a mixture of the predictor's own relationship with the response and the confounding from the omitted variable. This is the distinction between **collinearity** (which is a property of the included predictors) and **confounding**, which involves a predictor that should have been included but was not. [Chapter 16](#) delves into both problems in more detail.

14.5 OUTLIERS

Unusual observations can affect coefficient estimates and standard errors in multiple regression as they do in simple regression. The added difficulty is that an observation can be outlying in multivariate predictor space without appearing extreme on any

single variable alone. **Residuals versus leverage** plots and **Cook's distance** are therefore more informative diagnostics in multiple regression than they would be in a simple regression. Both are produced by the standard diagnostic plotting functions used in the examples below.

14.6 R FUNCTION

The `lm()` function fits a multiple regression model in the same way it fits a simple regression model, extended to include several terms in the formula.

For a model with only continuous predictors:

```
lm(response ~ predictor1 + predictor2 + predictor3,
    data = dataset)
```

Interaction effects between two predictors:

```
lm(response ~ predictor1 * predictor2, data = dataset)
```

This expands to `predictor1 + predictor2 + predictor1:predictor2`. The main effects are always included when an interaction is specified this way. I will explore interaction effects in more detail later in [Chapter 15](#).

When the model contains both continuous and categorical predictors, the formula is extended accordingly:

```
lm(response ~ continuous_predictor1 + continuous_predictor2 +
    factor(categorical_predictor),
    data = dataset)
```

R handles dummy-variable coding of factors automatically. This is one reason multiple regression and ANOVA belong to the same linear-model family.

14.7 EXAMPLE 1: A CLIMATE MODEL FOR THE EAST COAST TRANSITION ZONE

The seaweed dataset from Smit et al. (2017) contains:

- γ , mean Sørensen dissimilarity between adjacent seaweed assemblages along the South African coast;
- several continuous climatological predictors derived from sea-surface temperature records;
- bio, a categorical bioregional classification.

```
sw <- read.csv(here::here("data", "BCB743", "seaweed", "spp_df2.csv"))
rbind(head(sw, 3), tail(sw, 3))[, -1]
```

```
R>      dist bio    augMean  febRange    febSD    augSD    annMean
R> 1      0.000 BMP 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
R> 2     51.138 BMP 0.05741369 0.09884404 0.16295271 0.3132800 0.01501846
R> 3    104.443 BMP 0.15043904 0.34887754 0.09934163 0.4188239 0.02602247
R> 968 102.649 ECTZ 0.41496099 0.11330069 0.24304493 0.7538546 0.52278161
R> 969  49.912 ECTZ 0.17194242 0.05756093 0.18196664 0.3604341 0.24445006
R> 970   0.000 ECTZ 0.00000000 0.00000000 0.00000000 0.0000000 0.00000000
```



```
R>           Y           Y1           Y2
R> 1  0.000000000 0.0000000 0.000000000
R> 2  0.003610108 0.0000000 0.003610108
R> 3  0.003610108 0.0000000 0.003610108
R> 968 0.198728140 0.1948882 0.003839961
R> 969 0.069337442 0.0443038 0.025033645
R> 970 0.000000000 0.0000000 0.000000000
```

I begin with the East Coast Transition Zone (ECTZ) and ask which combination of climate variables best explains variation in Y .

```
sw_ectz <- sw %>%
  filter(bio == "ECTZ")

sw_sub1 <- sw_ectz[, c("Y", "augMean", "febRange", "febSD", "augSD", "annMean")]
```

14.7.1 Exploratory Data Analysis

Before fitting any model, I summarise the candidate variables and examine each predictor against the response.

```
summary(sw_sub1)
```

```
R>           Y           augMean           febRange           febSD
R> Min.      :0.00000   Min.      :0.0000   Min.      :0.0000   Min.      :0.0000
R> 1st Qu.:0.09615   1st Qu.:0.1817   1st Qu.:0.1866   1st Qu.:0.1820
R> Median :0.22779   Median :0.4686   Median :0.7531   Median :0.6256
R> Mean     :0.24393   Mean     :0.5313   Mean     :0.8329   Mean     :0.7209
R> 3rd Qu.:0.35690   3rd Qu.:0.7927   3rd Qu.:1.3896   3rd Qu.:1.2072
R> Max.     :0.63754   Max.     :1.7508   Max.     :2.2704   Max.     :2.1249
R>           augSD           annMean
R> Min.      :0.0000   Min.      :0.0000
R> 1st Qu.:0.4178   1st Qu.:0.1874
R> Median :1.0454   Median :0.5228
R> Mean     :1.3207   Mean     :0.5722
R> 3rd Qu.:2.0924   3rd Qu.:0.8085
R> Max.     :4.5267   Max.     :1.8941
```

```
predictors <- c("augMean", "febRange", "febSD", "augSD", "annMean")

models <- map(predictors, ~ lm(as.formula(paste("Y ~", .x)),
  data = sw_ectz))

plts1 <- map(predictors, function(predictor) {
  ggplot(sw_ectz, aes(x = .data[[predictor]], y = Y)) +
    geom_point(shape = 1, colour = "dodgerblue4") +
    geom_smooth(method = "lm", col = "magenta", fill = "pink") +
    labs(title = paste("Y vs", predictor),
         x = predictor,
         y = "Y") +
    theme_grey()
})

ggpubr::ggarrange(plotlist = plts1, ncol = 2,
  nrow = 3, labels = "AUTO")
```

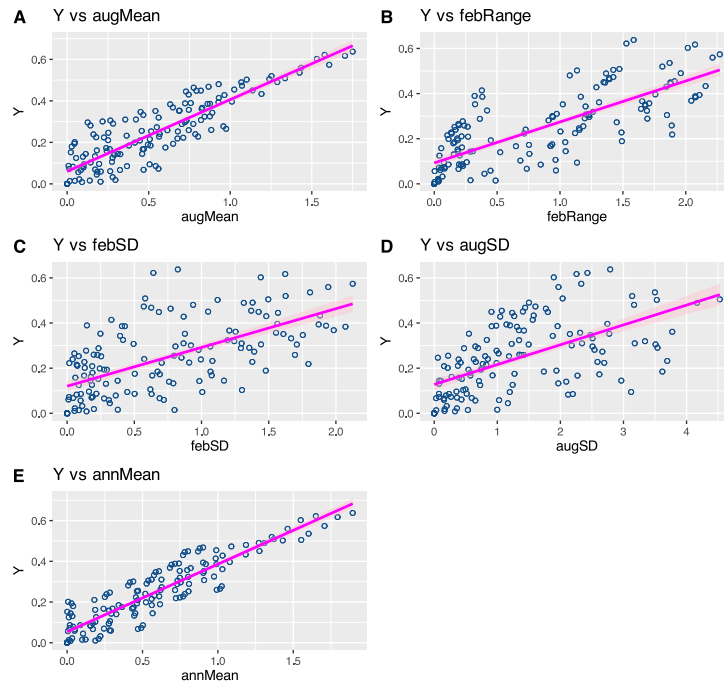


FIGURE 14.1. Simple linear regressions fitted separately for five candidate climate predictors against mean Sørensen dissimilarity (Y) in the East Coast Transition Zone.

Figure 14.1 shows that each predictor relates approximately linearly and positively to Y when examined alone. To make the strength of each marginal relationship precise, I extract the slope, its standard error, the coefficient of determination, and the overall model p -value from each fitted model and assemble them into Table 14.1.

```
fml_pval <- function(p) {
  case_when(
    p < 0.001 ~ "< 0.001",
    p < 0.01 ~ paste0("=", formatC(p, digits = 3, format = "f")),
    p < 0.05 ~ paste0("=", formatC(p, digits = 3, format = "f")),
    TRUE ~ paste0("=", formatC(p, digits = 3, format = "f"))
  )
}

slr_table <- map2_dfr(predictors, models, function(pred, mod) {
  tidy_mod <- tidy(mod)
  glance_mod <- glance(mod)
  tibble(
    Predictor = pred,
    Slope = tidy_mod$estimate[2],
    SE = tidy_mod$std.error[2],
    `R^2` = glance_mod$r.squared,
    p_val = fml_pval(tidy_mod$p.value[2])
  )
})
```

TABLE 14.1. Summary statistics for the five simple linear regressions of mean Sørensen dissimilarity (Y) against individual climate predictors in the East Coast Transition Zone. Slope and SE are the partial coefficient and its standard error. R^2 is the coefficient of determination; p is the overall model p -value from the F -test against the intercept-only null.

Predictor	Slope	SE	R^2	F	p
augMean	0.346	0.011	0.778	1,007.94	< 0.001
febRange	0.182	0.009	0.592	416.38	< 0.001
febSD	0.172	0.012	0.398	190.12	< 0.001
augSD	0.088	0.007	0.342	149.13	< 0.001
annMean	0.332	0.009	0.837	1,468.83	< 0.001

```

`F`      = glance_mod$statistic,
`p`      = fmt_pval(glance_mod$p.value)
)
})

slr_table >
  gt() >
  fmt_number(columns = c(Slope, SE, `R^2`), decimals = 3) >
  fmt_number(columns = `F`, decimals = 2)

```

All five predictors are individually significant at $\alpha = 0.05$, and each accounts for a substantial proportion of the variance in Y when fitted alone. These simple regressions are not checked for distributional assumptions; residual normality and homoscedasticity will be assessed properly on the later correctly specified model. Checking assumptions on exploratory marginal fits would be premature since the residuals from a single-predictor model still contain all the variance that the remaining predictors explain, so any assumption violations diagnosed here may simply reflect omitted predictors rather than genuine model failure. The relevant question at this stage is whether each predictor is plausibly related to the response and approximately linear in that relationship. Figure 14.1 and Table 14.1 together confirm that both conditions are met.

These one-predictor views do not show how the predictors behave when considered together. In particular, they give no indication of how much the predictors overlap with each other.

DO IT NOW!

Refer to Table 14.1, then answer:

1. Which predictor alone explains the most variance in Y ?
2. Does the predictor with the steepest slope also have the highest R^2 ? If not, why not?
3. Would you include all five in the multiple regression at this stage? What would be the risk?

14.7.2 *Hypotheses*

The null hypothesis for the overall model is that none of the climatological variables explains variation in Y :

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

The alternative is that at least one coefficient differs from zero:

$$H_a : \exists \beta_i \neq 0 \text{ for at least one } i$$

At the coefficient level, each predictor has its own hypothesis:

$$H_0 : \beta_i = 0 \quad H_a : \beta_i \neq 0$$

These individual tests must be read as partial tests because each evaluates the contribution of that predictor given the others in the model.

14.7.3 *Fit the Initial Model*

I fit a **full model** containing all five climate predictors and a **null model** containing only the intercept.

```
null_mod1 <- lm(Y ~ 1, data = sw_sub1)
full_mod1 <- lm(Y ~ augMean + febRange + febSD + augSD + annMean,
               data = sw_sub1)

summary(null_mod1)
```

```
R>
R> Call:
R> lm(formula = Y ~ 1, data = sw_sub1)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.24393 -0.14778 -0.01614  0.11297  0.39361
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.24393      0.00963   25.33  <2e-16 ***
```

```
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.1637 on 288 degrees of freedom
summary(full_mod1)
```

```
R>
R> Call:
R> lm(formula = Y ~ augMean + febRange + febSD + augSD + annMean,
R>     data = sw_sub1)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.131229 -0.042352 -0.003415  0.038601  0.144128
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.044886   0.006306   7.118 9.04e-12 ***
R> augMean     -0.079925   0.042630  -1.875 0.061841 .
R> febRange     0.112913   0.015949   7.080 1.14e-11 ***
R> febSD       -0.057174   0.016554  -3.454 0.000637 ***
R> augSD        0.003024   0.004886   0.619 0.536421
R> annMean      0.322776   0.041615   7.756 1.59e-13 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.05713 on 283 degrees of freedom
R> Multiple R-squared:  0.8804, Adjusted R-squared:  0.8782
R> F-statistic: 416.5 on 5 and 283 DF,  p-value: < 2.2e-16

ggplot(sw_sub1 > mutate(obs = row_number()), aes(x = obs, y = Y)) +
  geom_point(shape = 1, colour = "dodgerblue4") +
  geom_smooth(method = "lm", formula = y ~ 1,
             col = "magenta", fill = "pink", linewidth = 0.4, alpha = 0.4) +
  labs(x = "Observation", y = "Y") +
  theme_grey()
```

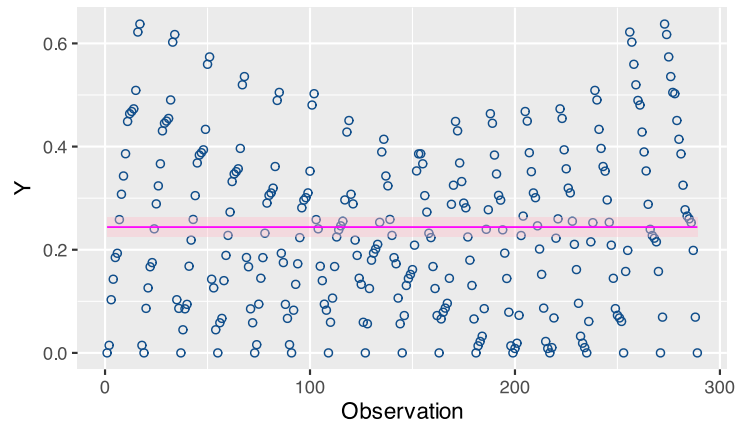


FIGURE 14.2. The null model fitted to Sørensen dissimilarity (Y) in the East Coast Transition Zone. Each point is one observation; the magenta line is the null model's single estimate — the grand mean of Y — and the pink band is its 95% confidence interval. Every residual is a vertical deviation from this horizontal line.

The **null model** (`lm( $Y \sim 1$ , data = sw_sub1)`) estimates a single parameter, the *grand mean* of Y (Figure 14.2). Here, the $Y \sim 1$ specifies the intercept-only, so there are no predictors on the right-hand side. The null's fitted value is identical for every observation and its residuals are the deviations of each observation from that mean. The null model is often used as the baseline against which any more fully-specified model can be compared using a nested test that asks whether adding predictors reduces residual variance by more than chance alone.

NOTE

`summary()` on a fitted `lm` object computes the overall F -statistic internally, by comparing the fitted model's residual sum of squares with the total sum of squares around the grand mean of Y . That total sum of squares is what the null model's residuals would give, and R derives it directly from the response vector without needing the null model object. I explicitly fit the `null_mod1` here for two reasons: i) to make the idea of the null model visible as a named model object, and to allow a direct nested comparison using `anova()`, which is demonstrated below.

Calling `anova(null_mod1, full_mod1)` does the nested comparison and asks whether adding all five climate predictors to the intercept-only baseline reduces residual variance by more than would be expected by chance.

```
anova(null_mod1, full_mod1)
```

```
R> Analysis of Variance Table
R>
R> Model 1: Y ~ 1
R> Model 2: Y ~ augMean + febRange + febSD + augSD + annMean
R>   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
R> 1      288 7.7192
R> 2      283 0.9235   5    6.7957 416.49 < 2.2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F -statistic and p -value here are numerically identical to the overall model F reported at the bottom of `summary(full_mod1)` above. The difference is that `anova()` reveals the comparison's details by naming both models and printing the residual sums of squares for each. Seeing both outputs side by side is useful because it shows that `summary()`'s overall F -test is exactly a nested model comparison against the intercept-only baseline, even when that baseline model is never explicitly fitted.

The full model explains a large proportion of the variance in Y , but the coefficient table already hints at a problem: two predictors that were individually significant in Table 14.1 are no longer clearly distinguishable from zero once all five are included simultaneously. This pattern is characteristic of multicollinearity.

14.7.4 Check for Multicollinearity

The **variance inflation factor** (VIF) for a predictor X_j measures how much the variance of $\hat{\beta}_j$ is inflated relative to what it would be if X_j were uncorrelated with the other predictors. A VIF of 1 indicates no collinearity with the remaining predictors. Values above 5 begin to suggest instability in the coefficient estimate, and values above 10 usually make interpretation unreliable. The formula is:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination from regressing X_j on all other predictors in the model. When R_j^2 is large, X_j can be reconstructed almost entirely from the other predictors, and its partial coefficient becomes imprecise.

```
vif(full_mod1)
```

```
R>   augMean febRange   febSD   augSD annMean
R> 27.947767 10.806635  8.765732  2.497739 31.061900
```

Several predictors have VIF values well above 10, indicating that the full five-predictor model is not interpretable as specified. The solution in this case is to remove the most problematic predictor and refit, iterating until VIF values are acceptable. This is a pragmatic response to collinearity within the data at hand, but it does not

resolve the deeper question of why the variables are intercorrelated or what is lost when a variable is removed. [Chapter 16](#) addresses those questions.

DO IT NOW!

Fit the full five-predictor model and check its VIFs. Then remove the single predictor with the highest VIF, refit, and check again. Record which predictor was removed and how the remaining VIF values changed.

```
library(car)
full_mod1 <- lm(Y ~ augMean + febRange + febSD + augSD + annMean,
               data = sw_sub1)
vif(full_mod1)
```

After removing the worst offender (say annMean):

```
mod_reduced <- lm(Y ~ augMean + febRange + febSD + augSD,
                 data = sw_sub1)
vif(mod_reduced)
```

Discuss with a partner:

1. Which predictor had the highest VIF, and what does that say about its relationship with the others?
2. Did removing one predictor bring all remaining VIFs below 10, or is another iteration needed?
3. What information is potentially lost by removing a predictor this way?

Note also the distinction from confounding: collinearity is a property of the included predictors since they carry overlapping information. Confounding involves a predictor that was omitted entirely but correlates both with an included predictor and with the response, causing the included predictor's coefficient to be biased. Both problems undermine interpretation, but through different mechanisms.

14.7.5 Reduce Multicollinearity

The following iterative function removes the predictor with the highest VIF at each step until all remaining VIF values fall below a threshold of 10.

```
threshold <- 10
vif_mod <- full_mod1
sw_vif <- sw_sub1

while (TRUE) {
  vif_values <- vif(vif_mod)
  max_vif <- max(vif_values)

  if (max_vif > threshold) {
    high_vif_var <- names(which.max(vif_values))
    sw_vif <- sw_vif[, !(names(sw_vif) %in% high_vif_var)]
    vif_mod <- lm(Y ~ ., data = sw_vif)
  } else {
```



```

    } break
  }
}

summary(vif_mod)

R>
R> Call:
R> lm(formula = Y ~ ., data = sw_vif)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.153994 -0.049229 -0.006086  0.045947  0.148579
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.028365   0.007020   4.040 6.87e-05 ***
R> augMean      0.283335   0.011131  25.455 < 2e-16 ***
R> febSD        0.049639   0.008370   5.930 8.73e-09 ***
R> augSD        0.022150   0.004503   4.919 1.47e-06 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.06609 on 285 degrees of freedom
R> Multiple R-squared:  0.8387, Adjusted R-squared:  0.837
R> F-statistic: 494.1 on 3 and 285 DF,  p-value: < 2.2e-16

vif(vif_mod)

R> augMean  febSD  augSD
R> 1.423601 1.674397 1.585055

```

The retained predictors are `augMean`, `febSD`, and `augSD`, with VIF values now well below the threshold. Regularisation methods such as ridge regression or LASSO can also handle collinearity without discarding predictors, and these are introduced in [Chapter 25](#).

14.7.6 Fit the Selected Model

Still, not all three predictors retained after VIF screening may be needed to explain the response. I use **forward stepwise AIC** to identify the most parsimonious model within the reduced candidate set.

VIF screening and stepwise AIC address different problems and must be applied in that order. VIF screening is a stability check to see whether the coefficient estimates are numerically reliable given the intercorrelations among the predictors. If two predictors share most of their information, their partial coefficients become imprecise and their standard errors inflate, regardless of how well either predicts the response.

AIC-based selection cannot fix multicollinearity because it will choose among poorly identified models and may retain collinear predictors simply because their combined redundancy happens to reduce the AIC criterion slightly. Therefore, re-

moving collinear predictors first produces a candidate variable set in which every remaining predictor carries reasonably independent information. Selection then operates on a set of well-identified models and the AIC differences between them are interpretable. Applying stepAIC to the original five-predictor variable set, before VIF screening, would risk retaining a subset of correlated predictors whose coefficients look stable in the selected model only because the worst offenders were excluded by chance during the search.

```
mod1 <- stepAIC(
  lm(Y ~ 1, data = sw_vif),
  scope = list(lower = lm(Y ~ 1, data = sw_vif), upper = vif_mod),
  direction = "forward",
  trace = FALSE
)
```

```
formula(mod1)
```

```
R> Y ~ augMean + febSD + augSD
```

```
summary(mod1)
```

```
R>
R> Call:
R> lm(formula = Y ~ augMean + febSD + augSD, data = sw_vif)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.153994 -0.049229 -0.006086  0.045947  0.148579
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.028365    0.007020   4.040 6.87e-05 ***
R> augMean      0.283335    0.011131  25.455 < 2e-16 ***
R> febSD        0.049639    0.008370   5.930 8.73e-09 ***
R> augSD        0.022150    0.004503   4.919 1.47e-06 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.06609 on 285 degrees of freedom
R> Multiple R-squared:  0.8387, Adjusted R-squared:  0.837
R> F-statistic: 494.1 on 3 and 285 DF, p-value: < 2.2e-16
```

```
anova(mod1)
```

```
R> Analysis of Variance Table
R>
R> Response: Y
R>      Df Sum Sq Mean Sq F value    Pr(>F)
R> augMean  1 6.0084  6.0084 1375.660 < 2.2e-16 ***
R> febSD    1 0.3604  0.3604  82.507 < 2.2e-16 ***
R> augSD    1 0.1057  0.1057  24.196 1.473e-06 ***
R> Residuals 285 1.2448  0.0044
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NOTE

We include the null model as part of the `stepAIC()` function. The fitted model therefore estimates a single value (the grand mean of Y) and uses it as the prediction for every observation. In `stepAIC()`, this sets the starting point (for forward selection) or the lower boundary (for both forward and stepwise) to prevent the search from stepping below an intercept-only baseline.

The selected model is:

$$Y = \alpha + \beta_1 \text{augMean} + \beta_2 \text{febSD} + \beta_3 \text{augSD} + \epsilon$$

14.7.7 Reading the Coefficient and ANOVA Tables

`summary()` and `anova()` each answer a different question, and both appear above.

The **coefficient table** from `summary()` tests each partial effect *conditional* on all other predictors in the model. The t -statistic for `augMean` asks: given that `febSD` and `augSD` are already in the model, does `augMean` still contribute? Each test conditions on the full model as specified.

Each $\hat{\beta}$ is the rate of change in Y per unit change in that predictor, with all others held constant. For `augMean`, $\hat{\beta}_1 = 0.2833$: each additional degree Celsius of August mean temperature corresponds to a 0.28-unit increase in Sørensen dissimilarity. Sørensen spans 0 to 1, so that is a steep gradient (roughly a quarter of the entire response range per degree) reflecting the strong thermal structuring of seaweed assemblages along the South African coast.

The **ANOVA table** from `anova()` is *sequential*. It partitions the total sum of squares in the order the predictors were entered: `augMean` is assessed first, then `febSD` gets credit for what `augMean` left unexplained, and `augSD` gets the remainder. Change the order of predictors in the formula and the sums of squares in the ANOVA table change.

For testing whether a predictor contributes after adjusting for all others, the coefficient table is the appropriate place to look. The ANOVA table is more informative when the question is about the total variance attributable to each predictor in a natural sequence, or when comparing nested models (see [Chapter 15](#)).

14.7.8 Scaling and Centring

The coefficient for `augMean` gives the expected change in Y for a one-unit increase in August mean temperature, in degrees Celsius. When predictors are on very different measurement scales, comparing their coefficients directly is not meaningful, so a

one-unit change in temperature and a one-unit change in temperature variability are not equivalent steps. **Standardising** predictors to zero mean and unit variance before fitting makes their coefficients comparable in the sense that each then represents the change in Y per one standard deviation of the predictor. **Centring** alone (subtracting the mean without rescaling) does not change the slopes but makes the intercept interpretable as the fitted value of Y at average predictor values rather than at predictor values of zero, which are often outside the data range entirely.

14.7.9 Check Diagnostics

Added-variable plots show the relationship between the response and each predictor after accounting for all the others in the model. They are the appropriate visual analogue to partial coefficients.

```
avPlots(mod1, col = "dodgerblue4", col.lines = "magenta")
```

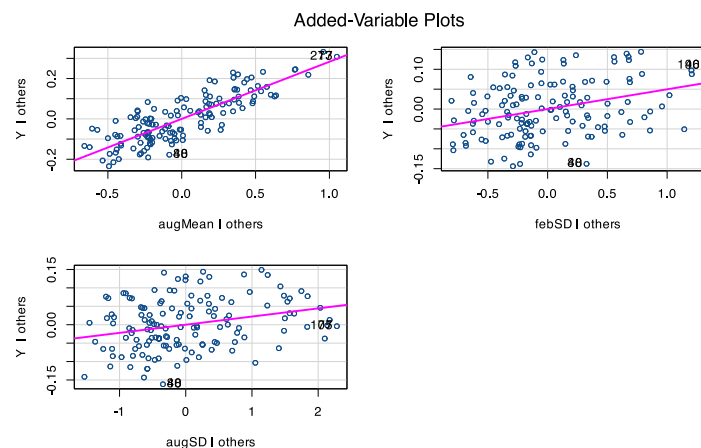


FIGURE 14.3. Added-variable (partial regression) plots for the selected model `mod1`. Each panel shows the relationship between the response and one predictor after removing the effects of the other two predictors.

In Figure 14.3, a tight, steeply sloped partial relationship indicates a strong predictor given the others. `augMean` shows a clear, strong partial relationship. `febSD` and `augSD` show weaker but positive partial associations. If a partial plot showed a nonlinear pattern, that would suggest the predictor needs a transformation or polynomial term. If a partial plot showed no relationship at all, that predictor contributes little once the others are accounted for.

We also want to view the normal diagnostics plots to assess the normality and variance expectations, and find outliers if they exist.

```
autoplot(mod1, shape = 21, colour = "dodgerblue4",
          smooth.colour = "magenta") +
  theme_grey()
```

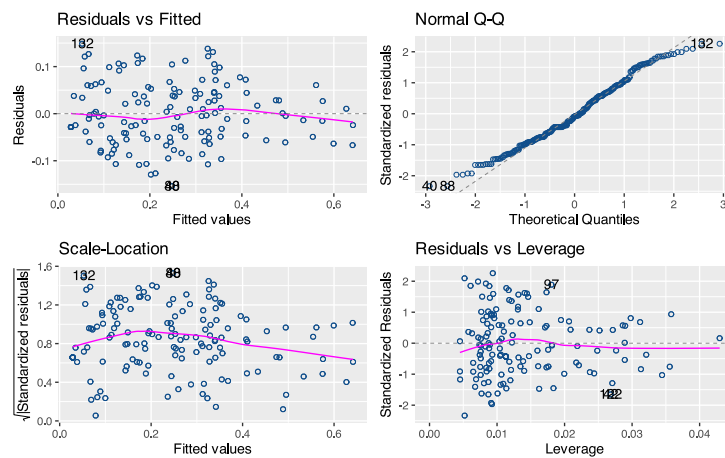


FIGURE 14.4. Standard diagnostic plots for mod1. Top left: residuals vs fitted values; top right: normal Q-Q; bottom left: scale-location; bottom right: residuals vs leverage.

Reading Figure 14.4:

- **Residuals vs fitted:** Points should scatter around zero without a systematic trend. A curve here suggests a missing predictor or a nonlinear term. Fanning, the tendency to display an increasing spread with fitted values, indicates heteroscedasticity and may require a response transformation.
- **Normal Q-Q:** Points should fall close to the diagonal. Systematic departure at both tails suggests non-normality; a heavy right tail in particular may require a log transformation of the response.
- **Scale-location:** The square root of absolute residuals should be roughly horizontal, and a rising trend confirms heteroscedasticity.
- **Residuals vs leverage:** Points with high leverage and large residuals are influential. The dashed contour lines mark Cook's distance of 0.5 and 1.0, and points outside those contours should be examined individually to confirm that they represent genuine observations and not data entry errors or atypical conditions.

My overall interpretation of Figure 14.4 is that the model looks broadly acceptable. There is no severe curvature, no strong fanning, and no obviously influential outliers. Adequacy is a judgement, not a binary pass or fail.

The last plt that can help us assess the fit of a multiple regression is a **component-plus-residual plot**.

```
crPlots(mod1, col = "dodgerblue4", col.lines = "magenta")
```

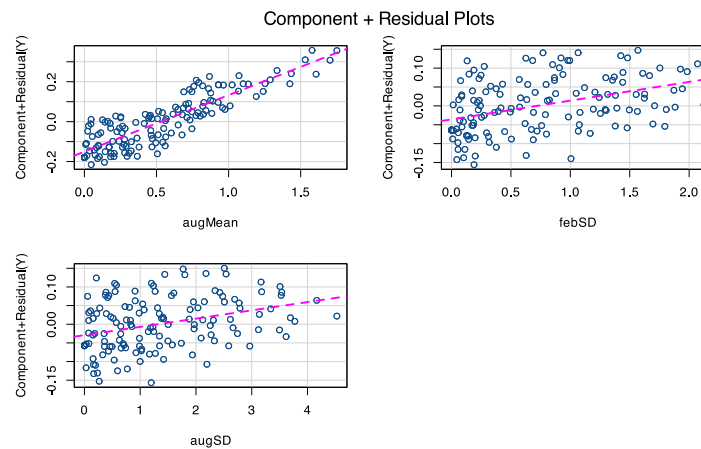


FIGURE 14.5. Component-plus-residual plots for mod1. Each panel shows whether the relationship between the response and that predictor is approximately linear in the fitted model.

The component-plus-residual plots in Figure 14.5 are an additional linearity check. A straight pink line against the data cloud confirms that a linear term is adequate for that predictor in the model. Curvature in any panel would suggest adding a polynomial or transformed version of that predictor.

NOTE

The added-variable plots in Figure 14.3 and the component-plus-residual plots in Figure 14.5 address different questions and plot different quantities against each other, which is why they can look similar without being equivalent.

An **added-variable plot** for predictor X_j plots two sets of residuals against each other: the residuals from regressing Y on all predictors *except* X_j , against the residuals from regressing X_j on all those same other predictors. Both axes have been purged of the other predictors' effects. The slope of the line in that panel is exactly the partial coefficient $\hat{\beta}_j$. What the plot reveals is the partial relationship; that is, the strength

and direction of X_j 's contribution to Y after everything else in the model has been removed from both variables. If X_j has no partial effect, the cloud in its panel will be horizontal. If X_j 's partial effect is nonlinear, the cloud will curve.

A **component-plus-residual plot** (also called a partial-residual plot) for predictor X_j plots $e_i + \hat{\beta}_j X_{ji}$ on the vertical axis — the raw residuals with the fitted linear component of X_j added back in — against X_{ji} itself on the horizontal axis. The idea is to reconstruct what the response would look like as a function of X_j alone, after adjusting for the other predictors. The slope of the fitted line in the panel is again $\hat{\beta}_j$, but now the vertical axis is on the original scale of the predictor rather than on a residualised scale.

The practical difference is that added-variable plots are better for assessing the *existence* and *strength* of a partial effect and for detecting influential observations, because both axes are free of the other predictors. Component-plus-residual plots are better for diagnosing the *functional form* of each predictor's contribution, because the vertical axis retains the scale of X_j and any smooth curve fitted through the cloud reveals whether a linear term is adequate or whether a nonlinear term is needed. In both plots, the line should be roughly straight; but it is the component-plus-residual plot where departures from linearity are most directly interpretable as evidence that a predictor needs to be transformed.

14.7.10 Interpret the Results

The three retained predictors are August mean temperature (`augMean`), February temperature standard deviation (`febSD`), and August temperature standard deviation (`augSD`). They are all positively associated with mean Sørensen dissimilarity in the East Coast Transition Zone. The strongest partial effect is `augMean`, while `febSD` and `augSD` make smaller but positive contributions.

Reading the partial coefficients: a 1 °C increase in August mean temperature is associated with an increase in mean Sørensen dissimilarity of approximately $\hat{\beta}_1$, holding February and August temperature variability constant. That “holding constant” qualification is the central point of the conditional interpretation. So, if the predictors were uncorrelated, this would simplify to the marginal effect seen in Figure 14.1. Because they are moderately correlated even after VIF screening, the partial coefficients differ somewhat from the simple-regression slopes.

The adjusted R^2 is approximately 0.84, indicating that the selected predictors account for a large proportion of the observed variation in Y . The ordinary R^2 is the proportion of total variance in the response explained by the model:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

It never decreases when a predictor is added, even if that predictor carries no real information, because adding any variable, however useless, can only reduce or hold steady the residual sum of squares SS_{res} . In a multiple regression with several predictors, this means R^2 will overstate explanatory power if the model was arrived at by adding predictors freely. The adjusted R^2 corrects for this by penalising for the number of predictors p relative to the sample size n :

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{res}}/(n - p - 1)}{SS_{\text{tot}}/(n - 1)}$$

The denominator of the penalty term, $n - p - 1$, is the residual degrees of freedom. As p increases, $n - p - 1$ shrinks, the ratio $SS_{\text{res}}/(n - p - 1)$ grows, and R^2_{adj} falls. That is, unless the additional predictor explains enough variance to compensate. The adjusted R^2 can therefore decrease when a weak predictor is added, which R^2 cannot. For comparing models with different numbers of predictors, the adjusted R^2 is the more honest summary of fit.

The partial coefficients are not causal estimates. The model does not rule out the possibility that some shared, unmeasured environmental driver explains both the temperatures and the dissimilarity. They are adjusted associations within the fitted model.

14.7.11 Report

WRITE-UP

Methods

Mean Sørensen dissimilarity in the East Coast Transition Zone was modelled as a function of candidate climate predictors using multiple linear regression. Collinearity among predictors was assessed using variance inflation factors; predictors with VIF above 10 were removed iteratively until the remaining predictors were below that threshold. Forward stepwise AIC was used to identify the most parsimonious model within the reduced candidate set. Assumptions were assessed using added-variable plots, standard diagnostic plots, and component-plus-residual plots.

Results

Mean Sørensen dissimilarity in the East Coast Transition Zone was strongly explained by a multiple linear regression containing August mean temperature, February temperature variability, and August temperature variability (adjusted $R^2 = 0.84$; overall model F -test $p < 0.001$). All three retained predictors had positive partial effects. The 95% confidence intervals confirm that the partial effects of `augMean` and `febSD` exclude zero comfortably, while the interval for `augSD` is positive but wider.

Discussion

The model identifies a small set of comparatively stable climatic predictors of compositional turnover in the East Coast Transition Zone. The dominant partial effect of August mean temperature is consistent with the pronounced temperature gradient along this coastline, which separates warm-temperate assemblages of the east from cool-water communities to the south. February temperature variability and August temperature variability contribute additional signal, reflecting that compositional differences are also sensitive to thermal range and seasonal extremes.

14.8 EXAMPLE 2: CATEGORICAL PREDICTORS — DISTANCE AND BIOREGION

Multiple regression can include both continuous and categorical predictors. In R this is done by entering the categorical variable as a factor. Once a factor is included, multiple regression and ANOVA become part of the same linear-model family, since both are fitted by `lm()` and both produce least-squares estimates within the same framework.

14.8.1 Dummy Variable Coding

When a *categorical variable* with k levels is included in a regression, R represents it using $k - 1$ dummy variables. One level is taken as the reference, and its contribution is absorbed into the model intercept. Each remaining level receives a coefficient that gives the difference in fitted intercept relative to that reference.

The four seaweed bioregions are AMP, B-ATZ, BMP, and ECTZ. With AMP as the reference level, the additive model containing both geographic distance and bioregion is:

$$Y_i = \alpha + \beta_1 \text{dist}_i + \gamma_1 D_{i,B-ATZ} + \gamma_2 D_{i,BMP} + \gamma_3 D_{i,ECTZ} + \epsilon_i$$

where each D is an indicator variable: it equals 1 for observations belonging to that bioregion and 0 for all others.

a) What the Equation Says. Look at the equation term by term. The term $\beta_1 \text{dist}_i$ contributes the same amount to the predicted Y regardless of which bioregion site i belongs to, so the slope β_1 is shared across all regions. The dummy terms switch on or off depending on the bioregion. For a site in AMP, all three dummies are zero, so the predicted value simplifies to $\alpha + \beta_1 \text{dist}_i$. For a site in B-ATZ, only $D_{i,B-ATZ} = 1$, so the predicted value is $\alpha + \beta_1 \text{dist}_i + \gamma_1$. The γ coefficients therefore describe *shifts in the intercept*, so they adjust the expected baseline level of dissimilarity for each bioregion up or down from the AMP reference, independently of how far apart the sites are.

Put differently: the coefficient γ_1 is not the mean Y for B-ATZ. It is the *difference* between the average level of Y in B-ATZ and the average level of Y in AMP, after accounting for distance. A positive $\hat{\gamma}_2$ for BMP means that, at any given geographic distance, sites in BMP show higher mean dissimilarity than sites in AMP. This is what the model is designed to address: **does the average level of Sørensen dissimilarity differ among bioregions once distance is accounted for?**

b) *What it Does Not Say.* The model has only one slope for dist , β_1 . That single value applies equally in AMP, B-ATZ, BMP, and ECTZ. If you drew the fitted regression line for each bioregion, all four lines would be parallel, that is, they share the same gradient but they have different heights. The model therefore cannot answer whether dissimilarity *accumulates with distance faster or slower in some bioregions than in others*. Whether the lines should instead have different slopes is a separate biological question entirely: it asks whether the rate of compositional turnover with distance differs among bioregions, not merely whether the baseline level differs. Answering that question requires adding an interaction term, $\text{dist}:\text{bio}$, which allows each bioregion to have its own slope. That extension is the subject of [Chapter 15](#).

For this chapter, the additive structure is the right tool for the questions being asked. The bioregion coefficients in `summary()` are intercept adjustments. They describe *how high or low the mean dissimilarity is* in each region relative to AMP, after removing the distance effect. They say nothing about the steepness of that distance effect in any particular region.

c) *A Visual Representation.* Figure 14.6 shows the reasoning in two steps. Panel A shows the simplest possible model: $Y \sim \text{bio}$ with no distance predictor at all. The fitted value for each bioregion is just its group mean, producing a horizontal line for every region. The lines have no slope because distance plays no role here. This is also precisely what a one-way ANOVA produces. The vertical gaps between the horizontal lines represent the raw mean differences among bioregions in Y .

Panel B adds dist to the model: $Y \sim \text{dist} + \text{bio}$. All four horizontal lines from panel A now tilt by the same angle, acquiring the shared slope $\hat{\beta}_1$. Their vertical separation is preserved, so the gaps between the lines are unchanged from panel A, because the bioregion coefficients still describe the same intercept differences. What the slope adds is the ability to account for the tendency of dissimilarity to increase with geographic distance, and in this model, it does so by the same amount in every bioregion.

That vertical gap between any two lines (constant across the entire x-axis in panel B) is what the bioregion coefficients measure. For example, the coefficient $\hat{\gamma}_2$

for BMP is the vertical distance between the BMP line and the AMP line at any point along the distance axis.

```
fit_anova <- lm(Y ~ bio, data = sw)
fit_ancova <- lm(Y ~ dist + bio, data = sw)

dist_range <- seq(min(sw$dist), max(sw$dist), length.out = 200)
bios <- unique(sw$bio)

pred_anova <- expand.grid(dist = dist_range, bio = bios)
pred_anova$fit <- predict(fit_anova, newdata = pred_anova)

pred_ancova <- expand.grid(dist = dist_range, bio = bios)
pred_ancova$fit <- predict(fit_ancova, newdata = pred_ancova)

p_a <- ggplot() +
  geom_point(data = sw, aes(x = dist, y = Y, colour = bio),
            shape = 1, alpha = 0.4) +
  geom_line(data = pred_anova, aes(x = dist, y = fit, colour = bio),
           linewidth = 0.9) +
  scale_colour_brewer(palette = "Set1", name = "Bioregion") +
  labs(x = "Distance (km)", y = "Mean Sørensen Dissimilarity",
       subtitle = "Y ~ bio (group means only; no slope)") +
  theme_grey()

p_b <- ggplot() +
  geom_point(data = sw, aes(x = dist, y = Y, colour = bio),
            shape = 1, alpha = 0.4) +
  geom_line(data = pred_ancova, aes(x = dist, y = fit, colour = bio),
           linewidth = 0.9) +
  scale_colour_brewer(palette = "Set1", name = "Bioregion") +
  labs(x = "Distance (km)", y = "Mean Sørensen Dissimilarity",
       subtitle = "Y ~ dist + bio (shared slope, different intercepts)") +
  theme_grey()

ggarrange(p_a, p_b, ncol = 2, labels = "AUTO", common.legend = TRUE,
          legend = "right")
```

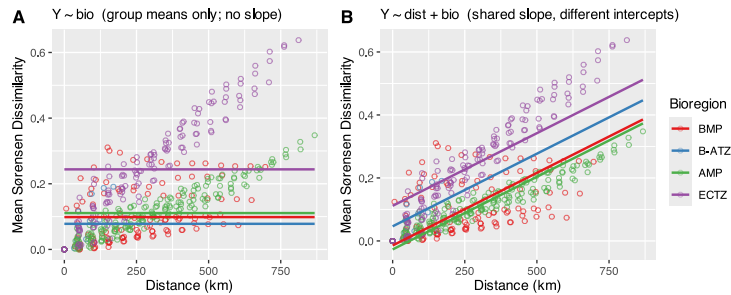


FIGURE 14.6. (A) A one-way ANOVA model ($Y \sim \text{bio}$) produces one horizontal fitted line per bioregion, sitting at each group's mean dissimilarity. Distance plays no role. (B) The additive model ($Y \sim \text{dist} + \text{bio}$) tilts all four lines by the same slope but preserves their vertical separation. The lines are strictly parallel: same gradient, different heights. The vertical gap between any two lines equals the difference in their bioregion coefficients.

The parallelism in panel B is the algebraic consequence of the additive structure. If the lines were *not* parallel (*i.e.*, if the slope of `dist` differed among bioregions), the additive model would be the wrong model. Whether that is the case here is the question that motivates the interaction models in [Chapter 15](#).

```
bio_demo <- tibble(
  bio = factor(c("AMP", "B-ATZ", "BMP", "ECTZ"),
    levels = levels(factor(sw$bio)))
)

bio_model_matrix <- model.matrix(~ bio, data = bio_demo) >
  as.data.frame() >
  mutate(bio = bio_demo$bio) >
  relocate(bio)

bio_model_matrix >
  pivot_longer(cols = -bio, names_to = "term", values_to = "value") >
  ggplot(aes(x = term, y = bio, fill = factor(value))) +
  geom_tile(colour = "white") +
  scale_fill_manual(values = c("0" = "grey85", "1" = "steelblue4")) +
  labs(x = NULL, y = "Bioregion", fill = "Value") +
  theme_grey()
```

TABLE 14.2. Dummy-variable coding for the four seaweed bioregions when AMP is the reference level. Each row shows the values of the three indicator variables for that bioregion. AMP is coded entirely as zeros, so its contribution is captured by the model intercept.

bio	(Intercept)	bioB-ATZ	bioBMP	bioECTZ
AMP	1	0	0	0
B-ATZ	1	1	0	0
BMP	1	0	1	0
ECTZ	1	0	0	1

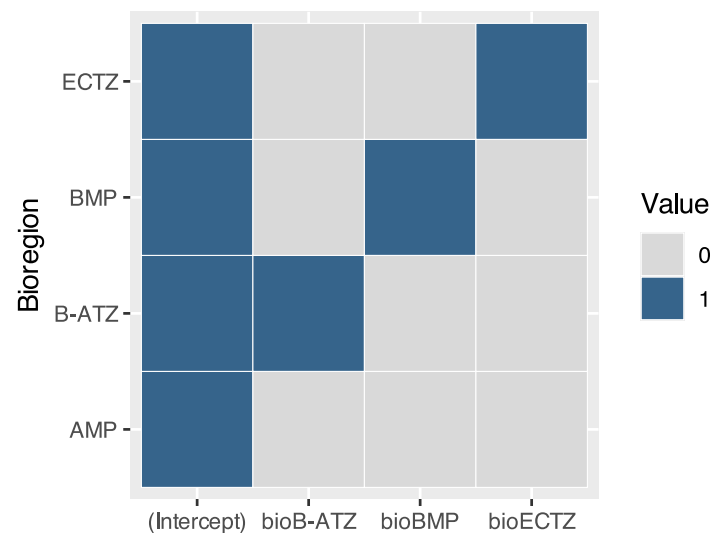


FIGURE 14.7. Visualisation of the dummy-variable coding shown in Table 14.2. AMP (reference level) has all zeros; the three other bioregions each have a single 1 in their corresponding indicator column.

Each row of Table 14.2 is a bioregion; each column after the first is a dummy variable. The first column, (Intercept), contains a 1 for every observation without exception. It is not a variable in the usual sense; it is the column that allows the model to estimate a constant term. Multiplying α by 1 for every row is how the intercept enters the linear predictor $\mathbf{X}\boldsymbol{\beta}$ in matrix notation. When you see (Intercept) in `summary()` output, the coefficient next to it is $\hat{\alpha}$, which is the fitted value of Y when

every other column in the model matrix is zero, which in this case means an AMP observation at $\text{dist} = 0$.

The three dummy columns come after it. The dummy for a given bioregion is 1 for that bioregion's row and 0 for every other row. AMP has a 0 in every dummy column — that is not an error or a gap; it is the definition of the reference level. When all dummies are zero, the model collapses to the intercept column alone, so $\hat{\alpha}$ is the fitted baseline for AMP.

Figure 14.7 shows the same information as a grid. The pattern clearly reveals that no two bioregions share a 1 in the same dummy variable column. Each bioregion is uniquely identified by exactly one blue cell, or, in the case of AMP, by the absence of blue cells.

Let's see what `summary()` does. When `sw$bio` is passed to `lm()` as a factor, R creates the three dummy columns shown in Table 14.2 and fits a coefficient to each. The coefficient table in `summary()` will contain rows named `bioB-ATZ`, `bioBMP`, and `bioECTZ`. It will *not* contain a row for `bioAMP`, because AMP is captured entirely by the intercept. Reading those output rows correctly requires remembering the table:

- For an AMP observation: all three dummies are 0. The fitted value from the categorical part of the model is α , i.e., the intercept alone.
- For a BMP observation: `bioBMP` = 1, the other two dummies are 0. The fitted value from the categorical part is $\alpha + \hat{\gamma}_2$, where $\hat{\gamma}_2$ is the coefficient of `bioBMP`. The model has shifted the fitted line for BMP up (or down) relative to AMP by exactly $\hat{\gamma}_2$.
- For an ECTZ observation: `bioECTZ` = 1, the others are 0. The fitted value is $\alpha + \hat{\gamma}_3$.

Changing the reference level with `relevel()` shifts which bioregion is absorbed into the intercept but does not change the fitted values, residuals, or the model's overall explanatory power. It only redistributes how the group differences are expressed in the coefficient table.

DO IT NOW!

Fit the additive model with `bio` as a factor and then refit it with ECTZ as the reference level instead of AMP. Compare the two `summary()` outputs.

```
sw$bio <- factor(sw$bio) # AMP is the default reference
mod_amp_ref <- lm(Y ~ dist + bio, data = sw)
summary(mod_amp_ref)

sw$bio <- relevel(sw$bio, ref = "ECTZ")
mod_ectz_ref <- lm(Y ~ dist + bio, data = sw)
summary(mod_ectz_ref)
```

Answer the following:

1. Do the fitted values change when you switch the reference level? Check with `all.equal(fitted(mod_amp_ref), fitted(mod_ectz_ref))`.
2. What does the intercept represent in each model?

3. The coefficient for bioECTZ in the first model is a comparison against AMP. What comparison does the coefficient for bioAMP represent in the second model?

14.8.3 Biological Question and Hypotheses

The two variables used in this example are:

- `dist` — geographic distance between adjacent sample sites along the South African coastline;
- `bio` — the bioregional classification of each site.

Now that the structure of the model equation is clear from the dummy-coding discussion above, I can precisely state the biological questions. The additive model is fitted to address two questions:

1. Is mean Sørensen dissimilarity related to geographic distance along the coast?
2. After accounting for distance, does mean dissimilarity differ among bioregions?

The null hypothesis for the distance effect is that geographic separation does not predict dissimilarity once bioregion is accounted for:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

The null hypothesis for the bioregion effect is that no bioregion differs from AMP in average dissimilarity after accounting for distance:

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$$

$$H_a : \gamma_i \neq 0 \text{ for at least one } i$$

The coefficient table from `summary()` tests each predictor's partial effect; the nested model comparison via `anova()` tests the overall contribution of `bio` beyond the distance-only model.

14.8.4 Visualise the Main Effects

```
p1 <- ggplot(sw, aes(x = dist, y = Y)) +
  geom_point(shape = 1, aes(colour = bio)) +
  geom_smooth(method = "lm", se = TRUE,
             colour = "black", fill = "grey60",
             linewidth = 0.4) +
  scale_colour_brewer(palette = "Set1") +
  labs(x = "Distance (km)",
```

```

y = "Sørensen Dissimilarity") +
theme_grey()

p2 <- ggplot(sw, aes(x = bio, y = Y)) +
  geom_boxplot(aes(colour = bio)) +
  labs(x = "Bioregional Classification",
       y = "Sørensen Dissimilarity") +
  theme_grey() +
  theme(legend.position = "none")

ggarrange(p1, p2, ncol = 2, labels = "AUTO")

```

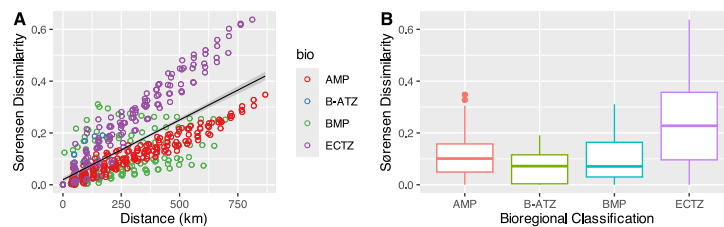


FIGURE 14.8. Main effects of (A) geographic distance along the coast and (B) bioregional classification on mean Sørensen dissimilarity. Points in panel A are coloured by bioregion to show how the distance–dissimilarity relationship maps onto regional structure.

Figure 14.8 shows that Y increases with distance and that the average level of Y differs visibly among bioregions. The additive model combines both patterns in a single regression without yet allowing the distance slope to vary by region.

14.8.5 Fit and Assess Nested Models

```

mod2a <- lm(Y ~ dist, data = sw)
mod2 <- lm(Y ~ dist + bio, data = sw)

anova(mod2a, mod2, test = "F")

R> Analysis of Variance Table
R>
R> Model 1: Y ~ dist
R> Model 2: Y ~ dist + bio
R>   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
R> 1      968 7.7388
R> 2      965 4.1156    3    3.6232 283.19 < 2.2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(mod2)

R>
R> Call:
R> lm(formula = Y ~ dist + bio, data = sw)

```



```

R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.189964 -0.038724 -0.002277  0.029329  0.256021
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept) -2.698e-02  4.547e-03  -5.933 4.15e-09 ***
R> dist         4.612e-04  1.059e-05  43.533 < 2e-16 ***
R> bioB-ATZ     7.340e-02  1.368e-02   5.364 1.02e-07 ***
R> bioBMP       1.258e-02  5.258e-03   2.393  0.0169 *
R> bioECTZ      1.385e-01  5.043e-03  27.461 < 2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.06531 on 965 degrees of freedom
R> Multiple R-squared:  0.7453, Adjusted R-squared:  0.7442
R> F-statistic: 706 on 4 and 965 DF, p-value: < 2.2e-16
anova(mod2)
R> Analysis of Variance Table
R>
R> Response: Y
R>      Df Sum Sq Mean Sq F value    Pr(>F)
R> dist   1  8.4199   8.4199 1974.25 < 2.2e-16 ***
R> bio     3  3.6232   1.2077  283.19 < 2.2e-16 ***
R> Residuals 965  4.1156   0.0043
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The nested comparison uses the sequential ANOVA approach described in [Example 1](#) and asks whether adding `bio` to the distance-only model reduces residual variance by more than chance. The residual sum of squares drops sharply, and the F -test is highly significant ($p < 0.001$), so the more comprehensive additive model fits significantly better.

The `summary()` output shows the partial effect of `dist` and the coefficient for each non-reference bioregion. Each `bio` coefficient is the difference in fitted intercept from AMP after conditioning on distance.

Table 14.3 translates those coefficients into effective intercepts for each bioregion. Read across each row: the 0/1 values determine which $\hat{\gamma}$ coefficient gets added to $\hat{\alpha} = -0.0270$.

TABLE 14.3. Dummy-variable coding, effective intercepts, and estimates for each bioregion in `mod2`, with AMP as the reference level. The full predicted $\hat{Y}$ for any observation adds the effective intercept to $\hat{\beta}_1 \cdot \text{dist}$, which is identical for all bioregions.

Bioregion	bioB-ATZ	bioBMP	bioECTZ	Effective intercept	Estimate
AMP	0	0	0	$\hat{\alpha}$	-0.0270
B-ATZ	1	0	0	$\hat{\alpha} + \hat{\gamma}_1$	-0.0270 + 0.0734 = 0.0464

Bioregion	bioB-ATZ	bioBMP	bioECTZ	Effective intercept	Estimate
BMP	0	1	0	$\hat{\alpha} + \hat{\gamma}_2$	$-0.0270 + 0.0126 = -0.0144$
ECTZ	0	0	1	$\hat{\alpha} + \hat{\gamma}_3$	$-0.0270 + 0.1385 = 0.1115$

The estimate for AMP is $\hat{\alpha}$ read directly from the (Intercept) row of `summary(mod2)`, and its SE is $SE(\hat{\alpha}) = 0.0045$. For any other bioregion the effective intercept is $\hat{\alpha} + \hat{\gamma}_i$, and its SE is **not** simply the sum of the two standard errors reported in `summary()`. Because $\hat{\alpha}$ and $\hat{\gamma}_i$ are estimated from the same data they are correlated, and that correlation must be included:

$$SE(\hat{\alpha} + \hat{\gamma}_i) = \sqrt{\text{Var}(\hat{\alpha}) + \text{Var}(\hat{\gamma}_i) + 2 \text{Cov}(\hat{\alpha}, \hat{\gamma}_i)}$$

The required variances and covariance come from the full variance–covariance matrix of the fitted model, accessible in R as `vcov(mod2)`. For B-ATZ, for example:

```
V <- vcov(mod2)
se_batz <- sqrt(V["(Intercept)", "(Intercept)"] +
               V["bioB-ATZ", "bioB-ATZ"] +
               2 * V["(Intercept)", "bioB-ATZ"])
se_batz
```

```
R> [1] 0.01308128
```

Equivalently, re-levelling the factor to place B-ATZ first and refitting moves its effective intercept into the (Intercept) row, from which the SE is read directly:

```
sw$bio <- relevel(factor(sw$bio), ref = "B-ATZ")
mod2_batz <- lm(Y ~ dist + bio, data = sw)
summary(mod2_batz)$coefficients["(Intercept)", ]
```

```
R>      Estimate Std. Error    t value    Pr(>|t|)
R> 0.0464209451 0.0130812846 3.5486533930 0.0004058626
```

The (Intercept) row of `mod2_batz` now reports the fitted baseline for B-ATZ (0.0464) together with its SE, without any matrix arithmetic. The fitted values, residuals, and overall model fit are unchanged; only the parameterisation of the intercept shifts.

14.8.6 Interpret the Additive Model

The additive model assumes one common slope for `dist` across all bioregions. The fitted distance effect is strongly positive, so as sample sites become more geographically separated along the coast, their seaweed assemblages become more dissimilar.

The bioregion coefficients show that baseline dissimilarity differs among regions even after that distance effect is removed.

The model does not yet address whether the rate at which dissimilarity increases with distance differs among bioregions. If that is the next biological question (it plausibly is, because distinct oceanographic provinces might produce different turnover rates) then the model must be extended to include an interaction between `dist` and `bio`. I will cover that step in [Chapter 15](#).

14.8.7 Where ANCOVA Fits

ANCOVA is often taught as a method distinct from regression and ANOVA. It is not. ANCOVA is a linear model that contains one factor and at least one continuous covariate in an additive structure. The model `mod2` above fits exactly that description since `bio` is the factor and `dist` is the covariate. The coefficient table and ANOVA table produced by `lm()` and `aov()` for an ANCOVA model are algebraically the same.

In R:

```
lm(Y ~ bio + dist, data = sw) # regression-style output
aov(Y ~ bio + dist, data = sw) # ANOVA-style output
```

Both fit the same model. The difference is in the emphasis of the output: `lm()` emphasises the coefficient table and makes the covariate slope explicit; `aov()` draws attention to the ANOVA table and omnibus F -tests. I prefer teaching the `lm()` approach over an ANCOVA.

The standard ANCOVA assumption is that the covariate slope is the **same across factor levels**, *i.e.*, the same `dist` slope applies in all bioregions. In model terms, that is the additive structure without a `bio:dist` interaction. If there is biological reason to expect that different bioregions show different turnover rates with distance, the interaction model becomes appropriate, and the simple ANCOVA framing no longer applies.

14.8.8 Report

WRITE-UP

Methods

Nested linear models were used to test whether mean Sørensen dissimilarity was related to geographic distance and whether mean dissimilarity differed among bioregions after accounting for distance. A distance-only model was compared with an additive model containing distance and bioregional classification using a sequential F -test.

Results

Mean Sørensen dissimilarity increased with geographic distance along the South African coast, and an additive model showed that dissimilarity also differed among bioregions independently. Adding bioregion to the distance-only model improved fit substantially ($p < 0.001$; Figure 14.8). The fitted distance effect was strongly positive, and the bioregion coefficients indicated that baseline dissimilarity differed among regions relative to the AMP reference level.

Discussion

Compositional turnover is jointly structured by geographic distance and bioregional context. The additive model attributes both effects to separate, non-interacting sources of variation. The assumption that distance has the same effect in all bioregions remains to be tested; that question is addressed in Chapter 15.

14.9 EXAMPLE 3: THE FULL ADDITIVE MODEL

I now combine the selected climate predictors from Example 1 with bio as an additive predictor, using the full dataset rather than the ECTZ subset alone. This produces a model in which climatic effects are estimated jointly with bioregional differences.

14.9.1 Fit the Model

```
sw_sub2 <- sw %>%
  dplyr::select(Y, augMean, febSD, augSD, bio)

full_mod3 <- lm(Y ~ augMean + febSD + augSD + bio, data = sw_sub2)
full_mod3a <- lm(Y ~ augMean + febSD + augSD, data = sw_sub2)
null_mod3 <- lm(Y ~ 1, data = sw_sub2)

anova(full_mod3, full_mod3a)

R> Analysis of Variance Table
R>
R> Model 1: Y ~ augMean + febSD + augSD + bio
R> Model 2: Y ~ augMean + febSD + augSD
R>   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
R> 1      963 4.834
R> 2      966 5.689 -3   -0.85507 56.781 < 2.2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC(full_mod3, full_mod3a)

R>           df      AIC
R> full_mod3    8 -2373.840
R> full_mod3a    5 -2221.852

summary(full_mod3)

R>
R> Call:
R> lm(formula = Y ~ augMean + febSD + augSD + bio, data = sw_sub2)
R>
R> Residuals:
```

```

R>      Min      1Q      Median      3Q      Max
R> -0.167665 -0.045280 -0.006433  0.039474  0.288426
R>
R> Coefficients:
R>      Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.014886   0.014271   1.043   0.297
R> augMean      0.281357   0.007892  35.653 < 2e-16 ***
R> febSD       -0.002945   0.002864  -1.028   0.304
R> augSD        0.014463   0.002918   4.956 8.49e-07 ***
R> bioAMP       0.012508   0.014915   0.839   0.402
R> bioBMP      -0.022034   0.014902  -1.479   0.140
R> bioECTZ      0.062590   0.015070   4.153 3.57e-05 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.07085 on 963 degrees of freedom
R> Multiple R-squared:  0.7008, Adjusted R-squared:  0.699
R> F-statistic: 376 on 6 and 963 DF, p-value: < 2.2e-16

```

```
anova(full_mod3)
```

```

R> Analysis of Variance Table
R>
R> Response: Y
R>      Df Sum Sq Mean Sq F value    Pr(>F)
R> augMean  1  9.9900   9.9900 1990.165 < 2.2e-16 ***
R> febSD    1  0.0530   0.0530  10.563  0.001194 **
R> augSD    1  0.4266   0.4266  84.983 < 2.2e-16 ***
R> bio      3  0.8551   0.2850  56.781 < 2.2e-16 ***
R> Residuals 963 4.8340   0.0050
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The climate-plus-bioregion model fits substantially better than the climate-only model: the AIC is lower and the nested comparison is highly significant. Adding bio captures broad regional differences in baseline dissimilarity that the climate predictors alone do not account for. This is ecologically plausible because the South African coast traverses distinct biogeographic provinces shaped by contrasting oceanographic features (the warm Agulhas current to the south-east and the cold Benguela upwelling to the west) that create regional biotic signatures beyond what the three retained temperature metrics capture.

14.9.2 Assess Assumptions

Before interpreting the coefficients, the residual assumptions need to be checked on the fitted model. The standard four-panel diagnostics and component-plus-residual plots for the continuous predictors are shown below.

```

autoplot(full_mod3, shape = 21, colour = "dodgerblue4",
          smooth.colour = "magenta") +
  theme_grey()

```

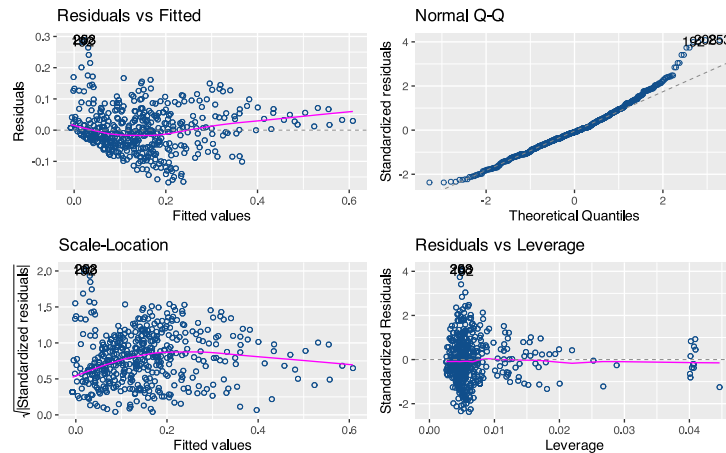


FIGURE 14.9. Standard diagnostic plots for `full_mod3`. Top left: residuals vs fitted values; top right: normal Q-Q plot of residuals; bottom left: scale-location; bottom right: residuals vs leverage with Cook's distance contours.

Reading Figure 14.9:

- **Residuals vs fitted:** In the residuals-versus-fitted panel, the residuals are centred near zero, but the fitted values fall into visible clusters because the model includes a categorical predictor. This is expected from the additive structure with bioregion-specific intercept shifts, so the main question is whether the residual spread changes systematically across the fitted range or whether a clear curved pattern appears. The plot does not indicate a strong departure from linearity, but it also does not support a stronger claim than that.
- **Normal Q-Q:** The normal Q-Q plot should be interpreted in the same cautious way. Most residuals appear to follow the reference line reasonably closely, but some deviations are concentrated in the tails. However, this type of pattern is common in ecological data and does not, by itself, invalidate the model. This suggests that normality is only approximate rather than exact.
- **Scale-location:** The scale-location panel is the main check on constant variance. Here the spread of the residuals appears broadly similar across fitted values, although minor changes in spread may remain. I therefore interpret the variance assumption as acceptable for ordinary least squares, without claiming perfectly constant variance.
- **Residuals vs leverage:** The residuals-versus-leverage panel does not suggest a single observation dominating the fit. Any points with moderately elevated leverage should still be checked individually, but the figure does not indicate

an obvious case in which one or a few observations determine the estimated coefficients.

```
crPlots(full_mod3, col = "dodgerblue4", col.lines = "magenta")
```

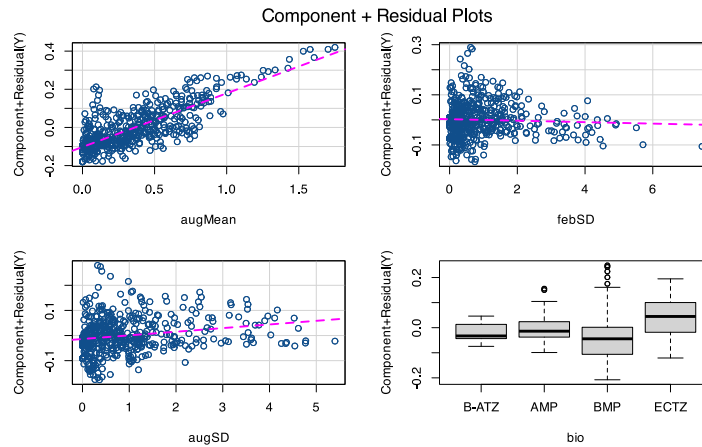


FIGURE 14.10. Component-plus-residual plots for the three continuous predictors in `full_mod3`. A straight fitted line confirms that a linear term is adequate for that predictor after accounting for the others.

The component-plus-residual plots in Figure 14.10 are most useful for assessing whether the continuous predictors require nonlinear terms. These plots do not show strong, systematic curvature, so the linear terms for `augMean`, `febSD`, and `augSD` appear adequate at the present level of resolution. Overall, I'd say the diagnostics support treating the additive linear model as a reasonable summary of the data, while recognising that model adequacy in regression is a matter of degree rather than a binary decision.

Figure 14.11 summarises the two main findings of `full_mod3`: the dominant positive relationship between August mean temperature and dissimilarity, expressed as parallel fitted lines that share a single slope across all bioregions, and the bioregion-specific intercept shifts relative to AMP.

```
bio_levels <- c("AMP", "B-ATZ", "BMP", "ECTZ")

aug_seq <- seq(min(sw_sub2$augMean), max(sw_sub2$augMean), length.out = 200)
pred_partial <- expand.grid(
  augMean = aug_seq,
  febSD   = mean(sw_sub2$febSD),
  augSD   = mean(sw_sub2$augSD),
  bio     = bio_levels
)
```

```

pred_partial$bio <- factor(pred_partial$bio, levels = bio_levels)
pred_partial$fit <- predict(full_mod3, newdata = pred_partial)

p_a <- ggplot() +
  geom_point(data = sw_sub2 > mutate(bio = factor(bio, levels = bio_levels)),
    aes(x = augMean, y = Y, colour = bio),
    shape = 1, alpha = 0.35, size = 1.2) +
  geom_line(data = pred_partial,
    aes(x = augMean, y = fit, colour = bio),
    linewidth = 0.9) +
  scale_colour_brewer(palette = "Set1", name = "Bioregion",
    breaks = bio_levels) +
  labs(x = "August mean temperature (°C)",
    y = "Mean Sørensen dissimilarity") +
  theme_grey()

coef_df <- broom::tidy(full_mod3, conf.int = TRUE) >
  dplyr::filter(stringr::str_starts(term, "bio")) >
  dplyr::mutate(
    bio = factor(stringr::str_remove(term, "bio"),
      levels = rev(bio_levels[-1]))
  )

p_b <- ggplot(coef_df, aes(x = estimate, y = bio, colour = bio)) +
  geom_vline(xintercept = 0, linetype = "dashed", colour = "grey50") +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.15) +
  geom_point(size = 2.5) +
  scale_colour_brewer(palette = "Set1", breaks = bio_levels, drop = FALSE) +
  scale_y_discrete(limits = rev(bio_levels[-1]), drop = FALSE) +
  labs(x = expression(hat(gamma)[i] ~ "(relative to AMP)",
    y = NULL) +
  theme_grey() +
  theme(legend.position = "none")

ggarrange(p_a, p_b, ncol = 2, labels = "AUTO",
  widths = c(1.6, 1), common.legend = TRUE, legend = "right")

```

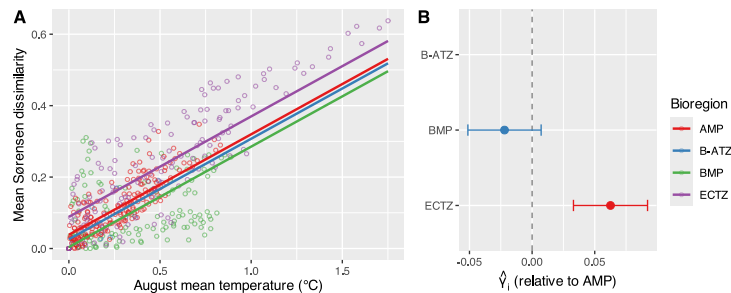



FIGURE 14.11. Results of the additive climate-plus-bioregion model. (A) Observed mean Sørensen dissimilarity plotted against August mean temperature, coloured by bioregion. Parallel lines show the partial effect of `augMean` with `febSD` and `augSD` held at their means; the shared slope reflects the additive structure of `full_mod3`. (B) Bioregion coefficients ($\hat{\gamma}_i$) with 95% confidence intervals, expressed as differences in baseline dissimilarity relative to the AMP reference level. A point to the right of the dashed line indicates higher mean dissimilarity than AMP at any given set of climate values.

14.9.3 Interpret the Model

`augMean` remains the dominant climate predictor. The `bio` coefficients show that after adjusting for the three climate variables, the baseline dissimilarity still differs among regions. The model is additive, so the climate slopes are assumed to be the same in all bioregions.

At this stage my analysis has reached the limit of the additive structure. The next biological question is whether climate relationships themselves differ among bioregions, *i.e.*, whether the slope of `augMean` is steeper in the ECTZ than in the BMP, for example. That question requires interactions, which are the subject of [Chapter 15](#).

DO IT NOW!

Compare the climate-only and climate-plus-bioregion models using both AIC and a nested *F*-test, then interpret what the numbers mean biologically.

```
sw_sub2 <- sw > dplyr::select(Y, augMean, febSD, augSD, bio)
mod_climate_only <- lm(Y ~ augMean + febSD + augSD, data = sw_sub2)
mod_climate_bio <- lm(Y ~ augMean + febSD + augSD + bio, data = sw_sub2)

AIC(mod_climate_only, mod_climate_bio)
anova(mod_climate_only, mod_climate_bio, test = "F")
```

Discuss:

1. By how many AIC units does the climate-plus-bioregion model improve over the climate-only model? Is that a meaningful difference (rule of thumb: $\Delta\text{AIC} > 2$ is noteworthy)?
2. What does a significant nested F -test tell you that AIC alone does not?
3. In plain biological language, what does it mean that adding `bio` to the model improves fit — even after the three climate predictors are already included?

14.9.4 Report

WRITE-UP

Methods

A comprehensive additive multiple regression was fitted to evaluate whether mean Sørensen dissimilarity was explained jointly by August mean temperature, February temperature standard deviation, August temperature standard deviation, and bioregional classification. A climate-plus-bioregion model was compared with a climate-only model using AIC and a nested F -test.

Results

The climate-plus-bioregion additive model fitted the data substantially better than the corresponding climate-only model (lower AIC; nested model comparison $p < 0.001$). August mean temperature remained the dominant positive predictor (Figure 14.11 A). After accounting for climate, baseline dissimilarity still differed among bioregions: ECTZ showed the highest dissimilarity relative to AMP and BMP the lowest, as shown by the bioregion coefficient estimates and their confidence intervals (Figure 14.11 B). The overall model explained approximately 70% of the variation in mean Sørensen dissimilarity.

Discussion

Compositional turnover along the South African coast is shaped both by the selected climate variables and by broader biogeographic structure. The additive model assumes that the climate slopes are uniform across bioregions. Whether that assumption is defensible — or whether the Agulhas and Benguela systems show distinct rates of climate-driven turnover — is a question for the interaction models developed in Chapter 15.

14.10 ALTERNATIVE CATEGORICAL VARIABLE CODING SCHEMES (CON-TRASTS)

The examples above use the default treatment coding in R, which takes the first factor level as the reference and expresses all other levels as deviations from it. Two other coding schemes are occasionally useful.

Effect coding (sum-to-zero coding) expresses each level's coefficient as a deviation from the grand mean rather than from one reference level. The last level is the one dropped, and its effect is the negative sum of the other coefficients.

Helmert coding compares each level to the average of all subsequent levels and can be informative when factor levels have a natural ordering.

The choice of coding does not affect the fitted values or residuals. It only changes how the coefficients are expressed and therefore how they are interpreted. The practical advice is to use the default treatment coding unless there is a specific scientific reason to express effects differently.

```
set.seed(123)
n <- 100
x <- rnorm(n)
cat_var <- factor(sample(c("A", "B", "C", "D"), n, replace = TRUE))
y <- 2 * x + ifelse(cat_var == "A", 3,
                    ifelse(cat_var == "B", 1,
                           ifelse(cat_var == "C", -1, -3))) + rnorm(n)

data_contr <- data.frame(y, x, cat_var)
head(data_contr)
```

```
R>      y      x cat_var
R> 1 0.6667876 -0.56047565      B
R> 2 1.3086873 -0.23017749      B
R> 3 0.4496192 1.55870831      D
R> 4 2.1326402 0.07050839      A
R> 5 -2.8608771 0.12928774      D
R> 6 0.1497346 1.71506499      D
```

Default treatment coding:

```
contrasts(data_contr$cat_var)
```

```
R>   B C D
R> A 0 0 0
R> B 1 0 0
R> C 0 1 0
R> D 0 0 1
```

Change the reference level:

```
data_contr$cat_var <- relevel(data_contr$cat_var, ref = "C")
contrasts(data_contr$cat_var)
```

```
R>   A B D
R> C 0 0 0
R> A 1 0 0
R> B 0 1 0
R> D 0 0 1
```

Effect coding:

```
contrasts(data_contr$cat_var) <- contr.sum(4)
contrasts(data_contr$cat_var)
```

```
R>   [,1] [,2] [,3]
R> C    1    0    0
R> A    0    1    0
R> B    0    0    1
R> D   -1   -1   -1
```

The key lesson is not that these coding schemes must be memorised. It is that the interpretation of categorical coefficients in a regression output is inseparable from the coding that produced them.

14.11 TOWARDS INTERACTION MODELS

The three examples in this chapter all use additive structure, in other words, the effect of each predictor is assumed to be independent of the others. In the context of the seaweed model, this means the climate slope does not depend on which bioregion an observation comes from, and the bioregion difference does not depend on climate.

That additivity assumption is an explicit claim about what structures seaweed diversity along South Africa. For the seaweed data, it may not hold. In fact, Smit et al. (2017) have fitted this model (a variation thereof) and concluded that region-specific slopes are indeed needed. If the slopes differ, the additive model will average across a real biological heterogeneity and produce a fitted `augMean` coefficient that is an average of two different relationships, which would be misleading.

In [Chapter 15](#), I relax the additivity assumption and fit models with explicit interactions. The same seaweed dataset is used there, so the biological context carries directly over. The step from $Y \sim \text{augMean} + \text{bio}$ to $Y \sim \text{augMean} * \text{bio}$ is small in code but large in interpretation, since it shifts from asking “what is the average climate effect?” to asking “does the climate effect differ among bioregions, and if so, how?”

14.12 SUMMARY

Multiple regression models a continuous response as a linear function of several predictors simultaneously. The main interpretive feature is the partial coefficient, that is, each $\hat{\beta}_j$ gives the expected change in Y for a one-unit increase in X_j while all other predictors are held at their observed values. This conditional interpretation is what distinguishes a coefficient from a multiple regression from the slope in a simple regression.

Model specification, such deciding which variables belong and what structure the model should have, is an act of scientific reasoning, not a data-processing step. Multicollinearity, confounding, and measurement error each undermine the reliability of fitted coefficients in different ways; the variance inflation factor detects collinearity but does not resolve confounding. This requires thinking carefully about which variables are absent from the model, so your expert knowledge of biological theory will rescue you at this point.

15

Interaction Effects

When the Effect of One Predictor Depends on Another

IN THIS CHAPTER

- what an interaction term means mathematically and biologically;
- how to interpret the slope of one predictor as a function of another;
- when an interaction is justified and when it is premature;
- why main effects become conditional once an interaction is present;
- how to interpret interaction models using worked algebra and plots;
- how centring continuous predictors clarifies coefficient interpretation;
- how to compare nested models with and without an interaction;
- how effect size and uncertainty matter alongside statistical significance.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Many ecological and biological processes are interactive. Nutrient supply may matter more at high temperature than at low temperature. The influence of distance may differ among bioregions. A treatment may affect males and females differently. These are situations where the effect of one variable is contingent on the state of another.

Linear regression handles these cases by including an **interaction term**. The idea is that one slope depends on another variable. Building this into a model requires care, because once an interaction is present, the main effects can no longer

be read in isolation and the coefficient table becomes harder to interpret without accompanying plots.

In [Chapter 14](#) I developed the idea of additive models, and in this chapter we build upon that foundation. In that chapter, I fitted models of the form $y \sim \text{group} + x$, sometimes called ANCOVA, where the fitted lines for each group were constrained to be parallel. Here I ask what happens when parallel lines are not enough, such as when the slope of x may differ among groups, and when the biology demands it.

One clarification before we start. The term *effect modification* is sometimes used in epidemiology for exactly the situation when an interaction is a claim that one variable *modifies* the effect of another. Stating the effect like this is useful, but it might suggest that some causal implication is implicit in the biology, and that is not something that regression alone can establish. This would require clever experimental manipulation. In the work below, I imply that the models describe conditional associations, and whether those associations reflect true causal modification depends on the study design and domain knowledge, not on model output.

15.1 IMPORTANT CONCEPTS

- **An interaction** means the effect of one predictor depends on another predictor.
- **Main effects become conditional** once an interaction term is present.
- **Plots now become indispensable in model assessment** because coefficients alone rarely communicate interactions clearly.
- **Nested model comparisons** are useful for asking whether the interaction improves the model beyond the main effects alone.
- **Interactions should be biologically motivated**, not generated mechanically because software allows it.

TWO TERMS WORTH CLARIFYING

Conditional means that the value of one thing depends on the value of another. In an interaction model, the effect of X_1 is conditional on X_2 : it is not a single fixed number but changes as X_2 changes. A main effect is likewise conditional because it describes the effect of a predictor at a specific value of the other predictor (usually zero or the reference level), not across all values simultaneously.

Nested models are a sequence of models where each simpler model is a special case of the next. A distance-only model is nested within a distance + bioregion model, which is itself nested within a distance $\times$ bioregion model. Because the simpler model can be recovered from the more complex one by setting extra coefficients to zero, the two can be compared directly using a formal test such as the likelihood ratio test or the F -test in `anova()`.

DO IT NOW!

Before reading further, look at the five bullet points in Important Concepts above and discuss with a partner which of the following biological scenarios are likely to involve an interaction, and which are probably additive.

1. Fish growth increases with food availability, and fish growth also increases with water temperature. There is no reason to think these effects depend on each other.
2. A herbicide reduces plant biomass more strongly in drought conditions than in well-watered plots.
3. The number of species on an island increases with island area, and also increases with proximity to the mainland.
4. The survival of juvenile fish depends on predator density, but that relationship is stronger in degraded habitats than in intact ones.

For each, write a brief justification (one sentence) and an R formula showing how you would code the model, *i.e.*, additive (+) or with interaction (*).

15.2 THE CORE EQUATION

For two predictors, a linear model with interaction is:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i \quad (15.1)$$

The coefficient β_3 multiplies the product $X_1 X_2$. When $\beta_3 = 0$, the model collapses to the additive form of Chapter 14. When $\beta_3 \neq 0$, the effect of X_1 on Y is not constant but it is conditional on X_2 .

In R, an interaction is specified with *:

```
lm(response ~ x1 * x2, data = df)
```

This expands to:

```
lm(response ~ x1 + x2 + x1:x2, data = df)
```

The : term is the interaction itself, while * includes both main effects and the interaction.

The important consequence of Equation 15.1 is that the slope of X_1 (which is the rate at which Y changes as X_1 increases) is not a single number. It is:

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2$$

This is what an interaction means in practice. Every unit increase in X_2 changes the slope of X_1 by exactly β_3 . If $\beta_3 > 0$, the effect of X_1 strengthens as X_2 increases but if $\beta_3 < 0$, the effect of X_1 weakens. If $\beta_3 = 0$, the slope of X_1 is β_1 everywhere, regardless of X_2 and the model reduces to the additive case.

This also implies that:

$$\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_3 X_1$$

The interaction coefficient β_3 simultaneously describes how X_2 modifies the slope of X_1 and how X_1 modifies the slope of X_2 . This is why you cannot interpret an interaction by looking at one predictor in isolation.

15.3 WHEN IS AN INTERACTION APPROPRIATE?

Interactions should follow from biology and should not simply be included in the model because the software allows it. So, before fitting a model with an interaction term, answer these questions.

Does the mechanism support effect modification? The strongest justification for an interaction is a biological or ecological mechanism that predicts it. Photosynthetic rate depends on both light and CO_2 , and the relationship between rate and light is steeper when CO_2 is not limiting. Here, that is a mechanistic prediction of an interaction. Without some supporting reasoning, an interaction term is just speculation.

Is there visual evidence of non-parallel fitted lines? Plotting the response against one predictor, coloured or faceted by the second, may reveal whether the slopes differ. Parallel lines support an additive model but diverging or converging lines suggest an interaction. This assessment is not a formal step, but it is a necessary to gain insight into the presence of effect modification.

Does the scale of measurement matter? Interactions can appear or disappear under transformation. A relationship that appears multiplicative on an arithmetic scale often becomes additive on a log scale. Before concluding that an interaction is real, consider whether it reflects the measurement scale rather than the biology. If a log transformation linearises both relationships and eliminates the interaction, the additive log-scale model may be the more parsimonious description.

Is the dataset large enough to support the interaction? Every interaction term consumes degrees of freedom. A continuous $\times$ continuous interaction adds one parameter. A factor with k levels interacting with a continuous predictor adds $k - 1$ parameters. In small datasets, interaction terms overfit readily and the estimated slopes for individual groups become unreliable. At minimum, ensure adequate observations across the full joint range of both predictors, not just in aggregate, but within each combination. If data cluster in one corner of the predictor space, the interaction estimate extrapolates everywhere else.

DEGREES OF FREEDOM

A degree of freedom is consumed each time the model estimates a parameter from the data. A dataset with n observations and p estimated parameters (intercept plus all slopes and interaction terms) has $n - p$ residual degrees of freedom. This is the quantity left over to estimate the error variance and run hypothesis tests. Fewer residual degrees of freedom means less precise estimates and less powerful tests.

In practice, count the parameters your model requires, subtract from n , and ask whether what remains is adequate. As a rough guide, aim for at least 10–20 observations per parameter. An interaction between a four-level factor and a continuous predictor adds three parameters; if your dataset has 40 rows, that interaction alone halves the residual degrees of freedom available for everything else.

Additive models are simpler and more stable. The burden of justification falls on the interaction, not on the additive alternative.

15.4 INTERPRETING CONDITIONAL COEFFICIENTS

When an interaction is present, the main-effect coefficients are conditional on the other predictor being at its reference value. This is, zero for continuous predictors and the reference level for factors (see [dummy-variable coding in Chapter 14](#)). Although the math makes sense to statisticians, it can produce biologically uninterpretable numbers, especially when zero is not a meaningful reference.

For example, consider a model of mussel filtration rate (L h^{-1}) as a function of shell length (mm) and seawater temperature ($^{\circ}\text{C}$), with their interaction included. The main effect of temperature is the slope of temperature on filtration rate when shell length equals zero. A mussel of zero length does not exist. The coefficient is mathematically well-defined but it describes the filtration performance of an animal that isn't there. Centring shell length on its mean (say, 47 mm, a real mussel of average size) shifts the reference to a biologically meaningful baseline, and the temperature coefficient then describes something an actual mussel does. The mechanics of centring (and the related idea of standardising to unit variance) are introduced in Chapter 14; here the motivation is reinforced by the presence of the interaction term.

CENTRING CONTINUOUS PREDICTORS

Rule: centre a continuous predictor when zero is not a meaningful or observable value in your study. After centring, the main effect of X_1 is its slope when X_2 is at its mean, and vice versa. The interaction coefficient and all fitted values are unchanged by centring.

15.4.1 *A Worked Algebra Example*

A simple numeric illustration makes this visible. Suppose we fit a model where group is a two-level factor (A and B, with A as the reference) and x is a continuous predictor. The factor is represented in the model matrix by a **dummy variable** D_i :

$$Y_i = \alpha + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + \epsilon_i$$

where $D_i = 0$ for group A and $D_i = 1$ for group B.

Substituting $D = 0$ gives the fitted line for group A:

$$\hat{Y} = \alpha + \beta_1 x$$

Substituting $D = 1$ gives the fitted line for group B:

$$\hat{Y} = (\alpha + \beta_2) + (\beta_1 + \beta_3) x$$

The slope of x in group A is β_1 . The slope of x in group B is $\beta_1 + \beta_3$. The interaction coefficient β_3 is therefore not the slope in group B, but the *difference in slopes* between groups B and A. If the slopes were the same ($\beta_3 = 0$), the lines would be parallel and the model would reduce to the additive ANCOVA of Chapter 14.

Now put in numbers. Suppose $\alpha = 2$, $\beta_1 = 0.5$, $\beta_2 = 3$, and $\beta_3 = 0.8$. Then:

- Group A slope: 0.5 per unit of x
- Group B slope: $0.5 + 0.8 = 1.3$ per unit of x
- At $x = 10$: Group A predicts $2 + 0.5(10) = 7$; Group B predicts $(2 + 3) + 1.3(10) = 18$

The difference in predicted response between groups A and B at $x = 10$ is 11 units, not 3. The 3 units (β_2) is only the between-group difference at $x = 0$. Once the interaction is present, the gap between groups changes with x , and reading the coefficients without tracing through the algebra leads to the wrong conclusion.

15.4.2 *The Role of the Reference Group*

In a factor $\times$ continuous interaction, the choice of reference level affects which coefficients appear in the table but not the fitted values or the model's predictions.

If the reference group happens to have a near-zero slope, the main effect of x will appear small or non-significant even when most groups show substantial slopes. This is one more reason to plot. Changing the reference level and refitting the model is also informative, for example, if the conclusions about the biology depend on which group was chosen as reference, they are being drawn from the parameterisation rather than from the data.

DO IT NOW!

Write out the full equation for a model with two predictors, `temp` and `bio` (a two-level factor with levels "warm" and "cold"), including the interaction. Use "cold" as the reference level.

1. Write the equation in the form $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 D_i + \beta_3 X_{1i} D_i + \epsilon_i$, replacing symbols with words where helpful.
2. What is the slope of `temp` for observations in the "cold" bioregion?
3. What is the slope of `temp` for observations in the "warm" bioregion? Express it in terms of the model coefficients.
4. If $\beta_3 = 0$, what does the model reduce to?

Verify your reasoning by fitting this model to any small dataset you create:

```
set.seed(42)
df <- tibble(
  temp = runif(60, 5, 20),
  bio = rep(c("cold", "warm"), 30),
  y = 2 + 0.5 * temp + ifelse(bio == "warm", 3 + 0.8 * temp, 0) + rnorm(60)
)
lm(y ~ temp * bio, data = df) > summary()
```

Check that the coefficient of `temp` matches your answer to question 2.

15.5 EXAMPLE 1: TWO CONTINUOUS PREDICTORS WITH AND WITHOUT CENTRING

Interactions do not only arise between a factor and a continuous predictor. They also occur when two continuous predictors occur together.

I model algal growth as a function of temperature and nutrient concentration, expecting that the effect of temperature depends on nutrient supply. Warming may be more influential when nutrients are abundant than when they are scarce. This is a biologically plausible interaction with a mechanistic basis, *i.e.*, we know that photosynthetic response to temperature is partly limited by nutrient availability.

15.5.1 Example Dataset

I simulate some data for this example.

```
set.seed(14)

int_demo <- tibble(
  temp = runif(120, 8, 24),
  nutrient = runif(120, 0.5, 5)
) >
mutate(
  growth = 4 + 0.25 * temp + 1.1 * nutrient + 0.18 * temp * nutrient +
    rnorm(n(), sd = 1.8),
  temp_c = temp - mean(temp),
  nutrient_c = nutrient - mean(nutrient)
)
```

15.5.2 Do an Exploratory Data Analysis (EDA)

Before fitting the interaction model, I inspect the scale of the predictors and the overall response pattern.

```
int_demo >
summarise(
  mean_temp = mean(temp),
  sd_temp = sd(temp),
  mean_nutrient = mean(nutrient),
  sd_nutrient = sd(nutrient),
  mean_growth = mean(growth),
  sd_growth = sd(growth)
)
```

```
R> # A tibble: 1 x 6
R>   mean_temp sd_temp mean_nutrient sd_nutrient mean_growth sd_growth
R>   <dbl>    <dbl>         <dbl>      <dbl>      <dbl>    <dbl>
R> 1    16.1    4.43           2.75      1.31      18.8    6.41
```

```
int_demo >
mutate(
  nutrient_band = cut(
    nutrient,
    breaks = quantile(nutrient, probs = c(0, 1/3, 2/3, 1)),
    include.lowest = TRUE,
    labels = c("Low nutrient", "Medium nutrient", "High nutrient")
  )
) >
ggplot(aes(x = temp, y = growth, colour = nutrient_band)) +
  geom_point(alpha = 0.65, size = 1.5) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 0.8) +
  labs(
    x = "Temperature",
    y = "Growth",
    colour = "Nutrient level"
  ) +
  theme_grey()
```

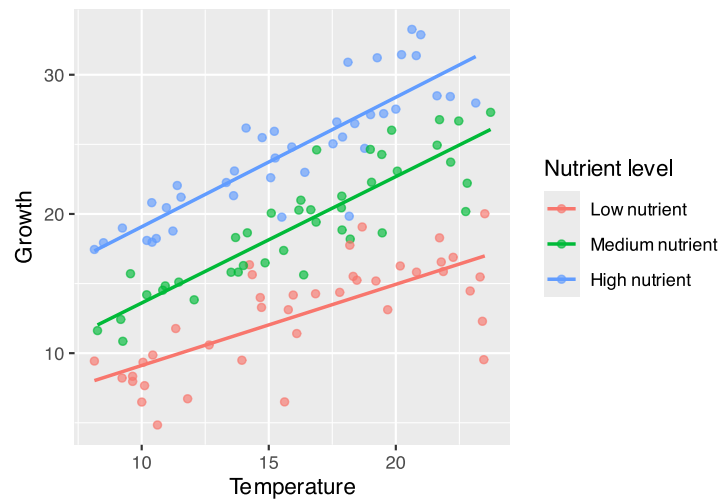


FIGURE 15.1. Algal growth as a function of temperature at low, medium, and high nutrient levels. The fitted lines suggest that the temperature-growth slope steepens as nutrient concentration increases.

The fitted lines in Figure 15.1 are not parallel. As nutrient concentration increases, the temperature-growth slope becomes steeper. This fan shape is the visual sign of an interaction; in other words, the same unit increase in temperature produces a larger growth response when nutrients are abundant than when they are scarce.

DO IT NOW!

Reproduce the fan-shaped plot using `int_demo`, but this time split by temperature bands instead of nutrient bands. Does the same interaction signature appear when the axes are swapped?

```
int_demo >
  mutate(
    temp_band = cut(
      temp,
      breaks = quantile(temp, probs = c(0, 1/3, 2/3, 1)),
      include.lowest = TRUE,
      labels = c("Low temp", "Medium temp", "High temp")
    )
  ) >
  ggplot(aes(x = nutrient, y = growth, colour = temp_band)) +
  geom_point(alpha = 0.65, size = 1.5) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 0.8) +
  labs(x = "Nutrient concentration", y = "Growth", colour = "Temperature level")
```

Answer:

1. Are the fitted lines parallel or diverging? What does that tell you about the interaction?
2. The same interaction is present regardless of which predictor you condition on. Why does that make sense mathematically? (Hint: the interaction term $\beta_3 X_1 X_2$ is symmetric in X_1 and X_2 .)

15.5.3 State the Model Question

The biological question is not whether temperature or nutrient concentration is influential alone, but it is whether or not the effect of temperature *changes across* the nutrient gradient.

For the interaction term, the hypothesis pair is:

$$H_0 : \beta_{\text{temp:nutrient}} = 0$$

$$H_a : \beta_{\text{temp:nutrient}} \neq 0$$

If the interaction coefficient is zero, the slope of temperature does not depend on nutrient concentration, so we can have one sloped line for all nutrient regimes. If it differs from zero, the temperature effect is conditional on nutrient supply and we will have a different line, each with its unique slope, for each nutrient condition.

Stating the question like this might suggest that the appropriate model is `lm(growth ~ temp:nutrient, data = int_demo)`, which specifies the interaction term alone and drops the main effects. That model is almost always wrong. The `:` operator without main effects constrains both β_1 and β_2 to zero; it asserts that temperature has no direct effect on growth and neither does nutrient concentration, only their product matters. That claim is too strong and usually unjustified biologically. The **hierarchy principle** says that whenever an interaction is in the model, its constituent main effects must be in the model too, regardless of whether the primary question concerns them. This is what `*` enforces: `temp * nutrient` expands to `temp + nutrient + temp:nutrient`. The main effects are there not because they are the focal hypothesis but because the interaction term is only meaningful in their presence. Dropping them changes the interpretation of β_3 entirely.

15.5.4 Fit the Uncentred and Centred Models

We fit the same interaction twice: once on raw predictor scales, and once after centring both predictors on their means.

```

mod_raw <- lm(growth ~ temp * nutrient, data = int_demo)
mod_cent <- lm(growth ~ temp_c * nutrient_c, data = int_demo)

tibble(
  term = names(coef(mod_raw)),
  uncentred = round(coef(mod_raw), 3),
  centred = round(coef(mod_cent), 3)
)

```

```

R> # A tibble: 4 x 3
R>   term          uncentred centred
R>   <chr>          <dbl>   <dbl>
R> 1 (Intercept)    3.89    18.9
R> 2 temp           0.23     0.8
R> 3 nutrient       0.76    4.09
R> 4 temp:nutrient  0.207   0.207

```

15.5.5 Interpret the Results

The interaction coefficient is identical in both models; so, centring does not change the interaction estimate. What centring changes are the main effects, and that is the important point.

In the uncentred model, the temperature coefficient is the slope of temperature when nutrient concentration is zero. Nutrient ranges from 0.5 to 5 in this dataset, and zero is outside the data range entirely. That coefficient describes an extrapolated, biologically unobservable condition, and interpreting it directly is misleading. In the centred model, the temperature coefficient is the slope of temperature at the mean nutrient concentration, and the nutrient coefficient is the slope of nutrient at the mean temperature. Both describe something that actually occurs in the data, and centring the data makes this possible.

The fitted values are unchanged because centring is a reparameterisation and not a different model. It repositions the reference point from an arbitrary zero to the sample mean to make the main effects describe something scientifically interpretable rather than a mathematically convenient but biologically meaningless baseline. The same reasoning informs factor $\times$ continuous interactions, ANCOVA-style models, and more complex multiple-regression scenarios. Whenever an interaction is present, ask whether the zero point of each continuous predictor gives you main effects that carry biologically relatable meaning. If it does not, centring is the way to go.

DO IT NOW!

Run both the uncentred and centred models and compare their coefficient tables side by side as shown above. Then answer the following questions:

```
mod_raw <- lm(growth ~ temp * nutrient, data = int_demo)
mod_cent <- lm(growth ~ temp_c * nutrient_c, data = int_demo)

coef(mod_raw)
coef(mod_cent)
```

1. The interaction coefficient (temp:nutrient or temp_c:nutrient_c) should be nearly identical in both models. Is it? Why should it be?
2. In mod_raw, the temperature main effect is the slope of temp when nutrient = 0. Is that a meaningful value for this dataset? Check the range of nutrient in the data.
3. In mod_cent, what does the temperature main effect now represent? Is this more biologically interpretable?
4. Confirm that the fitted values are the same from both models:

```
all.equal(fitted(mod_raw), fitted(mod_cent))
```

15.5.6 Scale Dependence

Centring is not the only scale consideration worth making. Interactions can appear or disappear depending on whether the response is modelled on an arithmetic or logarithmic scale. If a multiplicative process generates the data (for example, growth rates often multiply rather than add) then a log transformation of the response may convert an apparent interaction into an additive relationship. Before concluding that an interaction is biologically real, ask if it reflects the measurement scale. So, a log-scale additive model and an arithmetic-scale interaction model sometimes describe the same underlying biology.

15.6 EXAMPLE 2: COASTAL DISTANCE AND BIOREGIONAL CONTEXT

15.6.1 Example Dataset

I return to the seaweed dataset introduced in Chapter 14, where distance and bioregion were combined in an additive model. The additive model constrained the distance-dissimilarity slope to be the same across all bioregions, differing only in intercept. Here my question is more specific: does the relationship between Sørensen dissimilarity (Y) and coastal distance (dist) differ among bioregions (bio)?

```
sw <- read.csv(here::here("data", "BCB743", "seaweed", "spp_df2.csv"))
sw$bio <- factor(sw$bio)
```


15.6.2 Do an Exploratory Data Analysis (EDA)

We first inspect the response against distance within each bioregion.

```
ggplot(sw, aes(x = dist, y = Y, colour = bio)) +
  geom_point(alpha = 0.6, size = 1.6) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 0.9) +
  labs(x = "Distance along the coast (km)",
       y = "Sørensen dissimilarity",
       colour = "Bioregion")
```

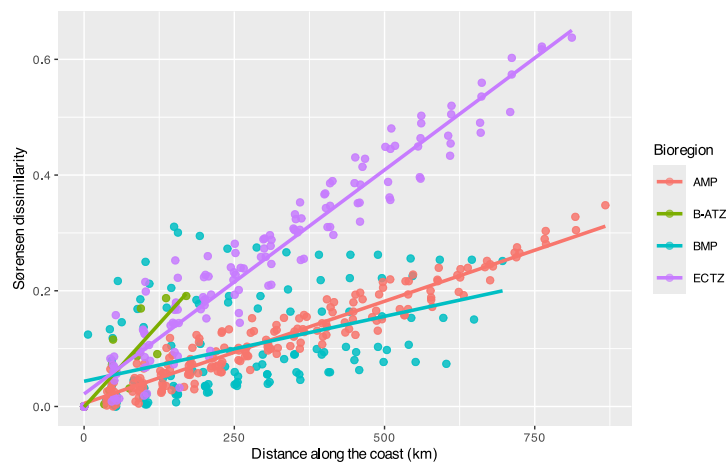


FIGURE 15.2. Sørensen dissimilarity as a function of coastal distance in four bioregions.

The fitted lines in Figure 15.2 are not parallel. The relationship between distance and dissimilarity appears steeper in some bioregions than in others. That visual evidence justifies the interaction model and I now see that the additive constraint I specified in Chapter 14 may be too simple.

15.6.3 State the Hypotheses

Three increasingly specific questions inform this analysis:

1. Is dissimilarity related to distance?
2. Do average dissimilarity levels differ among bioregions?
3. Does the effect of distance differ among bioregions?

The third question is the interaction question. In that case the null hypothesis is:

$$H_0 : \beta_{\text{interaction}} = 0$$

$$H_a : \beta_{\text{interaction}} \neq 0$$

for all interaction terms. Here $\beta_{\text{interaction}}$ refers to the set of coefficients that describe how the slope of distance changes among bioregions. If any of these differ from zero, the distance-dissimilarity slope is not the same in all bioregions.

15.6.4 Fit Nested Models

We fit three nested models:

- a distance-only model;
- a model with the main effects of distance and bioregion;
- a model that adds the interaction between distance and bioregion.

```
mod_dist <- lm(Y ~ dist, data = sw)
mod_main <- lm(Y ~ dist + bio, data = sw)
mod_int <- lm(Y ~ dist * bio, data = sw)

anova(mod_dist, mod_main, mod_int)

R> Analysis of Variance Table
R>
R> Model 1: Y ~ dist
R> Model 2: Y ~ dist + bio
R> Model 3: Y ~ dist * bio
R>   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
R> 1      968 7.7388
R> 2      965 4.1156   3    3.6232 516.21 < 2.2e-16 ***
R> 3      962 2.2507   3    1.8648 265.69 < 2.2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(mod_int)

R>
R> Call:
R> lm(formula = Y ~ dist * bio, data = sw)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.112117 -0.030176 -0.004195  0.023698  0.233520
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  5.341e-03  4.177e-03   1.279   0.2013
R> dist         3.530e-04  1.140e-05  30.958 < 2e-16 ***
R> bioB-ATZ     -6.140e-03  1.659e-02  -0.370   0.7114
R> bioBMP       3.820e-02  6.659e-03   5.737  1.29e-08 ***
R> bioECTZ      1.629e-02  6.447e-03   2.527   0.0117 *
R> dist:bioB-ATZ 7.976e-04  1.875e-04   4.255  2.30e-05 ***
R> dist:bioBMP  -1.285e-04  2.065e-05 -6.222  7.31e-10 ***
R> dist:bioECTZ  4.213e-04  1.801e-05  23.392 < 2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R>
R> Residual standard error: 0.04837 on 962 degrees of freedom
R> Multiple R-squared: 0.8607, Adjusted R-squared: 0.8597
R> F-statistic: 849.2 on 7 and 962 DF, p-value: < 2.2e-16
```

Adding bioregion to the distance-only model improves fit strongly. Adding the interaction improves fit again. The interaction changes the model materially. We can see this in the nested-ANOVA, where the $Y \sim \text{dist} * \text{bio}$ model has the smallest residual sum of squares (RSS), and is highly significant.

15.6.5 Interpret the Interaction

The interaction model $Y \sim \text{dist} * \text{bio}$ expands the additive model with a second set of dummy-coded terms (one per non-reference bioregion) that allow the slope of dist to differ across groups. The full equation is:

$$Y_i = \alpha + \beta_{\text{dist}} \text{dist}_i + \gamma_1 D_{i,\text{B-ATZ}} + \gamma_2 D_{i,\text{BMP}} + \gamma_3 D_{i,\text{ECTZ}} + \delta_1 (\text{dist}_i \cdot D_{i,\text{B-ATZ}}) + \delta_2 (\text{dist}_i \cdot D_{i,\text{BMP}}) + \delta_3 (\text{dist}_i \cdot D_{i,\text{ECTZ}}) + \epsilon_i$$

where each D is an indicator variable (1 for that bioregion, 0 otherwise). Substituting the 0/1 values for each bioregion collapses the equation to a simple regression line, but one with a bioregion-specific intercept **and** a bioregion-specific slope.

Table 15.1 shows which terms are active for each bioregion. For AMP every dummy is zero, so all γ and δ terms drop out entirely. For B-ATZ, $D_{i,\text{B-ATZ}} = 1$ and both bioB-ATZ and $\text{dist}:\text{bioB-ATZ}$ become active; the interaction term contributes $\delta_1 \cdot \text{dist}_i$, which scales with the actual distance value.

TABLE 15.1. Dummy-variable coding for the interaction model $Y \sim \text{dist} * \text{bio}$, with AMP as the reference level. For each non-reference bioregion, exactly one intercept dummy and one slope dummy are switched on.

Bioregion	bioB-ATZ	bioBMP	bioECTZ	dist:bioB-ATZ	dist:bioBMP	dist:bioECTZ
AMP	0	0	0	0	0	0
B-ATZ	1	0	0	dist	0	0
BMP	0	1	0	0	dist	0
ECTZ	0	0	1	0	0	dist

Because each bioregion activates at most one γ and one δ , every group's regression line reduces to a simple intercept-slope pair. Table 15.2 shows the symbolic expressions and Table 15.3 applies the coefficient estimates from `summary(mod_int)`.

TABLE 15.2. Symbolic effective intercepts and slopes for each bioregion in `mod_int`. The interaction coefficients $\hat{\delta}_i$ are slope *adjustments* relative to the reference slope, not bioregion slopes in their own right.

Bioregion	Effective intercept	Effective slope
AMP	$\hat{\alpha}$	$\hat{\beta}_{\text{dist}}$
B-ATZ	$\hat{\alpha} + \hat{\gamma}_1$	$\hat{\beta}_{\text{dist}} + \hat{\delta}_1$
BMP	$\hat{\alpha} + \hat{\gamma}_2$	$\hat{\beta}_{\text{dist}} + \hat{\delta}_2$
ECTZ	$\hat{\alpha} + \hat{\gamma}_3$	$\hat{\beta}_{\text{dist}} + \hat{\delta}_3$

TABLE 15.3. Numerical effective intercepts and slopes for each bioregion, computed from `summary(mod_int)`. Read the `dist` row for $\hat{\beta}_{\text{dist}}$ and the `dist:bioX` rows for the $\hat{\delta}_i$ adjustments, then add. Do not read the interaction rows as slopes — they are deviations from the AMP slope.

Bioregion	Effective intercept	Effective slope (per km)
AMP	$\hat{\alpha} = 0.00534$	$\hat{\beta}_{\text{dist}} = 3.53 \times 10^{-4}$
B-ATZ	$0.00534 + (-0.00614) = -0.00080$	$3.53 \times 10^{-4} + 7.98 \times 10^{-4} = 1.151 \times 10^{-3}$
BMP	$0.00534 + 0.03820 = 0.04354$	$3.53 \times 10^{-4} + (-1.29 \times 10^{-4}) = 2.25 \times 10^{-4}$
ECTZ	$0.00534 + 0.01629 = 0.02163$	$3.53 \times 10^{-4} + 4.21 \times 10^{-4} = 7.74 \times 10^{-4}$

The most important column is the effective slope. B-ATZ shows the steepest gradient (1.15×10^{-3} per km), more than three times the BMP slope (2.25×10^{-4} per km). Over a 500 km transect, B-ATZ accumulates roughly 0.58 units of Sørensen dissimilarity from distance alone, while BMP accumulates only 0.11. That is the biological content of the significant interaction.

The standard errors for the effective slopes follow the same idea presented in [Chapter 14](#). That is, the SE of $\hat{\beta}_{\text{dist}} + \hat{\delta}_i$ requires the covariance between the two estimates, available from `vcov(mod_int)`, or can be read directly by re-levering `bio` to place each target bioregion first and refitting.

The adjusted R^2 of around 0.86 indicates that the model accounts for most of the variation in dissimilarity across the coastline.

The model comparison establishes that the interaction is statistically detectable ($p < 0.001$). The more biologically informative question is how different the slopes actually are. Extract the slope for each bioregion from the coefficient table by adding $\hat{\beta}_{\text{dist}}$ to each interaction coefficient, then compare them. If one bioregion shows a slope of 0.003 per kilometre and another shows 0.008 per kilometre, that two-fold

difference may be substantial relative to the range of Sørensen dissimilarity (0 to 1), or it may be negligible depending on the spatial scale of the study. A significant interaction is not automatically an important one. Evaluate the magnitude of slope differences against what is biologically meaningful, not only against the null of zero difference.

DO IT NOW!

Run the three nested models and use `anova()` to compare them. Then plot the fitted interaction model to see what the coefficients actually imply. Now plot the fitted lines from `mod_int` on top of the raw data:

```
sw$fitted_int <- fitted(mod_int)

ggplot(sw, aes(x = dist, colour = bio)) +
  geom_point(aes(y = Y), alpha = 0.4, size = 1.4) +
  geom_line(aes(y = fitted_int), linewidth = 1) +
  labs(x = "Distance (km)", y = "Sørensen dissimilarity", colour = "Bioregion")
```

Questions:

1. By how much does the residual sum of squares drop when you add the interaction to the main-effects model? Is the improvement statistically significant?
2. From the plot, does the interaction look practically meaningful, or do the fitted lines for different bioregions appear nearly parallel despite the significant test?
3. Looking at the `summary(mod_int)` coefficient table, identify one bioregion whose interaction coefficient is near zero. What does that imply about that bioregion's distance-dissimilarity slope relative to the reference?
4. Extract the fitted slope for each bioregion by combining $\hat{\beta}_{\text{dist}}$ with the relevant interaction coefficient. How large is the biggest slope difference in absolute terms? Is that difference meaningful on the Sørensen scale?

15.6.6 Diagnostic Checks for Interaction Models

Interaction models are more complex than additive ones and create additional opportunities for misfit. Three checks are worth doing alongside any interaction analysis.

Residuals by group. Plot residuals against fitted values separately for each level of the grouping factor. If the residual spread differs substantially among groups, variance heterogeneity is present and may affect inference on the interaction coefficients. This does not invalidate the interaction, but it calls for caution in interpreting standard errors and *p*-values.

Leverage. Observations that are extreme in both predictors simultaneously exert high leverage on interaction estimates. A single unusual observation can drive the entire estimated difference in slopes between groups. Check Cook's distance and hat values, particularly for bioregions or conditions that are sparsely sampled.

Data coverage across predictor combinations. The interaction estimate is most reliable when observations cover the full joint range of both predictors. If one bioregion has few observations at large distances, or if one predictor varies only within a narrow range for certain factor levels, the interaction estimate in those combinations is extrapolation. Plot the joint distribution of predictors before interpreting edge-of-range behaviour.

15.6.7 Reporting

WRITE-UP

Methods

A linear model containing the interaction between geographic distance and bioregion was fitted to test whether the relationship between distance and Sørensen dissimilarity differed among regions. Nested model comparisons were used to evaluate whether the interaction improved fit beyond the additive model.

Results

Sørensen dissimilarity increased with coastal distance, but the slope of that relationship differed among bioregions. A model containing the interaction between distance and bioregion fit the data substantially better than models with distance alone or with only additive main effects ($p < 0.001$ for the interaction comparison; adjusted $R^2 = 0.86$). The slopes ranged from [smallest] to [largest] per kilometre across the four bioregions, indicating that the distance effect is not uniform across the South African coastline.

Discussion

The biologically important point is that distance does not have one universal effect on community turnover. Emphasis in the Discussion should fall on regional differences in turnover rate — their plausible ecological causes, and what they imply for managing distinct coastal zones — not merely on the existence of an interaction term in the model output.

DO IT NOW!

Read through the list of common mistakes that follows. For each one, describe a concrete biological example where a researcher might make that mistake. You do not need to fit any models. This is a thinking exercise.

Common mistake	Concrete example of when this would happen
Adding interactions without biological justification	

Interpreting main effects as if no interaction were present
 Failing to plot the fitted interaction
 Fitting many interaction terms in small datasets
 Treating a significant interaction as important without considering effect size

Then, using the `mod_int` model from Example 2, answer: is the interaction in that model biologically important, or merely statistically detectable? Justify your answer by looking at how much the slopes differ among bioregions in the plot you made above.

15.7 COMMON MISTAKES

- Adding interactions without biological justification — the most common path to overfitting.
- Interpreting main effects as if no interaction were present; in an interaction model, main effects are conditional on the reference value of the other predictor.
- Failing to plot the fitted interaction; the coefficient table communicates the parameterisation but not the biological pattern.
- Fitting many interaction terms simultaneously in small datasets; each term adds parameters, reduces degrees of freedom, and increases the risk that one interaction is significant by chance.
- Treating a statistically detectable interaction as biologically important without asking how large the slope differences actually are relative to meaningful variation in the response.

15.8 SUMMARY

- An interaction means that the effect of one predictor depends on another. Operationally, the slope of X_1 is $\beta_1 + \beta_3 X_2$, not a constant — it changes linearly with X_2 .
- Once an interaction is present, main effects must be interpreted conditionally. They describe the effect of one predictor at the reference value of the other, not its average effect across the data.
- Centring continuous predictors moves the reference to the sample mean and makes main effects interpretable at a scientifically meaningful baseline. The interaction coefficient and fitted values are unchanged.

- Nested model comparisons establish whether the interaction improves model fit, but statistical significance does not determine biological importance. Evaluate slope differences against the biological scale of the response.
- Interactions should be motivated by biological reasoning, checked against visual evidence of non-parallel fitted lines, and validated with attention to scale, sample size, and data coverage across predictor combinations.
- Diagnostic checks — residuals by group, leverage, joint data coverage — matter more in interaction models than in additive ones.

In the next chapter, I remain in the multiple-regression setting but turn to three major threats to interpretation: collinearity, confounding, and measurement error.

16

Collinearity, Confounding, and Measurement Error

Three Threats to Interpretable Regression Models

IN THIS CHAPTER

- what collinearity is, why shared variance destabilises coefficients, and how to diagnose it;
- how confounding differs from collinearity and why it is a problem of attribution rather than overlap;
- why measurement error weakens inference through attenuation bias;
- how proxy variables differ from mechanistic variables and what that implies for interpretation;
- a practical workflow for identifying and responding to each problem.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Regression models are often taught as though the hard part is getting the formula right. In practice, some of the most serious difficulties arise before and after the fitting step. Predictors may overlap so strongly that their individual effects cannot be separated. A third variable may distort a relationship that appears clean. The variables we fit may be noisy measurements, or only rough stand-ins for the biological processes we actually want to understand.

This chapter brings together three related problems:

1. **Collinearity**, where predictors share so much information that coefficients become unstable;
2. **Confounding**, where a missing variable creates or distorts an apparent effect; and
3. **Measurement error**, where noise in the predictor attenuates the estimated effect.

All three produce models that look numerically reasonable but carry weakened or misleading biological interpretations. A model may fit, produce a statistically significant result, and still tell you something false about the mechanism.

The problems are especially common in ecology. Environmental variables are rarely independent because they are generated by the same physical processes: temperature, oxygen, and productivity co-vary in the ocean; altitude, temperature, and radiation co-vary on land. Collinearity and confounding are the normal condition of field data, and ignoring them leads to overconfident conclusions.

In the last section of this chapter, I give you a practical workflow that ties the three problems together. It is worth reading first as an orientation before working through the examples.

Some connections to earlier chapters are worth making. In [Chapter 14](#), the seaweed climate model was trimmed because several candidate predictors were strongly correlated. That decision was an informal response to collinearity. Here, I show you how to go making this decision. In [Chapter 15](#), I noted that interactions can be confused with confounding when a third variable modifies the apparent relationship, and I'll return to that distinction when addressing confounding below.

16.1 IMPORTANT IDEAS

- **Collinearity** occurs when two or more predictor variables in a model are so strongly correlated that the model cannot cleanly separate their individual effects on the response.
- **The instability problem**: when predictors share variance, the model tries to partition the same signal between two parameters. Small changes in the data can produce large swings in those parameters, and that is what unstable coefficients look like in practice.
- **VIF (Variance Inflation Factor)** is the standard diagnostic that quantifies how much the variance of an estimated coefficient has been inflated by overlap with other predictors.
- **Confounding** is about attribution. A confounder influences both the predictor of interest and the response to create or distort an apparent association. Two variables can be weakly correlated and still produce serious confounding.

- **Measurement error** in a predictor attenuates its estimated effect and biases the slope towards zero. Weak effects are not always evidence of weak biology.
- **Proxy variables** are a related interpretive problem. For example, altitude, depth, and distance from shore are easy to measure but they do not represent single mechanisms. Their coefficients describe a spatial or environmental gradient, not a specific causal process.
- Compared to proxy variables, **mechanistic variables** such as temperature, nutrient concentration, light intensity are closer to the process of interest and produce more directly interpretable coefficients, though they too can be collinear.
- **Interpretation vs prediction:** collinearity, confounding, and poor measurement all weaken explanation. A collinear model may still predict well, but its coefficients do not support mechanistic claims. This distinction governs most of the decisions in this chapter and will not be restated for each example.

16.2 PROBLEMS OF NOTE

In the regression chapters so far, I have assumed that each predictor contributes something identifiable to the response, but that assumption is often too optimistic since several things can undermine it.

Two predictors may describe almost the same underlying process and this makes it impossible to say how much each contributes. A variable may appear important only because an unmeasured third variable drives both it and the response, or a predictor may be measured with enough noise that its effect is weakened and unstable. And even a well-measured predictor may be no more than a proxy for something else, a spatial stand-in rather than a mechanism.

These determine whether the biological story that a model tells is defensible.

16.3 R FUNCTIONS

The most useful functions in this chapter are:

- `cor()` and scatterplots for inspecting predictor overlap;
- `car::vif()` for diagnosing collinearity in a fitted model;
- `lm()` for comparing models that omit or include potentially confounding variables;
- `broom::tidy()` for comparing coefficients across models.

16.4 COLLINEARITY

16.4.1 *What Is Collinearity?*

Collinearity occurs when two predictor variables in a model are highly correlated. **Multicollinearity** is the more general term, referring to the situation where one predictor can be substantially predicted from the others.

In a marine environment, temperature and dissolved oxygen are strongly negatively correlated, with colder water holding more oxygen. Including both in a model to predict fish abundance forces the model to assign separate effects to variables that track the same underlying thermal gradient. That is multicollinearity.

Ecology is full of such paired relationships, such as altitude and temperature, nitrate and phosphate, rainfall and river flow. They co-vary because they share physical or biological underpinnings.

Graham's paper, *Confronting multicollinearity in ecological multiple regression* (Graham 2003), is an important reference. It is a methodological review of why collinearity causes trouble in ecological models, how it inflates coefficient uncertainty, and why interpretive confidence can collapse even when the fitted model still predicts reasonably well.

16.4.2 *The Mechanism: Why Shared Variance Destabilises Coefficients*

Collinearity's effects are not intuitive, and to see why it causes unstable coefficients, consider what the model is being asked to do.

When two predictors are highly correlated, they explain much of the same variation in the response. But, the model must nonetheless assign a separate coefficient to each. There is, in principle, a continuous range of coefficient combinations that fit the data nearly equally well. For example, one predictor can be given more credit and the other less, or the credits can be reversed, with only marginal change in the total fit. Which combination the algorithm picks depends on small differences in the data.

Simply removing one observation, adding a small amount of noise, or resampling the data could cause the algorithm to pick a quite different combination. It is this sensitivity that inflated standard errors measure; the coefficient estimates are real but unreliable because the data do not cleanly constrain them.

Sign reversals follow from the same reasoning. If two predictors both have positive effects on the response, the model may compensate by giving one a negative partial coefficient and the other an inflated positive one, while keeping the total prediction nearly unchanged. The coefficient signs can flip because the algorithm is navigating a poorly-constrained estimation problem. This will say nothing at all about the scientific problem being addressed.

In practice, collinearity is primarily a problem for interpretation. In this way, predictions can remain accurate even when individual coefficients are badly estimated.

16.4.3 *Proxy Variables and Mechanistic Variables*

Example 1 introduces this distinction clearly, but the idea belongs in any discussion of collinearity from the start.

A **mechanistic variable** is one that directly participates in the process of interest, such as temperature acting on enzyme kinetics, light driving photosynthesis, nutrient concentration limiting cell division. A **proxy variable** is a stand-in that correlates with the mechanism but does not represent it directly. Examples are altitude standing in for temperature, depth as a proxy for light, or distance from shore as a substitute for wave exposure.

Proxy variables are common in ecology because the mechanistic variable is often expensive or impossible to measure at scale. They are convenient, and people are sometimes lazy. They are not wrong to use, but they create a specific interpretive limit in that the coefficient of a proxy cannot be taken as the effect of the mechanism it represents.

Collinearity between a proxy and the mechanism it stands in for is almost guaranteed. Including both altitude and temperature in a plant-growth model is asking the model to separate the effect of a spatial gradient from the thing that gradient primarily affects. The model usually cannot do this reliably. The analysis of the seaweed biodiversity dataset is to some extent guilty of this (Smit et al. 2017).

16.5 EXAMPLE 1: PREDICTING PLANT GROWTH

Suppose we model plant growth using mean annual temperature and altitude. Both enter the model as predictors, but altitude is not a mechanism; it is simply a spatial index. Plant growth responds to temperature, oxygen partial pressure, UV radiation, and growing season length, all of which correlate with altitude. Temperature is closer to the mechanism, but it too is partly a proxy for the combined effects of the elevational gradient.

Including both predictors asks the model to separate a proxy from one of the processes the proxy represents. The outcome is that coefficients become unstable, standard errors increase, and biological interpretation becomes unclear.

Better modelling choices depend on the question. If the hypothesis is physiological, temperature is the more appropriate predictor. If the question is about broad spatial gradients, altitude may be sufficient alone. But fitting both simultaneously is hard to justify and produces results that are difficult to defend.

16.6 EXAMPLE 2: NUTRIENT LIMITATION IN COASTAL SYSTEMS

Phytoplankton production depends on nutrient supply. If we model it using both nitrate and phosphate, we typically encounter collinearity, because both nutrients are supplied by the same upwelled or remineralised water masses. Both predictors are also mechanistically meaningful, but unlike in Example 1, this is collinearity between two real variables, not between a proxy and a mechanism.

The model may fit well overall, but individual coefficients become unstable and hard to interpret. The model cannot distinguish how much of the phytoplankton response is due to nitrate specifically and how much to phosphate, because both vary together. Both nutrients are in fact required for phytoplankton growth, but which one most constrains production is context specific: it depends on demand relative to supply, so limitation shifts according to which nutrient is in shortest effective supply at that time and place. This is the principle of Liebig's Law of the Minimum, and it is why the apparent response to one nutrient or the other often depends on their balance rather than on either concentration alone.

Possible resolutions depend on the goal. If the question concerns stoichiometric limitation, an N:P ratio may be more informative than either nutrient separately. If the question is about prediction rather than mechanism, PCA or another dimension-reduction approach can represent the broader nutrient gradient in a single axis. If one nutrient is theoretically limiting based on Liebig's law, use that one and leave the other out.

This example shows that collinearity is not always a matter of bad variable choices. In fact, both variables *do* drive phytoplankton growth! So, collinearity can arise between the most sensible predictors imaginable. The decision is how to handle it given the goal. Throughout, it reinforces the notion that doing good science involves reliance on our deep understanding of the governing scientific principles behind our study system, as well of statistical model building.

16.7 DIAGNOSING COLLINEARITY

The standard diagnostic tool is the **Variance Inflation Factor (VIF)**:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (16.1)$$

In Equation 16.1, R_j^2 is the R^2 obtained by regressing predictor X_j on all the remaining predictors in the model. It measures how well X_j can be predicted from its companions; in other words, how much of X_j 's variance is already captured by the other predictors in the model.

When R_j^2 is high, X_j carries very little information that is not already present elsewhere. A R_j^2 of 0.9 means 90% of X_j 's variation is redundant given the other predictors. The VIF for that predictor would be $1/(1 - 0.9) = 10$: the estimated variance of $\hat{\beta}_j$ is ten times what it would be if X_j were uncorrelated with all other predictors. That is the inflation in uncertainty that collinearity imposes.

We rely on these rough thresholds when deciding which variables to discard: a VIF below about 5 is generally manageable; above 5 to 10, interpretation of individual coefficients becomes difficult and should be flagged. These are guides and not hard rules. Whether to act on a VIF analysis recommendation should be decided based on how strongly the analysis requires individual coefficients for interpretation.

Although VIF is an useful number, it does not diagnose a problem on its own. A high VIF means a coefficient is imprecisely estimated due to overlap. Whether that is something you need to be concerned with depends on whether the model is intended for explanation or prediction.

COLLINEARITY, SAMPLE SIZE, AND MODEL COMPLEXITY

Collinearity is more damaging in small samples. With many observations and a moderate correlation between predictors, the model has enough information to partially separate their contributions, and the inflated standard errors remain acceptable. With a small dataset and strongly correlated predictors, the estimation problem becomes very underconstrained. As a rough guide, collinearity that is tolerable with 200 observations can be crippling with 30.

When building models in small datasets, choose your predictors conservatively. The degrees of freedom available to estimate each additional parameter (see the discussion in [Chapter 15](#)) are quickly exhausted, and collinearity makes that situation worse.

16.8 EXAMPLE 3: WORKED COLLINEARITY EXAMPLE

Using the seaweed dataset from Chapter 14, we inspect the candidate climate predictors for overlap.

16.8.1 *Inspect the Predictor Set*

```
sw <- read.csv(here::here("data", "BCB743", "seaweed", "spp_df2.csv"))
sw_ectz <- sw %>%
  filter(bio == "ECTZ") %>%
  select(Y, augMean, febRange, febSD, augSD, annMean)

cor(sw_ectz[, -1])
```

```
R>      augMean  febRange  febSD  augSD  annMean
R> augMean  1.0000000  0.6678245  0.5039666  0.4603614  0.9711458
```

```
R> febRange 0.6678245 1.0000000 0.9212490 0.5415907 0.6860398
R> febSD     0.5039666 0.9212490 1.0000000 0.5744316 0.5533197
R> augSD     0.4603614 0.5415907 0.5744316 1.0000000 0.5851171
R> annMean   0.9711458 0.6860398 0.5533197 0.5851171 1.0000000
```

The correlation matrix already shows that several of the candidate predictors overlap strongly. Visually, this pattern is even more revealing.

```
ggplot(sw_ectz, aes(x = augMean, y = annMean)) +
  geom_point(alpha = 0.7, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = TRUE, colour = "magenta", linewidth = 0.8) +
  labs(
    x = "August mean temperature (°C)",
    y = "Annual mean temperature (°C)"
  ) +
  theme_grey()
```

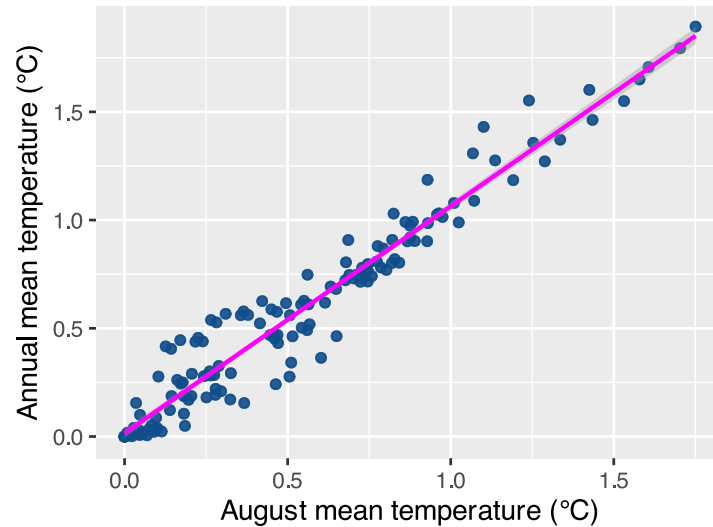


FIGURE 16.1. August mean temperature and annual mean temperature are tightly linearly related in the ECTZ subset. Two variables that fall along a line like this carry largely the same information; including both in a regression forces the model to partition the same signal between two parameters.

Figure 16.1 shows why collinearity is visually intuitive once you look for it. The two temperature variables follow almost the same line. A regression model that includes both must somehow divide this common variation between two separate coefficients, and as the mechanism section above explains, that division is unreliable. Mechanistically, you can also see why this high amount of collinearity exists: this is

because the two predictor variables are not entirely independent, since the August mean temperature is a part of the annual mean signal itself.

16.8.2 Diagnose the Overlap with VIF

```
col_mod <- lm(Y ~ augMean + febRange + febSD + augSD + annMean,
             data = sw_ectz)

vif(col_mod)
```

```
R> augMean febRange febSD augSD annMean
R> 27.947767 10.806635 8.765732 2.497739 31.061900
```

The VIF values confirm what the scatter and correlation matrix suggested. Predictors with VIF values well above 5-10 have R_j^2 values close to 1, so most of their variation is already explained by the other predictors in the model, and therefore their coefficients carry disproportionately large uncertainty.

16.8.3 Compare Simple and Multiple-Regression Slopes

The comparison below is a clear illustration of what collinearity actually does in practice.

```
preds <- c("augMean", "febRange", "febSD", "augSD", "annMean")

simple_models <- purrr::map(preds, ~ lm(as.formula(paste("Y ~", .x)),
                                     data = sw_ectz))

simple_slopes <- purrr::map2_dfr(simple_models, preds, \(mod, nm) {
  tidy(mod) %>
    filter(term != "(Intercept)") %>
    mutate(model = "Simple regression",
           predictor = nm)
})

multiple_slopes <- tidy(col_mod) %>
  filter(term != "(Intercept)") %>
  mutate(model = "Multiple regression",
         predictor = term)

bind_rows(simple_slopes, multiple_slopes) %>
  select(model, predictor, estimate, std.error, statistic, p.value)
```

```
R> # A tibble: 10 x 6
R>   model      predictor estimate std.error statistic    p.value
R>   <chr>      <chr>      <dbl>    <dbl>    <dbl>    <dbl>
R> 1 Simple regression augMean    0.346    0.0109    31.7 6.68e- 96
R> 2 Simple regression febRange    0.182    0.00890   20.4 8.31e- 58
R> 3 Simple regression febSD      0.172    0.0124    13.8 1.56e- 33
R> 4 Simple regression augSD      0.0879   0.00720    12.2 6.68e- 28
R> 5 Simple regression annMean    0.332    0.00867   38.3 6.82e-115
R> 6 Multiple regression augMean   -0.0799   0.0426    -1.87 6.18e- 2
R> 7 Multiple regression febRange    0.113    0.0159     7.08 1.14e- 11
R> 8 Multiple regression febSD     -0.0572   0.0166    -3.45 6.37e- 4
R> 9 Multiple regression augSD      0.00302   0.00489     0.619 5.36e- 1
R> 10 Multiple regression annMean    0.323    0.0416     7.76 1.59e- 13
```

We see that predictors that show strong, stable slopes in simple regressions become weaker, unstable, or sign-reversed once the overlapping predictors are fitted simultaneously.

The model is not necessarily wrong as far as its math is concerned since it is correctly trying to estimate partial effects given the available data. But those partial effects are being estimated under much greater uncertainty than any single slope suggests, because the model cannot cleanly separate what each predictor contributes. And this creates problems for when we need to *explain* the findings in terms of the underlying biological or ecological theory.

16.8.4 Interpret the Collinearity Problem

The difficulty lies in attribution, not in the existence of a fitted model. Several predictors here represent overlapping components of the same underlying climatic gradient. The model cannot assign a stable partial effect to any one of them. This is why collinearity is so often an interpretive problem rather than a purely computational one.

16.8.5 Reporting

WRITE-UP

Methods

Candidate climate predictors were screened for overlap using a correlation matrix, a scatterplot of the most strongly correlated pair, and variance inflation factors. Simple and multiple regressions were then compared to assess how strongly collinearity altered the apparent slopes of individual predictors.

Results

Several climate predictors in the seaweed model were strongly correlated, and the VIF analysis showed substantial overlap among them. Predictors that showed clear positive slopes in isolation became weaker or less stable once the overlapping variables were fitted together, demonstrating that the apparent partial effects were sensitive to collinearity.

Discussion

The model is not necessarily invalid, but coefficient-level interpretation becomes difficult when predictors represent overlapping components of the same underlying gradient. That is the key discussion point for a collinearity analysis.

16.9 ADDRESSING COLLINEARITY

Common responses include:

- removing one of the correlated predictors, choosing based on which better matches the biological question;
- using theory to select the variable closest to the mechanism of interest;
- combining correlated predictors into a composite index or ordination axis when the question concerns the overall gradient rather than any single variable;
- accepting some overlap if prediction is the goal, while avoiding strong causal language about individual coefficients;
- using regularisation methods such as ridge regression or the lasso when predictive stability is the primary objective.

No one VIF threshold automatically resolves the problem. The decision rests on whether the model needs interpretable individual coefficients. Ridge regression, for example, addresses collinearity by shrinking coefficients towards zero; it trades a small increase in bias for a large reduction in variance, making coefficients more stable but no longer unbiased. In [Chapter 25](#), I will demonstrate the regularisation methods.

SCALE AND TRANSFORMATION

Collinearity can change under transformation. Two predictors that are strongly correlated on arithmetic scales may be less correlated after log transformation, particularly when their relationship is multiplicative rather than additive. Before concluding that collinearity is an unavoidable feature of a dataset, consider whether a transformation suggested by the biology might also reduce the overlap between predictors.

16.10 CONFOUNDING

16.10.1 *What Is Confounding?*

Confounding is a different problem from collinearity.

Collinearity is about overlap among predictors that are already in the model. The variables are measured, present, and causing estimation difficulties because they share information.

Confounding is about a variable that is *not* in the model (either it was measured but omitted, or the researcher naively failed to sample it) that influences both the predictor of interest and the response. The confounder creates or distorts the apparent association between the predictor and response, because the model attributes to the predictor an effect that is partly or wholly due to the omitted variable.

A confounder does not need to be strongly correlated with the predictor to cause problems. Even a modest association between the confounder and both variables can shift the estimated coefficient substantially.

16.10.2 *Confounding and Causal Interpretation*

Confounding is fundamentally a causal concept, even though regression is not a causal tool by itself. The concern is whether the apparent effect of a predictor reflects a genuine influence on the response, or whether it reflects the action of an unmeasured third variable.

In field ecology, true causal claims require experimental manipulation or, at minimum, careful design that separates the process of interest from associated variables. Regression with observational data can identify associations and quantify how much they change when additional variables are included. It cannot establish causality on its own. The language to use is “the apparent effect of temperature after accounting for nutrient concentration,” not “the causal effect of temperature.” Whether the remaining partial effect reflects a true temperature mechanism requires domain knowledge and careful study design to assess.

16.11 EXAMPLE 4: WORKED CONFOUNDING EXAMPLE

Suppose we want to know whether seawater temperature affects algal productivity along a coastal gradient, but nutrient concentration also varies with temperature and itself drives productivity. In that situation, a temperature-only model attributes to temperature some of the variation that nutrient concentration actually produces. The temperature coefficient is confounded.

```
set.seed(144)

n <- 120
nutrient <- rnorm(n, mean = 10, sd = 2)
temperature <- 15 + 0.9 * nutrient + rnorm(n, sd = 1.2)
productivity <- 4 + 1.4 * nutrient + 0.15 * temperature + rnorm(n, sd = 1.5)

conf_dat <- tibble(
  nutrient = nutrient,
  temperature = temperature,
  productivity = productivity
)
```

16.11.1 *Do an Exploratory Data Analysis (EDA)*

```
conf_dat > summarise(
```

```

    mean_temperature = mean(temperature),
    sd_temperature = sd(temperature),
    mean_nutrient = mean(nutrient),
    sd_nutrient = sd(nutrient),
    mean_productivity = mean(productivity),
    sd_productivity = sd(productivity)
  )

R> # A tibble: 1 x 6
R>   mean_temperature sd_temperature mean_nutrient sd_nutrient
R>   <dbl>          <dbl>          <dbl>      <dbl>
R> 1      23.7        2.26          9.84       2.04
R> # i 2 more variables: mean_productivity <dbl>, sd_productivity <dbl>

cor(conf_dat)

R>           nutrient temperature productivity
R> nutrient    1.0000000    0.8403428    0.9035191
R> temperature 0.8403428    1.0000000    0.8066436
R> productivity 0.9035191    0.8066436    1.0000000

p1 <- ggplot(conf_dat, aes(x = temperature, y = productivity)) +
  geom_point(alpha = 0.7, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = FALSE, colour = "magenta") +
  labs(x = "Temperature", y = "Algal productivity") +
  theme_grey()

p2 <- ggplot(conf_dat, aes(x = nutrient, y = productivity)) +
  geom_point(alpha = 0.7, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = FALSE, colour = "magenta") +
  labs(x = "Nutrient concentration", y = "Algal productivity") +
  theme_grey()

p3 <- ggplot(conf_dat, aes(x = nutrient, y = temperature)) +
  geom_point(alpha = 0.7, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = FALSE, colour = "magenta") +
  labs(x = "Nutrient concentration", y = "Temperature") +
  theme_grey()

ggarrange(p1, p2, p3, ncol = 3)

```

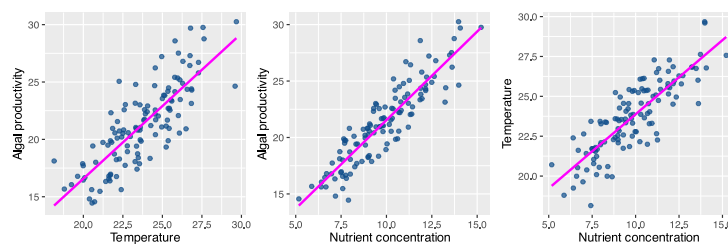


FIGURE 16.2. A confounding situation where nutrient concentration is associated with both temperature and productivity.

The EDA in Figure 16.2 already shows the confounding structure. Temperature and nutrient concentration are strongly positively correlated ($r = 0.84$), and both

are positively related to algal productivity. A temperature-only model will therefore absorb part of the nutrient effect into the temperature coefficient.

16.11.2 State the Model Question and Hypotheses

The biological question is not just whether temperature and productivity are associated. It is whether temperature still explains productivity **after nutrient concentration has been accounted for**.

$$H_0 : \beta_{\text{temperature}} = 0$$

$$H_a : \beta_{\text{temperature}} \neq 0$$

Here $\beta_{\text{temperature}}$ refers to the partial effect of temperature in the model that also includes nutrient concentration. If the temperature coefficient changes strongly when nutrient is added, and that change is evidence of confounding: the original temperature result was absorbing nutrient variation.

16.11.3 Fit Competing Models

```
mod_temp_only <- lm(productivity ~ temperature, data = conf_dat)
mod_with_nutrient <- lm(productivity ~ temperature + nutrient, data = conf_dat)

conf_compare <- bind_rows(
  tidy(mod_temp_only) >
    filter(term != "(Intercept)") >
    mutate(model = "Temperature only"),
  tidy(mod_with_nutrient) >
    filter(term != "(Intercept)") >
    mutate(model = "Temperature + nutrient")
) >
  select(model, term, estimate, std.error, statistic, p.value)

conf_compare
```

```
R> # A tibble: 3 x 6
R>   model      term      estimate std.error statistic  p.value
R>   <chr>      <chr>      <dbl>    <dbl>    <dbl>    <dbl>
R> 1 Temperature only temperature  1.27    0.0854    14.8 1.02e-28
R> 2 Temperature + nutrient temperature  0.253   0.112     2.25 2.61e- 2
R> 3 Temperature + nutrient nutrient     1.33   0.124    10.7 3.99e-19
```

```
anova(mod_temp_only, mod_with_nutrient)
```

```
R> Analysis of Variance Table
R>
R> Model 1: productivity ~ temperature
R> Model 2: productivity ~ temperature + nutrient
R>   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
R> 1      118 522.57
R> 2      117 263.30  1    259.26 115.2 < 2.2e-16 ***
R> ---
```

```
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

16.11.4 *Interpret the Results*

In the temperature-only model, algal productivity appears to increase strongly with temperature ($\beta = 1.27$, $SE = 0.09$, $p < 0.001$). Once nutrient concentration is included, the estimated temperature effect shrinks to about 0.25 ($SE = 0.11$, $p = 0.026$), while nutrient concentration has a strong positive effect ($\beta = 1.33$, $SE = 0.12$, $p < 0.001$).

The temperature-only model was answering the wrong biological question. It estimated the marginal association between temperature and productivity, which includes everything that co-varies with temperature, including nutrient concentration. The partial temperature coefficient in the richer model is the effect of temperature after holding nutrient concentration constant, which is closer to the physiological question.

So, we see that the coefficient did not just shrink; it changed its meaning.

16.11.5 *Practical Implications of Confounding*

Confounding arises in field studies when variables of biological importance are omitted from the model, whether because they were not measured, not recognised, or excluded by an overly simple design. The best defence is domain knowledge: before fitting a model, ask which unmeasured variables could influence both the predictor and the response, and then measure or control for them. Statistical adjustment after the fact, as illustrated here, is only possible when the confounder is measured.

CONFOUNDING, INTERACTION, AND SCALE

Confounding and interaction are sometimes confused. An interaction means the effect of one predictor depends on the value of another. Both predictors are present in the model, and their joint effect is the point (see [Chapter 15](#)). A confounder is a third variable whose omission distorts the estimated effect of the predictor of interest.

Confounding can also change under transformation. If the true relationship between a confounder and the response is multiplicative, a log transformation of the response may reduce the confounder's influence on the temperature coefficient because the scale on which the relationship is additive has changed. This can alter the apparent magnitude of the confounding effect, but doesn't resolve the scientific problem.

16.11.6 *Reporting*

WRITE-UP**Methods**

A simulated coastal algal-abundance dataset was used to examine confounding by nutrient concentration in a temperature-abundance relationship. A simple linear regression of abundance on temperature alone was compared with a multiple regression that included both temperature and nutrient concentration, and the fitted temperature coefficient was evaluated for change between the two models.

Results

In the temperature-only model, algal abundance appeared to increase strongly with temperature ($\beta = 1.27 \pm 0.09$ SE, $t = 14.83$, $p < 0.001$). However, once nutrient concentration was included, the fitted temperature effect decreased markedly to $\beta = 0.25 \pm 0.11$ SE ($t = 2.25$, $p = 0.026$), while nutrient concentration had a strong positive effect on abundance ($\beta = 1.33 \pm 0.12$ SE, $t = 10.73$, $p < 0.001$). Temperature and nutrient concentration were also strongly correlated ($r = 0.84$), indicating that the simple temperature-abundance relationship was substantially confounded by nutrient availability.

Discussion

The temperature-only model estimated the marginal temperature-abundance association, which conflated temperature effects with nutrient effects. The partial temperature coefficient in the richer model is a more defensible estimate of the temperature effect given this dataset, though a causal interpretation would require an experimental design that independently manipulates temperature and nutrients.

16.12 MEASUREMENT ERROR

Standard linear regression assumes that predictors are measured without serious error. That assumption is almost never exactly true in the field. Instruments are imperfect, environmental conditions fluctuate between sensor and organism, and the temporal or spatial scale of a measurement may not match the scale of the biological process.

When a predictor is measured with substantial noise, the estimated regression slope is biased towards zero. This is called **attenuation bias**. The model is relating the response to an imperfect version of the underlying variable, and the slope is therefore pulled downward. The effect is real but appears weaker than it is.

Attenuation bias must be considered for three reasons. It weakens the apparent effect of a predictor that may genuinely be important. It reduces statistical power, increasing the risk of false negatives. And it creates an asymmetry because predictors measured with less precision appear less important than those measured more accurately, which can mislead decisions about which variables to include.

16.13 EXAMPLE 5: WORKED MEASUREMENT-ERROR EXAMPLE

Suppose a response truly depends on temperature, but the observed temperature values are noisy because of instrument error or a mismatch between sensor location and the thermal environment experienced by the organism.

16.13.1 *Simulate the True and Noisy Predictors*

```
set.seed(414)

n <- 150
true_temp <- runif(n, min = 8, max = 24)
observed_temp <- true_temp + rnorm(n, sd = 4.5)
growth <- 1.2 + 0.55 * true_temp + rnorm(n, sd = 1.2)

me_dat <- tibble(
  true_temp = true_temp,
  observed_temp = observed_temp,
  growth = growth
)

mod_true <- lm(growth ~ true_temp, data = me_dat)
mod_observed <- lm(growth ~ observed_temp, data = me_dat)

bind_rows(
  tidy(mod_true) > mutate(model = "True predictor"),
  tidy(mod_observed) > mutate(model = "Observed noisy predictor")
) >
  filter(term != "(Intercept)") >
  select(model, term, estimate, std.error, statistic, p.value)
```

```
R> # A tibble: 2 x 6
R>   model      term      estimate std.error statistic  p.value
R>   <chr>      <chr>      <dbl>    <dbl>    <dbl>    <dbl>
R> 1 True predictor true_temp    0.555    0.0210    26.4 5.89e-58
R> 2 Observed noisy predictor observed_~    0.340    0.0293    11.6 1.43e-22
```

The measurement error standard deviation (4.5 °C) is set close to the standard deviation of the true predictor (≈ 4.6 °C), giving a noise-to-signal variance ratio near 1:1. Under classical measurement error, this implies that the fitted slope for the noisy predictor should be substantially attenuated relative to the true slope, with an expected reduction of roughly one-half. The fitted value is consistent with this expectation, though not identical, reflecting ordinary sampling variability in a finite dataset.

16.13.2 *Compare the Fitted Relationships*

```
p3 <- ggplot(me_dat, aes(x = true_temp, y = growth)) +
  geom_point(alpha = 0.7, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = TRUE, colour = "magenta", linewidth = 0.8) +
  scale_x_continuous(limits = c(0, 30)) +
  labs(x = "True temperature (°C)", y = "Growth", title = "True predictor")
```

```
p4 <- ggplot(me_dat, aes(x = observed_temp, y = growth)) +
  geom_point(alpha = 0.7, colour = "dodgerblue4") +
  geom_smooth(method = "lm", se = TRUE, colour = "magenta", linewidth = 0.8) +
  scale_x_continuous(limits = c(0, 30)) +
  labs(x = "Observed temperature (°C)", y = "Growth", title = "Noisy predictor")
ggarrange(p3, p4, ncol = 2, labels = "AUTO")
```

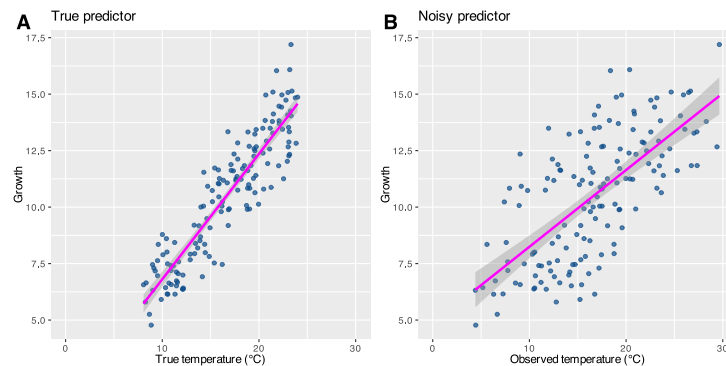


FIGURE 16.3. A) shows growth regressed against the true temperature, with a tight, steeply rising relationship. B) uses the same response but replaces the true predictor with a version with measurement noise of comparable magnitude to the signal itself. The slope is reduced and the cloud is visibly broader in the x-direction. The biology has not changed; only the measurement quality has.

Figure 16.3 makes the effect very obvious. The left panel shows what the relationship looks like when temperature is measured at the organism's scale: we see a clear, steep line. The right panel shows what happens when the sensor is poorly placed, averaged over the wrong spatial scale, or simply imprecise. Now we see the slope is approximately halved and the scatter is much wider. The fitted line in the right panel would lead a you to conclude that temperature is a weaker driver than it actually is. Both panels use the same 150 observations and the same biological process, so the difference is entirely in measurement quality.

16.13.3 Interpret the Measurement-Error Problem

The lesson is that a weak fitted slope is not necessarily evidence of a weak biological relationship. It may instead simply reflect poor measurement.

This is a concern that must be taken seriously when comparing studies. A study that measured temperature precisely at the organism's microhabitat will estimate a

steeper, more significant slope than one that used remote-sensed data or a regional weather station. The difference in results would be attributed to the biology if the measurement quality is ignored, but it may be largely an artefact of data quality or appropriateness.

Practically, you'd want to invest in measurement quality, replicate measurements to estimate and potentially correct for noise, and be explicit about the precision of predictors when reporting results.

16.13.4 *Reporting*

WRITE-UP

Methods

A simulation was used to examine the effect of measurement error in a predictor variable by comparing the fitted relationship obtained from the true predictor with that obtained from a noisy observed version of the same predictor.

Results

Adding measurement error to the predictor reduced the apparent slope of the fitted relationship and weakened the visual strength of the association, illustrating the characteristic attenuation expected when predictors are measured with error.

Discussion

Weak fitted effects do not always imply weak biology. Poor measurement systematically flattens otherwise meaningful relationships. That is the interpretive warning to carry into any Discussion section that relies on observational field data.

16.13.5 *Model II Regression as a Solution to Measurement Error*

Ordinary least squares (OLS) estimates the conditional mean of a response given a predictor under an error structure concentrated in the response. This is appropriate when the predictor is measured with relatively high precision. When the predictor contains substantial measurement error, the fitted slope is attenuated because variability in the predictor is treated as noise rather than signal.

Model II regression provides an alternative when both variables contain measurement error. Two commonly used forms are major axis (MA) regression and standardised major axis (SMA) regression. These methods fit a line that reflects the shared variation between the variables rather than a directional dependence of one on the other.

Model II approaches are useful when the goal is to describe the relationship between two noisy variables and when neither can be treated as measured without

error. They can therefore serve as a practical complement when attenuation bias is a concern.

However, the interpretation differs from OLS. The fitted line describes the structure of association rather than the conditional effect of a predictor on a response. For explanatory models, improving measurement quality or explicitly modelling the error process remains the more direct solution.

16.14 REVISITING PROXY VARIABLES

Proxy variables appear throughout ecology, and by this point in the chapter several examples have already appeared: altitude in the plant-growth model, distance from shore as a proxy for exposure, chlorophyll-*a* as a proxy for productivity, body size as a proxy for condition or age.

Let's integrate the consequences for interpretation.

A proxy is always an imperfect representation of the process of interest. It may correlate well with the mechanism, but it also correlates with other processes along the same gradient. Altitude covaries with temperature, radiation, precipitation, and soil properties. Depth in the ocean covaries with light, pressure, and oxygen. A coefficient estimated for a proxy variable describes the gradient well, but it says little about the driving mechanism.

This has a direct consequence for interpretation, because a significant coefficient for a proxy variable tells you that the gradient matters. Since it does not tell you which component of that gradient is the active process, you still need to measure the mechanism directly.

MECHANISM VERSUS PROXY

If the question is physiological, use a physiological variable where possible. If the question concerns a broad spatial or environmental gradient, a proxy may be perfectly appropriate. The important thing is that the variable should match the scale and aim of the question, and the coefficient should be interpreted at the same scale, not treated as a mechanistic effect.

When a proxy and its mechanistic counterpart are both in the model, collinearity is almost guaranteed. The overlap between altitude and temperature, or between distance and exposure, is a consequence of the proxy relationship. This is why proxy variables belong in the collinearity discussion and not only in a separate section: the interpretive problems they create are inseparable from the instability problems that collinearity creates.

16.15 A PRACTICAL WORKFLOW

When building and interpreting a regression model, the following sets address the problems in this chapter before they become interpretive traps.

1. **State the biological hypothesis first.** Is the question about a mechanism or a pattern? Your answer will determine which variables belong in the model.
2. **Identify which predictors are mechanistic and which are proxies.** Use mechanistic variables when the question is explanatory; proxies are acceptable for description or prediction.
3. **Ask which variables could confound the relationship of interest.** If a plausible confounder was not measured, that is a design limitation that should be stated.
4. **Inspect predictor correlations and calculate VIF before fitting the full model.** Strongly correlated predictors should not both enter the model without justification.
5. **Compare simpler and more fully-specified models when confounding is plausible.** A substantial change in a coefficient when a new variable is added is evidence of confounding, not only of model improvement.
6. **Remove or combine redundant predictors when interpretation is the goal.** If prediction is the goal and collinearity is moderate, some overlap may be tolerable, but then you should avoid strong mechanistic claims about individual coefficients.
7. **Be honest about measurement quality.** If an important predictor was measured with substantial noise, it will limit the strength of the conclusions, and this should be acknowledged.

An additional benefit to implementing these steps is that it also clarifies the explanation–prediction distinction that underpins most decisions we had to make in this chapter. Collinearity, confounding, and measurement error all weaken explanation, and sometimes they will all be present in the same dataset! A collinear model may still predict well, and an attenuated slope may still rank observations correctly, but the problems become serious when the question is biological interpretation rather than forecasting.

16.16 SUMMARY

- Collinearity makes coefficients unstable because predictors share variance. The model cannot cleanly partition the same signal between two parameters, so small data changes produce large coefficient changes, and sign reversals are possible.

- VIF quantifies this instability. A R_j^2 close to 1 means the predictor contributes little unique information and its coefficient variance is correspondingly inflated.
- Proxy variables and mechanistic variables are susceptible to collinearity when both enter the same model. Their interpretive constraints are different, so a proxy coefficient describes a gradient, not the mechanism.
- Confounding is a problem of attribution. An omitted variable that influences both the predictor and the response distorts the apparent effect, and correction requires measuring the confounder and including it in the model.
- Measurement error biases slopes towards zero, so w effects may reflect poor measurement rather than weak biology.
- These problems must be handled, and avoided if they can be avoided, primarily through strong biological reasoning (first), careful variable selection, and honest acknowledgement of what the data cannot establish.

Here, I looked at model construction and in the next chapter model evaluation will be the focus. This will include diagnostics, comparison, and the question of whether a fitted model is actually behaving well.

17

Model Checking and Evaluation

Diagnostics, Comparison, and Generalisation

IN THIS CHAPTER

- the distinction between model checking and model evaluation;
- how residuals are used to assess whether a model is adequate for the data;
- what leverage and influence mean in practice;
- why AIC and cross-validation answer different questions about model quality;
- how model adequacy connects to the interpretive risks discussed in Chapter 16.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Fitting a model is only the middle of the analysis. Then we need to ask:

1. Is the model adequate for the data?
2. How does this model compare with plausible alternatives?

The first question is **model checking**, which is examining whether the model is consistent with its assumptions and whether the fitted structure is appropriate for the data at hand. The second is **model evaluation**, *i.e.*, comparing competing explanations and assessing how well the model generalises beyond the data used to fit it.

These tasks are important in different ways. A model may pass all diagnostic checks and still be the wrong model for your scientific question. A model may fit poorly in-sample but capture the right structure and generalise well. Keeping the

two roles separate (as in, they are different steps with different methods) prevents a common failure, which is treating clean diagnostics as evidence of a good answer when the question was wrong or the predictors were poorly measured.

Chapter 16 addressed problems that arise before fitting (we looked at collinearity, confounding, measurement error). Here, I pick up after fitting. Residual diagnostics cannot reveal problems that were introduced before the model was fitted; for example, a model with collinear predictors or an attenuated slope may produce perfectly well-behaved residuals. Model checking and the assessments in Chapter 16 are therefore both necessary.

17.1 IMPORTANT CONCEPTS

- **Model checking** asks whether a fitted model is adequate relative to its assumptions and the data. It uses residuals, influence diagnostics, and assumption checks.
- **Model evaluation** compares a set of plausible models and assesses how well a model generalises using information criteria, model comparison, and cross-validation.
- **Residuals** are the main checking tool because they show what the model failed to explain and where structure may remain.
- **Leverage and influence** identify observations that contribute disproportionately to the fitted result.
- **AIC** compares models based on in-sample fit penalised for complexity, whereas **cross-validation** evaluates predictive performance on unseen data. They are complementary, but different.
- **Good diagnostics do not guarantee correct interpretation.** Collinearity and measurement error, discussed in Chapter 16, may be invisible in residual plots.

17.2 NATURE OF THE DATA AND ASSUMPTIONS

Model checking begins only after a model has been fitted, but the logic depends on the same assumptions introduced earlier in the regression sequence.

Independence remains fundamental because residual dependence can make a model look better behaved than it really is. **Linearity** still matters unless a more flexible form has been justified. **Homoscedasticity** affects the reliability of standard errors and confidence intervals. And **approximate normality of residuals** supports the inferential use of t - and F -tests, even though mild deviations are common and usually unimportant.

The practical aim to decide whether the fitted model is adequate for the scientific question being asked, and that it is good enough that the inferences it produces can be defended.

17.3 R FUNCTIONS

The main functions used in this chapter are:

- `plot()` for the standard diagnostic panels from a fitted `lm()` object;
- `cooks.distance()` for a summary of influence;
- `AIC()` for comparing biologically plausible candidate models;
- `augment()` from **broom** when fitted values, residuals, or leverage diagnostics need to be joined back to the data.

17.4 A WORKED DIAGNOSTIC EXAMPLE

I continue with the seaweed example from the previous chapters and fit the selected multiple regression model again.

```
sw <- read.csv(here::here("data", "BCB743", "seaweed", "spp_df2.csv"))
sw_ectz <- sw %>%
  filter(bio == "ECTZ")

mod_eval <- lm(Y ~ augMean + febSD + augSD, data = sw_ectz)
summary(mod_eval)
```

```
R>
R> Call:
R> lm(formula = Y ~ augMean + febSD + augSD, data = sw_ectz)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -0.153994 -0.049229 -0.006086  0.045947  0.148579
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  0.028365   0.007020   4.040 6.87e-05 ***
R> augMean      0.283335   0.011131  25.455 < 2e-16 ***
R> febSD        0.049639   0.008370   5.930 8.73e-09 ***
R> augSD        0.022150   0.004503   4.919 1.47e-06 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.06609 on 285 degrees of freedom
R> Multiple R-squared:  0.8387, Adjusted R-squared:  0.837
R> F-statistic: 494.1 on 3 and 285 DF, p-value: < 2.2e-16
```

The structure of what follows mirrors the split above. Sections 5.1 through 5.3 address model checking (residual structure, assumption limits, and influence). Sections 5.4 through 5.6 address model evaluation, in other words, model comparison and generalisation.

17.4.1 Residual Diagnostics

Residual plots are the first step in model checking. They reveal whether systematic structure remains after fitting and whether the distributional assumptions are broadly satisfied.

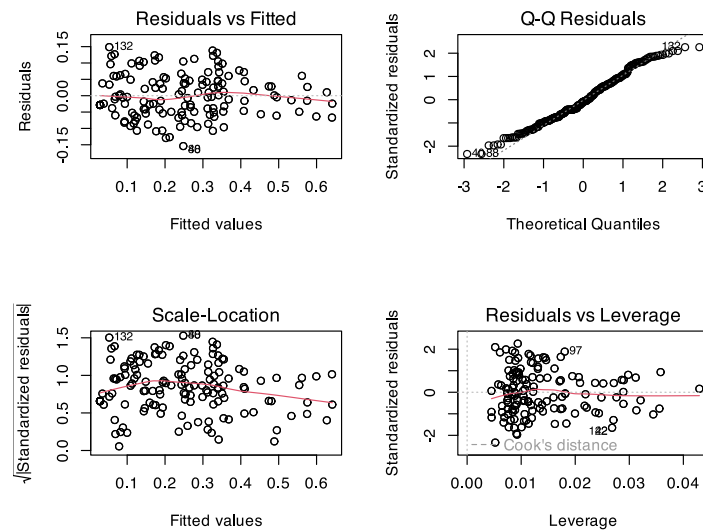


FIGURE 17.1. Standard diagnostic panels for the selected three-predictor seaweed model. Clockwise from top left: residuals versus fitted values (linearity and constant variance), normal Q-Q plot (residual normality), residuals versus leverage with Cook's distance contours (influential observations), and scale-location plot (homoscedasticity).

Figure 17.1 shows the four standard diagnostic panels. These allow assessment of several things simultaneously:

- **Residuals vs fitted values** for curvature or systematically changing variance;
- **Normal Q-Q plot** for approximate normality of residuals;
- **Scale-location plot** for heteroscedasticity (non-constant spread);
- **Residuals vs leverage** for observations that may be disproportionately influential.

No diagnostic plot should be thinking there will be clear-cut signs of danger. The question is whether the pattern is severe enough to undermine the scientific use of the model. In this example the plots suggest that the model is broadly adequate:

residuals are roughly centred around zero with no strong curvature, the Q-Q plot shows only mild tail deviation, and the leverage panel does not flag any single observation as severely influential.

A **failure pattern** would look quite different. For example, a strong funnel shape in the residuals-vs-fitted panel indicates heteroscedasticity; systematic curvature indicates that a predictor needs a nonlinear term or transformation; a small cluster of points with Cook's distance substantially above the rest indicates that the fitted coefficients are being driven by a few observations rather than by the bulk of the data. You should keep these visual failure signatures in mind when reading any residual plot.

17.4.2 *Assumption Checks and Their Limits*

Formal tests for normality or heteroscedasticity can be useful supplementary tools, but they should not replace residual plots and biological judgement.

The practical points are that tiny deviations from normality are usually unimportant since regression inference is robust to mild non-normality, especially in larger samples. Large datasets make this worse in the opposite direction, because even trivial deviations from the ideal distribution will produce a highly significant formal test result, even when the deviation has no practical consequence for inference. Visual residual structure is usually more informative than a single assumption test because a plot conveys the pattern *and* its magnitude simultaneously.

So, the question is actually whether the model is adequate for the inferential goal, not whether it is mathematically perfect.

There is one further limitation worth mentioning outright. Residual diagnostics cannot reveal problems that were introduced at the predictor level. So, a model with highly collinear predictors will typically produce residuals that look perfectly well-behaved, but the collinearity is hidden in the instability of the individual coefficients, not in the residual structure. Similarly, measurement error in a predictor attenuates the estimated slope but leaves the residuals looking reasonable. Model checking must therefore be interpreted alongside the assessments in [Chapter 16](#).

In the end, clean diagnostics indicate that the model is adequate for the data it was given but they do not guarantee that the data were right for the question.

17.4.3 *Influence and Leverage*

Not all unusual observations should be equally concerning.

A data point with a large **residual** is poorly fitted, so the model does not predict it well. A point with high **leverage** occupies an extreme position in predictor

space because it has an unusual combination of predictor values. A point with high **influence** materially changes the fitted coefficients when it is included or excluded.

Cook's distance is one commonly used summary of influence.

```
cook_summary <- summary(cooks.distance(mod_eval))
cook_summary
```

```
R>      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
R> 2.020e-08 5.316e-04 1.712e-03 3.061e-03 4.956e-03 1.876e-02
```

An influential point should not just be deleted automatically. It may reflect a data entry error, a measurement problem, a biologically informative extreme case, or a model that is too simple for the pattern in the data. The right response depends on which of these it is, which requires investigation rather than mechanical trimming.

17.4.4 Model Comparison

Biological analysis often involves several plausible models. Model comparison should therefore be used for comparing **competing explanations** and not as a numerical optimisation problem.

AIC (Akaike Information Criterion) compares models based on their in-sample fit penalised for complexity. In this framework, models with more parameters are penalised because they explain the present data at the cost of fitting more noise. A lower AIC is preferred, but the scale is relative; only differences between models should be concerned with, not the raw value.

I compare three models from the seaweed example:

- a null model (intercept only, no predictors);
- the selected three-predictor model;
- the larger five-predictor model considered in Chapter 16.

```
null_mod <- lm(Y ~ 1, data = sw_ectz)
full_mod <- lm(Y ~ augMean + febRange + febSD + augSD + annMean, data = sw_ectz)

AIC(null_mod, mod_eval, full_mod)
```

```
R>      df      AIC
R> null_mod  2 -222.8193
R> mod_eval  5 -744.1734
R> full_mod  7 -826.4472
```

The selected model is far better than the null model, so the predictors explain substantial variation in Sørensen dissimilarity. Its AIC is competitive with the larger model while remaining simpler. The important general lesson is that a more complex model is not automatically a better scientific model, especially when its complexity carries collinear predictors whose coefficients cannot be cleanly interpreted.

17.4.5 Overfitting and Generalisation

A model can fit the current data very well while performing poorly on new data. This is **overfitting**, *i.e.*, the model has “learned” specific features of this dataset, including its noise, that do not generalise.

Overfitting is especially likely when the sample size is modest relative to the number of predictors, when interaction terms are added freely, or when model choice is heavily data-driven rather than theory-guided. This is one reason why the stepwise selection approaches that dominated earlier statistical practice are now viewed with scepticism. That is, they optimise in-sample fit through a process that naturally finds the noise.

The main distinction is between **in-sample fit** (how well the model describes the data used to estimate it) and **generalisation** (how well it would perform on genuinely new data). These can diverge substantially. Explanation and prediction have different requirements regarding model checking.

17.4.6 Cross-Validation

Cross-validation assesses predictive performance on held-out data, directly measuring generalisation rather than inferring it from in-sample penalties.

The approach involves dividing the data into parts, fitting the model on all-but-one part, testing it on the held-out part, and repeating the process. The prediction error averaged across all held-out portions is a direct estimate of how well the model performs on new data.

AIC and cross-validation answer different questions. AIC compares models on in-sample fit penalised for complexity, so it is a model selection tool for the current dataset. Cross-validation estimates how well a fitted model would perform on data it has not seen, making it a generalisation tool. They are not interchangeable. A model with the lowest AIC is not necessarily the best predictor on new data, and a model that cross-validates best may not have the most interpretable coefficients.

```
set.seed(74416)

fold_id <- sample(rep(1:10, length.out = nrow(sw_ectz)))

cv_rmse <- function(formula, data, folds) {
  tibble(fold = sort(unique(folds))) %>
    mutate(
      rmse = map_dbl(fold, \(k) {
        train_dat <- data[folds != k, , drop = FALSE]
        test_dat <- data[folds == k, , drop = FALSE]

        mod <- lm(formula, data = train_dat)
        pred <- predict(mod, newdata = test_dat)
```

```

    sqrt(mean((test_dat$Y - pred) ^ 2))
  })
}

cv_mod_eval <- cv_rmse(Y ~ augMean + febSD + augSD, sw_ectz, fold_id) >
mutate(model = "Selected three-predictor model")

cv_full_mod <- cv_rmse(Y ~ augMean + febRange + febSD + augSD + annMean,
  sw_ectz,
  fold_id) >
mutate(model = "Larger five-predictor model")

cv_results <- bind_rows(cv_mod_eval, cv_full_mod)

cv_results >
group_by(model) >
summarise(
  mean_rmse = mean(rmse),
  sd_rmse = sd(rmse),
  .groups = "drop"
)

R> # A tibble: 2 x 3
R>   model                                mean_rmse sd_rmse
R>   <chr>                                <dbl>    <dbl>
R> 1 Larger five-predictor model          0.0577 0.00615
R> 2 Selected three-predictor model       0.0663 0.00577

ggplot(cv_results, aes(x = model, y = rmse, fill = model)) +
  geom_boxplot(alpha = 0.8, width = 0.55, show.legend = FALSE) +
  geom_jitter(width = 0.08, alpha = 0.8, size = 1.4, show.legend = FALSE) +
  labs(x = NULL, y = "Cross-validated RMSE") +
  theme_grey() +
  theme(axis.text.x = element_text(angle = 12, hjust = 1))

```

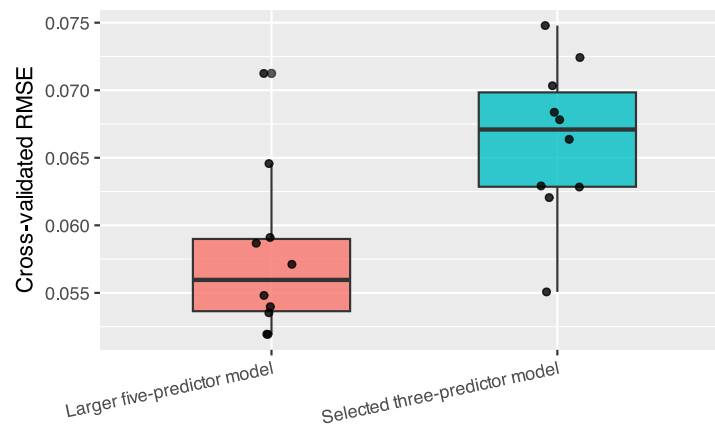


FIGURE 17.2. Ten-fold cross-validation RMSE for the selected three-predictor seaweed model and a larger five-predictor alternative.

The cross-validation results in Figure 17.2 show that the larger model does improve out-of-sample prediction, but only modestly. The mean RMSE decreases from about 0.066 for the selected three-predictor model to about 0.058 for the larger five-predictor model. That is a real predictive gain, but it is not dramatic. Whether to prefer the simpler or the more complex model depends on whether the goal is biological interpretation — where simpler and more stable is usually better — or prediction, where the marginal gain in RMSE may justify the additional complexity.

Neither AIC nor cross-validation resolves this trade-off automatically. AIC says the models are competitive; cross-validation says the larger model predicts fractionally better. The scientific question remains the arbiter.

17.4.7 What the Diagnostics Tell Us

The model is adequate in the sense that no major assumption violation is evident and no single observation dominates the fitted result. The selected three-predictor model explains substantially more than the null and performs comparably to the five-predictor alternative on both in-sample and out-of-sample criteria.

What the diagnostics do not tell us is whether the coefficients carry clean biological interpretations. The predictor selection steps in Chapter 14 already identified strong collinearity among the candidate climate variables, and the selected model still contains predictors that partly track the same climatic gradient. That collinearity does not appear in the residual plots; they are well-behaved regardless. This is the point made in Section 5.2, where I concluded that clean diagnostics do not guarantee interpretable coefficients.

The biological conclusion is that Sørensen dissimilarity in the East Coast Transition Zone is systematically related to the selected climate predictors. The model is adequate for describing that relationship. Whether the individual predictor effects represent distinct biological mechanisms (rather than overlapping signals from a shared climatic gradient) cannot be resolved by diagnostics alone.

17.5 A PRACTICAL WORKFLOW

After fitting a regression model:

1. inspect the residual plots and look specifically for curvature, funnel shapes, and strongly influential points;
2. assess leverage and influence and investigate any observation with substantially elevated Cook's distance;
3. check whether the residual structure suggests a missing predictor or a misspecified functional form;

4. compare only biologically sensible candidate models using AIC or a formal test, not data-driven searches;
5. use cross-validation if generalisation and predictive performance are part of the scientific question;
6. balance fit, complexity, and interpretability, so the best-fitting model is not always the most useful one;
7. report limitations clearly, including what the diagnostics cannot reveal.

17.6 REPORTING

A journal article should not present model checking as a raw list of diagnostics. Instead, it should state what was checked, what problems were looked for, and what conclusion was reached about the adequacy of the model.

WRITE-UP

Methods

After fitting the selected multiple regression model, diagnostic plots were examined to assess linearity, homoscedasticity, and the approximate normality of residuals. Leverage and influence were also inspected to identify observations that might disproportionately affect the fitted model. Model performance was compared with that of both an intercept-only model and a larger alternative model using information-theoretic criteria (AIC), and predictive performance was further assessed using 10-fold cross-validation.

Results

Diagnostic evaluation of the selected multiple regression revealed no major departures from model assumptions. Residuals were distributed reasonably evenly across the fitted range, the Q-Q plot suggested only mild deviation from normality, and no individual observation appeared sufficiently influential to undermine interpretation. Model comparison indicated that the selected three-predictor model was strongly preferred to the intercept-only model and achieved a fit comparable to that of the larger five-predictor model while retaining a simpler and more interpretable structure. Ten-fold cross-validation showed that the larger model reduced mean out-of-sample RMSE from about 0.066 to 0.058, indicating a modest predictive advantage at the cost of additional complexity.

Discussion

These results suggest that the selected model provides an adequate statistical description of the data and that its biological interpretation is not driven by obvious assumption violations or isolated influential observations. The comparison with the larger candidate model reinforces a central modelling principle: additional complexity is justified only when it yields a clearer or substantially better explanation, not simply because it reduces AIC. Adequacy of model diagnostics does not, however, guarantee

correct interpretation if predictors are poorly measured, collinear, or confounded with unmeasured variables; these are limitations that diagnostics alone cannot reveal.

17.7 SUMMARY

- Model checking and model evaluation are conceptually distinct tasks. Checking asks whether the model is adequate for the data; evaluation compares competing models and assesses generalisation.
- Residual diagnostics are the primary checking tool. Look for curvature, heteroscedasticity, and influential observations, not just for the absence of obvious problems.
- AIC compares models on in-sample fit penalised for complexity. Cross-validation estimates predictive performance on unseen data. They are complementary, not interchangeable.
- Clean diagnostics do not imply correct interpretation. Collinearity and measurement error, introduced in Chapter 16, can be invisible in residual plots while still distorting coefficient estimates.
- Model comparison should weigh competing biological explanations, not just minimise a criterion. The best-fitting model is not always the most scientifically informative one.

At this point the core modelling spine is in place: simple regression, multiple regression, interactions, threats to interpretation, and model evaluation. The later chapters extend that same logic to more specialised modelling situations.

Part IV

Extensions

The previous part established the core linear-model framework. This part extends that framework to the situations where the basic assumptions no longer describe the data well enough: non-independence, grouped structure, non-Gaussian responses, smooth nonlinear trends, mechanistic nonlinear forms, conditional quantiles, and prediction-oriented regularisation.

These chapters should not be read as a collection of advanced tricks. Each method responds to a particular kind of mismatch between question, design, data, and model. The point is therefore not only to learn a new function or model family, but to recognise the specific problem that makes the extension necessary.

By this stage, the statistical toolkit is broad. The challenge is judgement: knowing when the simple model is still defensible, and when the structure of the data requires something richer.

18

Pseudoreplication

When Sample Size Is Illusory

IN THIS CHAPTER

- What pseudoreplication is and why it matters, in both experiments and observational surveys
- The experimental or observational unit: what it is, how to identify it, and why it governs inference
- Why units of study must cover the full domain of intended inference
- The types of pseudoreplication and how to recognise them
- How pseudoreplication distorts statistical inference
- How better design prevents false replication
- When mixed-effects models are required

TASKS TO COMPLETE IN THIS CHAPTER

- None

I asked in Chapter 17 whether your model fits the data. Now I ask a prior question, whether the data structure allows the model's assumptions to hold. Independence of observations is central to every test covered in this course, including t-tests, ANOVA, and regression. When observations are not independent, the assumption is violated before the model is fitted, and no amount of diagnostic checking will rescue the inference. The violation originates in the study design, not the analysis.

Pseudoreplication is the most common form of this failure in ecological research. It occurs when multiple measurements from a single unit of study are treated as

independent replicates of a treatment or condition. Hurlbert (1984) framed the problem in terms of manipulative experiments, but the error appears just as often in observational surveys where no manipulation takes place. Whenever subsamples drawn from one site, one organism, or one time period are counted as independent representatives of a broader group or condition, the same failure is in play. The problem has not gone away.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2), 187–211.

18.1 IMPORTANT CONCEPTS

- **Pseudoreplication:** A fundamental error in experimental design where samples or measurements that are not statistically independent are treated as if they are independent replicates.
- **Experimental / Observational Unit:** In a manipulative experiment, the experimental unit is the smallest entity to which a treatment can be independently applied. In an observational study, the equivalent is the observational unit, *i.e.*, the entity independently selected to represent a condition or group. Correctly identifying this unit is essential regardless of whether we have an experimental study or an observational one.
- **Invalid Error Estimation:** Pseudoreplication uses the variation *among subsamples* (typically small, because subsamples share an environment and a treatment application) as a proxy for the variation *among true replicates* (typically larger), leading to an underestimate of the true experimental error.
- **Inflated Type I Error:** By underestimating the error term, pseudoreplication makes it far more likely to find a statistically significant result by chance (a false positive).
- **Subsample vs Replicate:** A subsample provides additional description of one experimental unit; a replicate provides an independent estimate of the treatment effect.
- **Domain of Inference:** The population or conditions over which a statistical conclusion is intended to generalise. Replicates must be distributed across this domain, so a sample drawn from a restricted subset cannot support broader inference.

NOTE

The central question is, what is the unit (experimental or observational) that was independently assigned to, or selected to represent, a treatment or condition? Every design decision and every analysis follows from the answer.

18.2 WHAT IS PSEUDOREPLICATION?

Hurlbert defines pseudoreplication as...

“...the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent” (Hurlbert 1984).

Pseudoreplication occurs when multiple samples from the same unit of study are treated as though they are independent observations of a treatment or condition. This applies equally to manipulative experiments and to observational surveys. If you take ten sediment cores from a single estuary that receives high agricultural runoff, you do not have ten independent observations of the runoff effect. You have ten subsamples from one estuary. There is one estuary, and therefore one unit of study. Whether that estuary was assigned a treatment by you or simply selected because it represented a particular condition makes no difference to this logic: a statistical test that treats the cores as independent replicates commits pseudoreplication either way.

18.3 THE EXPERIMENTAL UNIT**18.3.1 What It Is**

Three kinds of units appear in any ecological study, and they are often conflated. The **unit of study** is the entity that represents an independent observation of the condition or group being compared. These may be an estuary, a plot, a tank, an individual organism. In a manipulative experiment, this is called the **experimental unit**, so the entity to which a treatment is independently applied. In an observational survey (where you measure rather than manipulate), the equivalent is the **observational unit**, *i.e.*, the entity independently selected to represent a category, condition, or location. The **sampling unit** is what is drawn from within each unit of study for measurement: organisms, cores, or quadrats within an estuary or plot. The **observation unit** is the thing actually measured, which is often the same as the sampling unit but sometimes it can be a sub-component of it (such as a gill tissue

sample from a fish or a leaf from a plant). Pseudoreplication arises when sampling units are analysed as though they were units of study.

Not all ecological research involves experiments. If you are comparing fish assemblages across estuaries affected by different levels of agricultural runoff, you are not assigning runoff levels to estuaries. You are selecting estuaries that already differ in that regard. But the logic of replication is identical either way, and the conclusion is the same: **the estuary is the replicate, not the fish.**

Each estuary is an independently selected unit representing a particular runoff condition. The fish you sample within it are subsamples of that estuary. They tell you about fish in that estuary; they do not constitute independent observations of the runoff effect. If you have five high-runoff estuaries and five low-runoff estuaries, and you sample 30 fish from each, your effective sample size for testing the runoff effect is five per group, not 150. The 30 fish within each estuary replicate nothing except that estuary. This holds whether the runoff difference arose from your experimental manipulation or from a pre-existing gradient you chose to exploit in a sampling campaign.

IMPORTANT

In both manipulative experiments and observational sampling campaigns, the replicate is the independently selected or assigned unit of study, not the individual organisms, cores, or measurements collected within it. Additional sampling within a unit increases your description of that unit; it does not increase your replication of the condition or treatment it represents.

Whether your study is experimental or observational affects how causation can be inferred, and this does not change what constitutes valid replication.

The distinction between unit of study and sampling unit is critical because units of study are independent of one another in a way that sampling units are not. Two estuaries assigned to different treatments, or selected because they represent different conditions, could in principle have been in opposite groups. One organism within an estuary could not. An organism shares that estuary's salinity regime, turbidity, tidal signature, and history with every other organism sampled there. It is not independent.

18.3.2 *How to Identify It*

Two questions reliably locate the unit of study.

First: could this unit have been in a different group, or could a different unit have been selected in its place, without altering the status of any other unit? If yes, it is a

candidate for the unit of study. In an experiment, one estuary could have been the control rather than the treated unit. In an observational survey, a different estuary could have been selected to represent high-runoff conditions. An individual fish within an estuary could not have been classified differently from its neighbours. It belongs to that estuary's condition by virtue of being there.

Second: if you know the response of one unit, does that tell you anything about the response of another? If yes, the two units are not independent, and at most one of them contributes an independent data point. Fish from the same estuary share water temperature, food supply, predator assemblage, and pollution history. Their measurements are correlated. Weekly chlorophyll-*a* readings from the same embayment are correlated across time. Quadrats from adjacent halves of the same field share underlying spatial gradients. In each case, the correlation structure reveals that the unit of study sits at a higher level than the individual measurement.

18.3.3 *The Domain of Inference*

Identifying the correct unit of study is necessary but not sufficient. Your units must also be distributed across the full domain over which inference is intended to reach. This requirement is separate from independence, but equally binding.

Suppose you want to make a claim about rocky intertidal communities along the South African coastline. Ten plots at a single site near Betty's Bay, however carefully replicated and independently treated, support inference only about communities at that site, or sites very similar to it in exposure, substratum, and thermal regime. They cannot ground a claim about Hermanus, Tsitsikamma, Port Nolloth, or Sodwana Bay. Your experimental units exist; they are independent; but they do not span the domain.

The same applies in time. A study conducted over a single summer cannot support inference about summer conditions in general, because that particular summer may have been anomalously warm, calm, or productive. Replication across years is what converts a one-season snapshot into a statement about the season. One year's summer is a subsample of "summer", not a replicate of it.

The principle extends to any axis of variation that defines the target population. If your inference is intended to reach across species, your units must include representative species. If it is intended to reach across habitats, your units must span habitat types. A study whose units cluster in one corner of the domain can describe that corner accurately and still fail to support the broader claim being made.

No study covers its domain completely, and extrapolation beyond the sampled range is sometimes defensible if the biology justifies it. Your role is to be explicit about the domain actually sampled, to distinguish it from the domain over which

your conclusions are stated, and to flag the gap. An transparent Discussion section names this limitation.

18.4 TYPES OF PSEUDOREPLICATION

Hurlbert distinguishes several forms, of which three recur most often in ecological practice.

Simple pseudoreplication arises when only one experimental unit per treatment exists but multiple subsamples are drawn from it. Ten water samples from a single fertilised pond are ten subsamples, not ten replicates. There is one pond. There us one treatment application. Therefore, there is only one replicate. Period.

Temporal pseudoreplication arises when repeated measurements from the same unit across time are treated as independent replicates. Measuring a plant's photosynthetic rate every hour for a day produces 24 correlated observations of one plant, not 24 independent replicates of the treatment applied to that plant. The same mistake appears when a single estuary is sampled monthly and the 12 months are analysed as though they were 12 independent estuaries exposed to the same condition. Repeated measurements through time can be scientifically valuable, but they are repeated observations on the same unit, not new replicates of that unit. Genuine temporal replication requires independent units observed across time, or a design and model that explicitly treat time as a repeated-measures structure rather than as a source of independent sample size.

Interspersion pseudoreplication arises when two treatments occupy spatially contiguous, non-interspersed areas (one half of a field sprayed, the other half unsprayed) so that any spatial gradient is confounded with the treatment. The two "conditions" are not independent; they simply differ in position and treatment.

18.5 WHY IS IT A PROBLEM?

Variation among subsamples within a single unit of study is almost always smaller than variation among independent units, because subsamples share the same environment, history, and treatment application. When this within-unit variation is used as the error term in a t -test or ANOVA, the denominator of the test statistic is too small. Standard errors shrink, test statistics inflate, and the nominal α level no longer controls Type I error at its stated rate. The result is a false positive, so a conclusion of a significant treatment effect that reflects within-unit subsample variation, not a real treatment response.

Hurlbert argued, and nothing since has contradicted him, that it is a **fundamental design failure that invalidates inference**, and so a p -value calculated

from pseudoreplicated data does not mean what you think it means. Unfortunately, this failure continues to be perpetuated even by seasoned biologists.

18.6 WORKED EXAMPLES

The following examples follow the same structure. They identify the experimental unit, show how the incorrect design substitutes subsamples for replicates, and describe what a valid design would require. In each case, the error arises because subsamples are treated as independent replicates of a treatment applied at a higher level.

18.6.1 Example 1: Comparing Two Ponds

a) Incorrect Design (Simple Pseudoreplication). Suppose you want to test whether a new fish food increases fish growth. Two ponds are available. You add the new food to Pond A and use Pond B as the control. After three months, you sample 50 fish from each pond and run a t-test on their weights.

b) The Flaw. You have treated the 50 fish as independent replicates, but your experimental unit is the pond. All fish within a pond share the same environment and the same treatment. They do not provide independent information about the treatment effect. Your test therefore uses within-pond variation as if it were between-replicate variation, which underestimates the error term and inflates the test statistic. The result cannot distinguish a treatment effect from inherent differences between the two ponds.

c) Correct Design. You need multiple ponds for each treatment: say, five ponds receiving the new food and five control ponds. Fish sampled within each pond estimate mean growth for that pond. The pond mean, not the individual fish, is your unit of analysis ($n = 5$ per group).

d) Reporting.

WRITE-UP

Methods

Fish growth was compared between treatment and control using replicate ponds as the experimental units. Multiple fish were sampled within each pond to estimate mean growth per pond, but the pond rather than the individual fish was the replicate in the analysis.

Results

Mean fish growth differed between treatment and control ponds only to the extent supported by variation among ponds. The analysis therefore assessed the treatment effect against the correct error term.

Discussion

Treating individual fish as replicates would have overstated the effective sample size and inflated confidence in the treatment effect. The design and analysis must match the scale at which the treatment was applied.

18.6.2 Example 2: Interspersion

a) Incorrect Design (Segregated Treatments). Suppose you want to test whether an insecticide reduces an insect population in a field. You divide the field in half, spray one half (Treatment), and leave the other unsprayed (Control). You then take 20 sweep-net samples from each half.

b) The Flaw. You have treated the 20 samples within each half as independent replicates, but your experimental unit is the field plot at the scale of treatment application. Because the two conditions occupy contiguous, non-interspersed halves, any underlying environmental gradient (soil type, drainage, prevailing wind) is confounded with the treatment. Your test attributes spatial variation to the treatment effect, which biases the error estimate and makes inference unreliable.

c) Correct Design. Divide the field into a grid of smaller plots and randomly assign treatment and control in an interspersed pattern. Each plot becomes an independent experimental unit. Sweep-net samples within each plot characterise local abundance; the plot is your replicate.

*d) Reporting.***WRITE-UP****Methods**

Insect abundance was assessed in replicate field plots to which treatment and control conditions were randomly assigned in an interspersed design. Sweep-net samples within each plot were used to characterise local abundance, but the plot was the experimental unit in the analysis.

Results

Differences in insect abundance between treatment and control were evaluated among replicate plots, separating the treatment effect from background spatial heterogeneity.

Discussion

Had one half of the field been treated and the other used as control, any underlying environmental gradient could have been mistaken for an insecticide effect. Interspersion prevents this confounding.

18.6.3 Example 3: Temporal Structure

a) Incorrect Design (Temporal Pseudoreplication). Suppose you want to test whether a nutrient addition increases algal biomass in a lake. You apply the treatment to one lake, measure chlorophyll-*a* concentration monthly over one year, and compare the 12 post-treatment measurements to 12 measurements from the preceding year (the control period) using a t-test.

b) The Flaw. You have treated each monthly observation as an independent replicate of the treatment effect, but all measurements come from one lake. Your experimental unit is the lake, not the month. Monthly observations within a lake are temporally autocorrelated: conditions in one month influence the next. They do not provide independent information about the treatment. Your test uses within-lake temporal variation as the error term, which underestimates the variance relevant for testing the treatment effect.

c) Correct Design. You need multiple independent lakes, assigned to treatment and control. Repeated measurements within each lake across time remain legitimate for describing within-lake dynamics, but the lake mean (or a model that uses lake as a grouping factor) must be the basis of inference. The month is a subsample; the lake is the replicate.

d) Reporting.

WRITE-UP

Methods

The effect of nutrient addition on algal biomass was evaluated using replicate treatment and control lakes. Chlorophyll-*a* was measured repeatedly through time within each lake, but the analysis treated lake as the experimental unit and time as repeated sampling within units.

Results

Algal biomass differed between treatment and control only insofar as the replicated lake-level responses supported that conclusion. Temporal variation within individual lakes was not used as a substitute for replication of the treatment effect.

Discussion

Repeated monthly observations improve description of within-lake dynamics but do not create additional independent replicates of a treatment applied to a single lake.

18.6.4 Example 4: Seasonal Comparison

a) Incorrect Design (Seasonal Pseudoreplication). Example 3 examined a before-after comparison within a single unit, where repeated measurements in time are treated as independent replicates. This example examines a related but distinct failure: comparison between categories where the categories themselves are not replicated. Suppose you want to test whether algal biomass differs between summer and winter. You sample chlorophyll-*a* weekly from a single coastal bay over one year and compare all summer observations (December–February) to all winter observations (June–August) using a t-test.

b) The Flaw. You have treated each weekly observation as an independent replicate of the seasonal effect, but all observations come from one bay in one year. “Season” is not replicated: there is one summer and one winter. The observations are temporally autocorrelated, and any between-season difference is confounded with other between-period variation in this particular bay and year. Your test uses within-season temporal variation as if it were between-replicate variation, underestimating the error term and inflating the apparent strength of the seasonal effect.

c) Correct Design. To test for seasonal differences, you need to replicate the seasonal contrast across independent units: multiple bays sampled within each season, or the comparison repeated across multiple years with year treated as a replicate. Weekly samples within each unit remain repeated observations, not replicates.

d) Reporting.

WRITE-UP

Methods

Seasonal differences in algal biomass were assessed across replicated bay or year units so that the seasonal contrast was evaluated against independent replication rather than against repeated observations within a single annual time series. Weekly samples within each unit were treated as repeated observations rather than independent replicates.

Results

Summer algal biomass was higher than winter biomass only insofar as this difference was consistent among independent units. The seasonal contrast was judged against among-unit variation, not against week-to-week fluctuation within one annual record.

Discussion

Without replication of the seasonal contrast, an apparent seasonal difference could reflect the idiosyncrasies of one bay or one year. Repeated sampling within a system refines description; it does not by itself justify inferential comparison among categories.

18.7 INTERPRETING THE PROBLEM

When pseudoreplication occurs, the effective sample size is inflated. If you sample 50 fish from one pond and report $n = 50$, you are claiming far more independent information about the treatment than your design provides. The actual inferential leverage is one pond, and a single pond cannot distinguish a treatment effect from inherent differences between ponds.

The consequence for the model is equally direct. The error term (the denominator of the F-statistic or t-statistic) is built from within-unit subsample variation rather than from variation among independent replicates. Subsample variation is almost always smaller, because subsamples share an environment, a history, and the same treatment application. Standard errors are therefore too small, test statistics are too large, and p-values are too optimistic. Inference is miscalibrated at every level: the point estimate of the effect may be reasonable, but the uncertainty attached to it is not.

This is also why pseudoreplication does not show up clearly in standard residual diagnostics. Residuals from a pseudoreplicated model can look well-behaved: normally distributed, homoscedastic, free of obvious patterns. That is because the within-unit variation being modelled is genuinely modest and internally consistent. The problem is structural, not distributional. It lives in the design, not in the fitted model. Chapter 17's tools are necessary but not sufficient: a model can pass every diagnostic check and still be built on a pseudoreplicated foundation.

18.8 SUMMARY

Two requirements govern valid inference. First, your units of study must be independent: each unit must have received its treatment independently of every other unit, and the response of one unit must not be predictable from the response of another. Second, your units must span the domain over which inference is claimed: conclusions drawn from a geographically restricted sample cannot reach a broader coastline; conclusions drawn from a single year cannot reach across seasons in gen-

eral. Pseudoreplication violates the first requirement. A sample that fails to cover its intended domain violates the second. Both failures produce conclusions that exceed what the data support.

Pseudoreplication cannot be corrected in the analysis once the data have been collected. The solution is correct design before data collection: identify the unit of study, ensure the treatment is independently applied to multiple units per treatment level, intersperse or randomise units to prevent confounding with spatial or temporal gradients, and distribute those units across the full domain of intended inference.

When repeated measurements or grouped data cannot be avoided, because the system permits only a small number of independent units or because temporal structure is intrinsic to the question, models that explicitly represent this nesting are required. Mixed-effects models provide the framework for analysing data in which observations are nested within higher-level units. Chapter 19 addresses these models directly.

As Hurlbert stated, “The question of what is the experimental unit... is the most important one that an experimenter has to answer.” It remains so.

19

Dependence and Mixed Models

When Observations Are Not Independent

IN THIS CHAPTER

- why non-independence creates misleadingly confident analyses;
- how grouped and nested sampling structures arise in ecology;
- the difference between fixed and random effects;
- how to fit and interpret a simple random-intercept mixed model in R;
- how to report a mixed model in the same journal style used elsewhere in the course.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Most of the models in the earlier chapters assume that each observation contributes genuinely new information. In ecological sampling that is often false, because measurements taken from the same site, transect, colony, quadrat, or individual tend to resemble one another more than measurements taken from different sampling units.

That is what **dependence** means in practice. So, the observations are not fully independent because they share context, history, or structure.

If we ignore that structure, the analysis becomes overconfident. Standard errors shrink, degrees of freedom are effectively exaggerated, and the model can start reporting strong-looking results that are partly artefacts of the sampling design.

Mixed-effects models provide one of the main solutions to this challenge. They allow us to keep the fixed effects that address the biological question while accounting for variation among grouped sampling units through **random effects**.

19.1 KEY CONCEPTS

- **Dependence** means observations are not fully independent of one another.
- **Grouped structure** arises when observations are clustered within sites, years, individuals, or other units.
- **Fixed effects** represent systematic effects of direct biological interest.
- **Random effects** represent structured variation among sampling units.
- **A random intercept model** allows each site or group to have its own baseline level.
- **Mixed models are tailored to a specific sampling structure or design.**

19.2 WHEN THIS METHOD IS APPROPRIATE

You should start thinking about mixed models when:

- observations are nested within sites, transects, plots, years, or individuals;
- the same unit is measured repeatedly;
- the sampling design clearly has hierarchical structure;
- treating all observations as independent would create pseudoreplication.

Typical cases include:

- quadrats within sites;
- birds sampled within beaches;
- repeated measurements on the same animals;
- monthly observations within monitoring stations.

19.3 NATURE OF THE DATA AND ASSUMPTIONS

Mixed models do not remove the need for ordinary model checking. So, for a Gaussian linear mixed model, the main assumptions are still:

1. independence **between** higher-level units such as sites or individuals;
2. approximately normal residuals;
3. reasonably constant residual variance;
4. a sensible specification of the random-effects structure.

The important difference is that we now no longer assume independence among observations inside the same group.

19.4 THE CORE EQUATIONS

A simple random-intercept mixed model can be written as:

$$Y_{ij} = \alpha + \beta_1 X_{1ij} + \cdots + \beta_p X_{pij} + b_j + \epsilon_{ij} \quad (19.1)$$

In Equation 19.1, Y_{ij} is the response for observation i in group j , the β terms are the fixed effects, b_j is the random intercept for group j , and ϵ_{ij} is the residual error.

The important idea is that the model has two sources of variation:

- **between-group variation**, captured by the random effect b_j ;
- **within-group variation**, captured by the residual term ϵ_{ij} .

Mixed models help with dependence because they explicitly allow observations from the same site or individual to share a common baseline.

19.5 R FUNCTIONS

The most common functions are:

- `lmer()` from **lme4** for Gaussian linear mixed models;
- `glmer()` from **lme4** for generalised mixed models;
- `nlme::lme()` when more explicit correlation structures are needed.

A simple random-intercept model looks like this:

```
lmer(response ~ predictors + (1 | group), data = df)
```

The term `(1 | group)` means that each group has its own intercept.

19.6 EXAMPLE 1: SHOREBIRD LANDING DISTANCE WITH SITE-LEVEL DEPENDENCE

19.6.1 Example Dataset

We use the `BeachBirds.csv` dataset from the course data folder. These data contain flight-initiation distances and landing distances for several shorebird species sampled across multiple beach sites. The important point emerging from this example is that observations collected at the same site are likely to be more similar to one another than observations collected at different sites.

Here we model `land.dist` as a function of bird `Species` and `Sex`, while allowing each `Site` to have its own baseline intercept.

```
birds <- read_csv(here::here("data", "BCB744", "BeachBirds.csv"),
  show_col_types = FALSE) >
mutate(
  Site = factor(Site),
  Species = factor(Species),
```

TABLE 19.1. A subset of the `BeachBirds.csv` data used in the mixed-model example.

Site	Species	Sex	flush.dist	land.dist
1	Oystercatcher	Female	12.8	150.9
1	Oystercatcher	Male	4.4	114.1
1	Oystercatcher	Female	9.5	153.7
1	Oystercatcher	Female	12.2	137.7
1	Oystercatcher	Female	10.1	143.3
1	Oystercatcher	Male	6.6	142.8
1	Oystercatcher	Male	15.4	134.1
1	Oystercatcher	Male	11.3	159.5
1	Oystercatcher	Male	13.8	150.4
1	Oystercatcher	Female	12.9	161.9

```
Sex = factor(Sex)
)
gt(head(birds, 10))
```

```
birds > summarise(
  n = n(),
  n_sites = n_distinct(Site),
  n_species = n_distinct(Species),
  mean_land = mean(land.dist),
  sd_land = sd(land.dist)
)
```

```
R> # A tibble: 1 x 5
R>       n n_sites n_species mean_land sd_land
R>   <int>   <int>   <int>     <dbl>   <dbl>
R> 1   399       5         4       73.1    44.2
```

```
ggplot(birds, aes(x = Site, y = land.dist)) +
  geom_boxplot(outlier.shape = NA, alpha = 0.4) +
  geom_jitter(width = 0.1, alpha = 0.35, size = 0.8) +
  labs(
    x = "Site",
    y = "Landing distance (m)"
  )
```

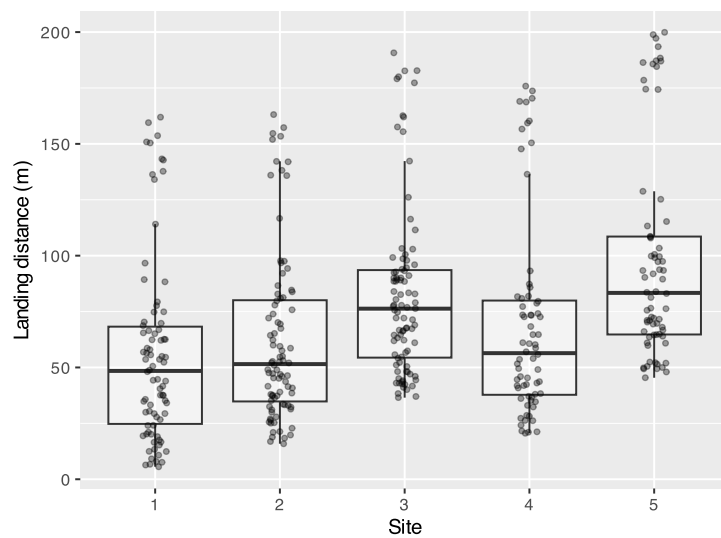


FIGURE 19.1. Landing distance by beach site for the sampled shorebirds.

```
ggplot(birds, aes(x = Species, y = land.dist, fill = Species)) +  
  geom_boxplot(alpha = 0.5, show.legend = FALSE) +  
  labs(  
    x = "Species",  
    y = "Landing distance (m)"  
  )
```

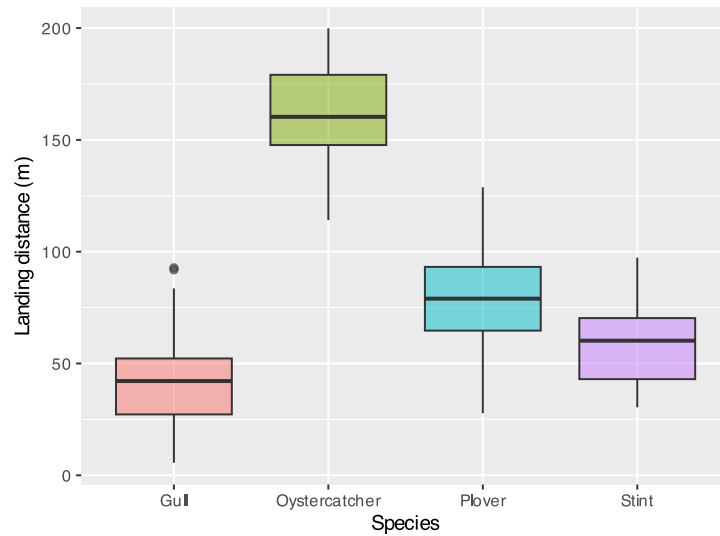


FIGURE 19.2. Landing distance differs strongly among shorebird species, but values also vary among sites.

The EDA in Figure 19.1 and Figure 19.2 already shows why an ordinary model may be too simple. Landing distance differs among species, but there is also clear site-to-site variability. That suggests a grouped structure in which site should not be ignored.

19.6.3 State the Model Question and Hypotheses

The main biological question is whether landing distance differs among species after accounting for the fact that birds were sampled within sites.

For the primary fixed effect, the hypothesis pair is:

H_0 : landing distance does not differ among species once site-level grouping is accounted for

H_a : landing distance differs among species after accounting for site-level grouping

The random effect is not interpreted through a simple null-hypothesis pair in the same way. Its role is to account for the grouped sampling structure.

19.6.4 Fit the Model

We first fit an ordinary linear model that ignores site, and then a mixed model with a random intercept for Site.

```
fit_lm <- lm(land.dist ~ Species + Sex, data = birds)
fit_lmer <- lmer(land.dist ~ Species + Sex + (1 | Site), data = birds)
summary(fit_lm)
```

```
R>
R> Call:
R> lm(formula = land.dist ~ Species + Sex, data = birds)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -51.779 -14.413   0.021  12.883  51.754
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)      42.637      1.852  23.028 < 2e-16 ***
R> SpeciesOystercatcher 120.509      2.938  41.014 < 2e-16 ***
R> SpeciesPlover       36.942      2.232  16.554 < 2e-16 ***
R> SpeciesStint        17.228      3.355   5.135 4.45e-07 ***
R> SexMale            -1.591      1.933  -0.823   0.411
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 19.15 on 394 degrees of freedom
R> Multiple R-squared:  0.8139, Adjusted R-squared:  0.812
R> F-statistic: 430.8 on 4 and 394 DF,  p-value: < 2.2e-16
```

```
summary(fit_lmer)

R> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
R> lmerModLmerTest]
R> Formula: land.dist ~ Species + Sex + (1 | Site)
R>   Data: birds
R>
R> REML criterion at convergence: 3178.2
R>
R> Scaled residuals:
R>      Min       1Q   Median       3Q      Max
R> -2.9082 -0.7167 -0.0061  0.6730  3.2716
R>
R> Random effects:
R>   Groups   Name      Variance Std.Dev.
R>   Site     (Intercept) 256.0    16.00
R>   Residual                167.9    12.96
R> Number of obs: 399, groups: Site, 5
R>
R> Fixed effects:
R>              Estimate Std. Error    df t value Pr(>|t|)
R> (Intercept)      43.061      7.265   4.177   5.927  0.00354 **
R> SpeciesOystercatcher 119.638      1.991 390.017  60.089 < 2e-16 ***
R> SpeciesPlover       37.689      1.515 390.044  24.870 < 2e-16 ***
R> SpeciesStint        18.858      2.275 390.024   8.290 1.85e-15 ***
R> SexMale            -2.260      1.311 390.029  -1.724  0.08550 .
R> ---
```

```

R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Correlation of Fixed Effects:
R>      (Intr) Spcs0y SpcsPl SpcsSt
R> Spcs0ystrct -0.074
R> SpeciesPlvr -0.100  0.334
R> SpeciesStnt -0.072  0.228  0.300
R> SexMale     -0.104  0.045  0.085  0.106

vc <- as.data.frame(VarCorr(fit_lmer))
site_var <- vc$vcov[vc$grp == "Site"]
resid_var <- vc$vcov[vc$grp == "Residual"]
icc <- site_var / (site_var + resid_var)

tibble(
  model = c("ordinary linear model", "mixed model"),
  AIC = c(AIC(fit_lm), AIC(fit_lmer))
)

```

```

R> # A tibble: 2 x 2
R>   model      AIC
R>   <chr>      <dbl>
R> 1 ordinary linear model 3495.
R> 2 mixed model          3192.

icc

```

```

par(mfrow = c(2, 2))
# plot(fit_lmer)
qqnorm(residuals(fit_lmer))
qqline(residuals(fit_lmer))
hist(residuals(fit_lmer), main = "Residuals", xlab = "Residual")
plot(fitted(fit_lmer), residuals(fit_lmer),
     xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2)
par(mfrow = c(1, 1))

```

```
R> [1] 0.603998
```

19.6.5 Test Assumptions / Check Diagnostics

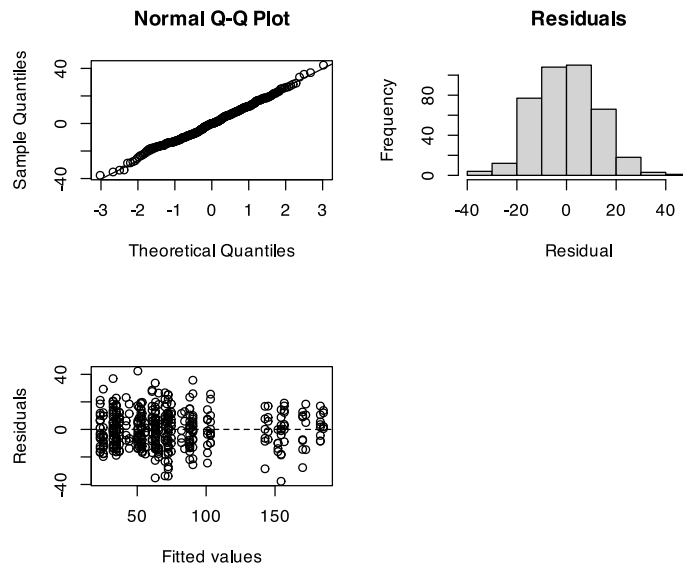


FIGURE 19.3. Standard residual diagnostics for the mixed model.

In a first-pass mixed-model workflow, the diagnostics in Figure 19.3 help answer the main questions:

- are the residuals grossly non-normal or strongly patterned;
- does the random effect explain a meaningful share of the variance;
- does the mixed model fit better than the independence-assuming alternative.

Here the random site effect is not trivial. The intra-class correlation is about 0.6, which means that roughly 60% of the variance represented by the model's variance components is at the site level rather than purely within sites.

19.6.6 Interpret the Results

The mixed model fit is clearly better than the ordinary model by AIC, which already suggests that site-level dependence matters.

The species effects are also large. Relative to the reference species, landing distance is substantially greater for Oystercatchers and Plovers, with Stints intermediate. The sex effect is much weaker and does not provide strong evidence for a difference at the conventional 0.05 level.

The ecological finding is that species differ, and those differences should be interpreted **after accounting for site-level dependence**, not before. If site is ignored, the analysis behaves as though all birds were sampled independently from one large homogeneous pool.

19.6.7 Reporting

WRITE-UP

Methods

Landing distance was analysed for shorebirds sampled across multiple beach sites using a linear mixed-effects model. Species and sex were included as fixed effects, and site was included as a random intercept to account for the grouped sampling structure. An ordinary linear model that ignored site was fitted for comparison.

Results

The mixed model fit the data better than the ordinary linear model (AIC 3495.2 vs 3192.2), indicating that site-level dependence was important. The estimated site-level variance was substantial, with an intra-class correlation of approximately 0.6. After accounting for this grouped structure, landing distance remained much greater in Oystercatchers ($\beta = 119.64$, $p < 0.001$), Plovers ($\beta = 37.69$, $p < 0.001$), and Stints ($\beta = 18.86$, $p < 0.001$) than in the reference species. The sex effect was weaker and not clearly different from zero ($p > 0.05$).

Discussion

Species differed in landing distance, but those differences were estimated while accounting for the non-independence of observations within sites. This makes the inference more defensible and avoids treating site-level similarity as if it were independent replication.

19.7 EXAMPLE 2: REPEATED SNAKE HABITUATION WITHIN INDIVIDUALS

The snake habituation dataset used earlier in the module is a better fit for this chapter than for the core ANOVA chapter. Each snake was measured repeatedly across days, so the observations are not independent. Treating snake as just another fixed factor in an additive ANOVA misses the main design point: repeated measurements on the same individuals create within-snake dependence.

```
snakes <- read_csv(here::here("data", "BCB744", "snakes.csv"),
                  show_col_types = FALSE) >
mutate(
  snake = factor(snake),
  day = factor(day)
)
snakes >
```

```

group_by(day) %>
summarise(
  mean_openings = mean(openings),
  sd_openings = sd(openings),
  .groups = "drop"
)

R> # A tibble: 4 x 3
R>   day mean_openings sd_openings
R>   <fct>      <dbl>      <dbl>
R> 1 1          63.3         30.5
R> 2 2          47          12.2
R> 3 3          34.5         26.0
R> 4 4          25.3         18.1

```

19.7.1 Why This Is a Dependence Problem

The biological question is whether the number of box openings required for habituation changes across repeated days. But each snake contributes several observations, one on each day. That means the data have a grouped structure: measurements from the same snake are likely to be more similar to one another than measurements from different snakes.

That grouped structure suggests a random-intercept model with snake identity as the grouping factor.

19.7.2 Fit a Naive Model and a Mixed Model

```

snakes_lm <- lm(openings ~ day, data = snakes)
snakes_lmer <- lmer(openings ~ day + (1 | snake), data = snakes)

summary(snakes_lm)

R>
R> Call:
R> lm(formula = openings ~ day, data = snakes)
R>
R> Residuals:
R>     Min       1Q   Median       3Q      Max
R> -41.333 -13.833  -3.333  11.500  43.667
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept)   63.333      9.304   6.807 1.28e-06 ***
R> day2          -16.333     13.158  -1.241  0.2289
R> day3          -28.833     13.158  -2.191  0.0404 *
R> day4          -38.000     13.158  -2.888  0.0091 **
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 22.79 on 20 degrees of freedom
R> Multiple R-squared:  0.3195, Adjusted R-squared:  0.2174
R> F-statistic:  3.13 on 3 and 20 DF,  p-value: 0.04852

```

```
summary(snakes_lmer)
```

```
R> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
R> lmerModLmerTest]
R> Formula: openings ~ day + (1 | snake)
R> Data: snakes
R>
R> REML criterion at convergence: 188.9
R>
R> Scaled residuals:
R>      Min       1Q   Median       3Q      Max
R> -1.8167 -0.6088 -0.1673  0.4153  1.9075
R>
R> Random effects:
R> Groups   Name              Variance Std.Dev.
R> snake    (Intercept)    29.68      5.448
R> Residual                    489.73   22.130
R> Number of obs: 24, groups: snake, 6
R>
R> Fixed effects:
R>              Estimate Std. Error    df t value Pr(>|t|)
R> (Intercept)    63.333      9.304   19.806   6.807 1.35e-06 ***
R> day2           -16.333     12.777   15.000  -1.278  0.22055
R> day3           -28.833     12.777   15.000  -2.257  0.03938 *
R> day4           -38.000     12.777   15.000  -2.974  0.00946 **
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Correlation of Fixed Effects:
R>      (Intr) day2    day3
R> day2 -0.687
R> day3 -0.687  0.500
R> day4 -0.687  0.500  0.500
```

```
tibble(
  model = c("ordinary linear model", "mixed model"),
  AIC = c(AIC(snakes_lm), AIC(snakes_lmer))
)
```

```
R> # A tibble: 2 x 2
R>   model      AIC
R>   <chr>    <dbl>
R> 1 ordinary linear model 224.
R> 2 mixed model          201.
```

The ordinary model behaves as though all snake-day observations are independent. The mixed model recognises that each snake has its own baseline tendency to habituate faster or more slowly.

19.7.3 Interpret the Results

The day effect remains the main biological signal: the number of openings required for habituation decreases across repeated days. But the mixed model estimates that change while accounting for repeated observations within snakes.

That is the crucial design lesson. This is not mainly a “two-way ANOVA without replication” problem. It is a repeated-measures problem in which snake identity induces dependence.

19.7.4 Reporting

WRITE-UP

Methods

The number of box openings required for rattlesnakes to habituate was analysed as a repeated-measures problem. Day was included as a fixed effect, and snake identity was included as a random intercept to account for repeated observations on the same individuals.

Results

The number of box openings required for habituation declined across repeated days, indicating that snakes habituated more rapidly with repeated exposure to the apparatus. Modelling snake identity as a random effect accounted for the non-independence created by repeated measurements on the same individuals, making the day effect interpretable at the correct design level.

Discussion

The important point is not only that habituation changed through time, but that the analysis respected the repeated-measures structure of the experiment. This is why the example belongs with dependence and mixed models rather than as a core factorial ANOVA illustration.

19.8 WHAT TO DO WHEN ASSUMPTIONS FAIL / ALTERNATIVES

- If the response is not Gaussian, move to a generalised mixed model with `glmer()`.
- If repeated measurements through time create autocorrelation, consider a model with explicit correlation structure rather than relying only on a random intercept.
- If the random-effects structure is biologically richer than a simple intercept, extend it cautiously to random slopes where the design supports that.
- If the design is fundamentally pseudoreplicated, a mixed model cannot rescue the study completely; it can only model the dependence that is actually observed.

19.9 SUMMARY

- Mixed models are needed when observations are grouped or otherwise dependent.
- The random effect is not an afterthought; it represents the sampling structure directly.

- In the beach-bird example, site-level dependence was strong enough to improve model fit markedly.
- Species differences in landing distance remained important even after that dependence was accounted for.

20

Generalised Linear Models

Extending Regression Beyond Normal Responses

IN THIS CHAPTER

- why ordinary linear models fail for binary, proportional, and count responses;
- how families and link functions define a generalised linear model;
- how simple proportion tests connect to the binomial GLM framework;
- how to fit and interpret a real logistic regression with ecological data;
- how to detect overdispersion in count data and what it means for interpretation.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Ordinary linear regression assumes a normally distributed response with constant variance. Many ecological responses do not meet those conditions at all. Presence-absence data are binary, proportions are constrained between 0 and 1, and counts are non-negative integers whose variance often changes with the mean.

Generalised linear models (GLMs) extend regression to handle those cases while retaining the same basic reasoning:

1. define the biological question;
2. choose predictors;
3. fit a model;
4. interpret the fitted effects.

The difference is that the response distribution and link function must now match the data structure.

20.1 KEY CONCEPTS

- **The family** describes the response distribution, such as binomial for binary data or Poisson for counts.
- **The link function** connects the linear predictor to the expected response scale.
- **A GLM still has a linear predictor**, but it acts on the link scale rather than directly on the response scale.
- **Interpretation depends on scale**: a coefficient may describe a change in log-odds or log-counts rather than a direct additive change in the response.
- **Overdispersion** is a warning that the simplest count model may be too optimistic.

20.2 WHEN THIS METHOD IS APPROPRIATE

Use a GLM when:

- the response is binary, such as present/absent or alive/dead;
- the response is a proportion derived from successes and failures;
- the response is a count;
- the variance changes with the mean in a way that an ordinary linear model cannot sensibly represent.

In this chapter, I pick up the earlier proportion-testing material and place it inside the broader regression sequence.

20.3 NATURE OF THE DATA AND ASSUMPTIONS

The assumptions depend on the GLM family, but some ideas are common and/or familiar:

1. observations should be independent;
2. the response family should match the data-generating structure;
3. the link function should make biological and statistical sense;
4. the linear predictor should be appropriately specified.

Unlike ordinary least squares, GLMs do not assume normal residuals with constant variance. Instead, the variance structure is part of the model family itself.

20.4 THE CORE EQUATIONS

A generalised linear model can be written as:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \quad (20.1)$$

In Equation 20.1, μ_i is the expected response for observation i , $g(\cdot)$ is the link function, and η_i is the linear predictor.

For binomial logistic regression, the most common link is the logit:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \quad (20.2)$$

Equation 20.2 says that predictors act linearly on the **log-odds** scale.

For Poisson regression, the most common link is the log link:

$$\log(\mu_i) = \alpha + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \quad (20.3)$$

Equation 20.3 means that predictors act linearly on the **log-count** scale.

20.5 R FUNCTIONS

The main function is `glm()`.

```
glm(response ~ predictors, family = ..., data = df)
```

Typical uses are:

```
glm(binary_response ~ predictors, family = binomial, data = df)
glm(cbind(successes, failures) ~ predictors, family = binomial, data = df)
glm(count ~ predictors, family = poisson, data = df)
glm(count ~ predictors, family = quasipoisson, data = df)
```

20.6 SIMPLE PROPORTION TESTS AS A BRIDGE TO GLMS

Before fitting full binomial models with predictors, it is useful to recognise that the simple proportion tests encountered earlier are part of the same broader reasoning framework.

If we only want to compare an observed proportion against a null expectation, or compare the proportions of two groups without additional predictors, `prop.test()` is still useful.

20.6.1 One-Sample Proportion Tests

For a two-sided one-sample proportion test, the hypotheses are:

$$H_0 : p = p_0$$

$$H_a : p \neq p_0$$

Here x is the number of successes, n is the number of trials, and p is the hypothesised probability:

```
prop.test(x = 45, n = 100, p = 0.5)

R>
R> 1-sample proportions test with continuity correction
R>
R> data: 45 out of 100, null probability 0.5
R> X-squared = 0.81, df = 1, p-value = 0.3681
R> alternative hypothesis: true p is not equal to 0.5
R> 95 percent confidence interval:
R> 0.3514281 0.5524574
R> sample estimates:
R> p
R> 0.45
```

```
prop.test(x = 33, n = 100, p = 0.5)

R>
R> 1-sample proportions test with continuity correction
R>
R> data: 33 out of 100, null probability 0.5
R> X-squared = 10.89, df = 1, p-value = 0.0009668
R> alternative hypothesis: true p is not equal to 0.5
R> 95 percent confidence interval:
R> 0.2411558 0.4320901
R> sample estimates:
R> p
R> 0.33
```

20.6.2 Two-Sample Proportion Tests

For a two-sided comparison of two proportions:

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$

```
mosquito <- matrix(c(70, 85, 50, 35), ncol = 2)
colnames(mosquito) <- c("yes", "no")
rownames(mosquito) <- c("Jack", "Jill")
mosquito

R>      yes no
R> Jack  70 50
R> Jill  85 35

prop.test(mosquito)

R>
R> 2-sample test for equality of proportions with continuity
R> correction
R>
R> data: mosquito
R> X-squared = 3.5704, df = 1, p-value = 0.05882
```

```
R> alternative hypothesis: two.sided
R> 95 percent confidence interval:
R> -0.253309811 0.003309811
R> sample estimates:
R> prop 1 prop 2
R> 0.5833333 0.7083333
```

These tests are useful when the question is simple and there are few or no predictors. Once predictors enter the problem, the binomial GLM becomes the natural extension.

20.7 EXAMPLE 1: LOGISTIC REGRESSION FOR PRESENCE-ABSENCE DATA

20.7.1 Example Dataset

We use the `Sessile_Benthic_Invertebrates.csv` dataset from the rocky intertidal course data. These observations were collected along a shore-position gradient (`Location_m`) and from the left and right sides of the shore platform. Here we derive a binary response indicating whether acorn barnacles were present in a sampled unit.

```
sessile <- read_csv(
  here::here("data", "BCB744", "Rocky Intertidal Data",
    "Sessile_Benthic_Invertebrates.csv"),
  show_col_types = FALSE
) >
mutate(
  Side = factor(Side),
  acorn_present = as.integer(Acorn.Barnacles > 0)
)
gt(head(sessile, 10))
```

20.7.2 Do an Exploratory Data Analysis (EDA)

```
sessile >
  summarise(
    n = n(),
    n_present = sum(acorn_present),
    prop_present = mean(acorn_present),
    min_loc = min(Location_m),
    max_loc = max(Location_m)
  )

R> # A tibble: 1 x 5
R>   n n_present prop_present min_loc max_loc
R>   <int>   <int>      <dbl>   <dbl>   <dbl>
R> 1  120      NA          NA      0     40

ggplot(sessile, aes(x = Location_m, y = acorn_present)) +
  geom_jitter(height = 0.05, width = 0, alpha = 0.35) +
  labs(
    x = "Location on shore gradient (m)",
```

TABLE 20.1. A subset of the sessile-benthic invertebrate data used for the logistic-regression example.

Location_m	Acorn.Barnacles	Goose.Neck.Barnacles	Anemones	Mussels	Side	acorn_present
0	2800		0	240	0 Left	1
0	2800		0	480	0 Right	1
0	0		0	0	0 Right	0
0	0		0	0	0 Left	0
0	4500		0	0	5 Left	1
0	13495		0	0	0 Right	1
0	300		0	0	0 Left	1
0	342		0	0	0 Right	1
0	5000		0	0	0 Right	1
0	1152		0	22	0 Left	1

```
y = "Acorn barnacles present"
)
```

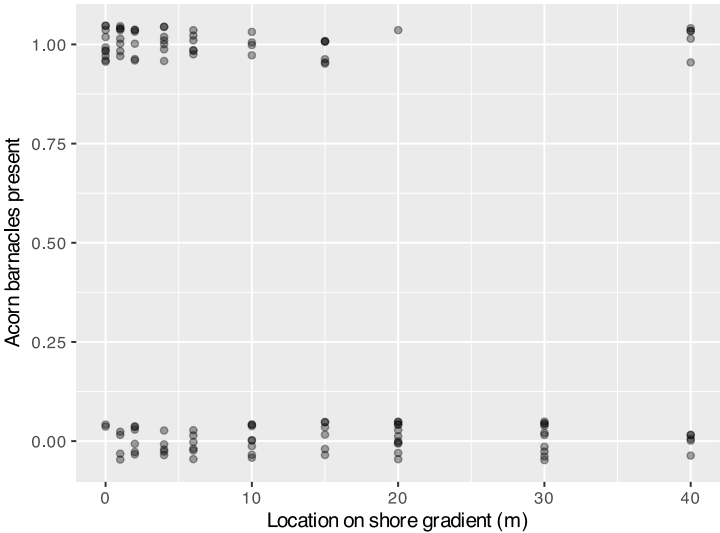


FIGURE 20.1. Presence and absence of acorn barnacles across the shore-position gradient.

The response pattern in Figure 20.1 is binary and clearly unsuitable for ordinary linear regression. That immediately points us to a binomial GLM.

20.7.3 State the Model Question and Hypotheses

The biological question is whether the probability of acorn-barnacle presence changes along the shore-position gradient.

$$H_0 : \beta_{\text{Location}} = 0$$

$$H_a : \beta_{\text{Location}} \neq 0$$

20.7.4 Fit the Model

```
fit_logistic <- glm(acorn_present ~ Location_m + Side,
  family = binomial,
  data = sessile)

summary(fit_logistic)

R>
R> Call:
R> glm(formula = acorn_present ~ Location_m + Side, family = binomial,
R>      data = sessile)
R>
R> Coefficients:
R>             Estimate Std. Error z value Pr(>|z|)
R> (Intercept)  0.46201    0.33287   1.388  0.16514
R> Location_m  -0.05270    0.01790  -2.944  0.00324 **
R> SideRight   -0.07575    0.38928  -0.195  0.84572
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> (Dispersion parameter for binomial family taken to be 1)
R>
R> Null deviance: 159.95  on 115  degrees of freedom
R> Residual deviance: 149.77  on 113  degrees of freedom
R> (4 observations deleted due to missingness)
R> AIC: 155.77
R>
R> Number of Fisher Scoring iterations: 4
```

20.7.5 Test Assumptions / Check Diagnostics

```
new_logistic <- expand.grid(
  Location_m = seq(min(sessile$Location_m), max(sessile$Location_m), length.out = 200),
  Side = levels(sessile$Side)
) >
as_tibble() >
mutate(prob = predict(fit_logistic, newdata = cur_data(), type = "response"))
```

```
ggplot(sessile, aes(x = Location_m, y = acorn_present, colour = Side)) +
  geom_jitter(height = 0.05, width = 0, alpha = 0.3) +
  geom_line(data = new_logistic, aes(y = prob), linewidth = 0.9) +
  labs(
    x = "Location on shore gradient (m)",
    y = "Probability of presence"
  )
)
```

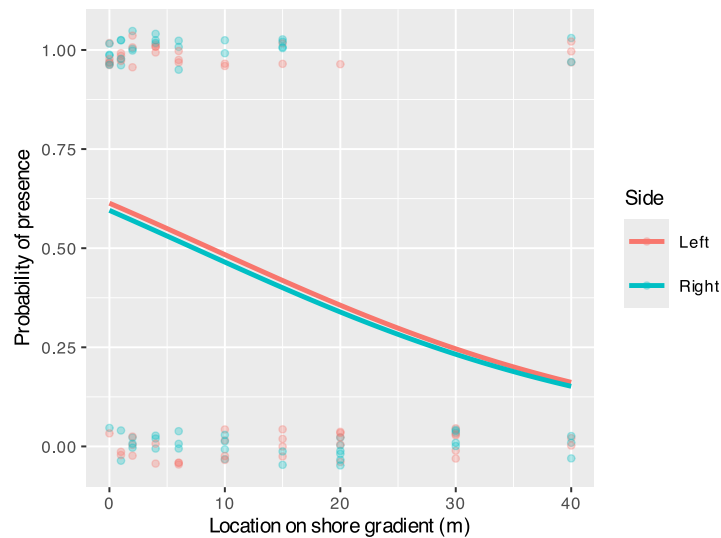


FIGURE 20.2. Fitted logistic-regression curves for acorn-barnacle presence along the shore-position gradient.

For introductory logistic regression, the most important checks are whether the response family is appropriate and whether the fitted probability pattern in Figure 20.2 is biologically sensible.

20.7.6 Interpret the Results

The fitted location coefficient is negative (-0.053 , $p < 0.01$), which means the log-odds of acorn-barnacle presence decline as shore position increases. In practical terms, the predicted probability of acorn-barnacle presence declines from about 0.61 near the lower end of the sampled gradient to about 0.16 near the upper end of the gradient.

The side effect is negligible in this example, so the dominant pattern is a decline in occurrence across shore position rather than a left-right platform difference.

20.7.7 Reporting

WRITE-UP

Methods

Acorn-barnacle occurrence was analysed using a binomial generalised linear model with logit link. Presence-absence was modelled as a function of shore position (`Location_m`) and platform side (`Side`) using rocky intertidal observations from the course dataset.

Results

The probability of acorn-barnacle presence declined significantly along the shore-position gradient (logit coefficient for `Location_m` = -0.053 , $z = -2.94$, $p < 0.01$). The fitted probability of presence on the left side of the platform fell from about 0.61 at the low end of the sampled gradient to about 0.16 at the high end. There was no evidence that left and right sides differed strongly once location was accounted for ($p > 0.05$).

Discussion

The ecological interpretation is that acorn barnacles became less likely to occur higher along the sampled shore gradient. In a manuscript, it is more helpful to phrase this as a change in probability of occurrence than to leave the result on the log-odds scale alone.

20.8 EXAMPLE 2: COUNT DATA AND OVERDISPERSION

20.8.1 Example Dataset

We now use the `Small_Mobile_Invertebrates.csv` rocky intertidal dataset. Here the response is the count of snails in each sampled unit.

```
mobile <- read_csv(
  here::here("data", "BCB744", "Rocky Intertidal Data", "Small_Mobile_Invertebrates.csv"),
  show_col_types = FALSE
) >
mutate(Side = factor(Side))
gt(head(mobile, 10))
```

20.8.2 Do an Exploratory Data Analysis (EDA)

```
mobile >
  summarise(
    n = n(),
    mean_snails = mean(Snails),
    var_snails = var(Snails),
    max_snails = max(Snails)
```

TABLE 20.2. A subset of the small mobile invertebrate data used for the count-data GLM example.

Location_m	Hermit.Crabs	Chitons	Snails	Limpets	Side
0	0	0	0	0	Left
0	0	0	3	9	Right
0	0	0	0	0	Right
0	0	0	0	0	Left
0	0	0	625	0	Right
0	0	1	0	1	Left
0	0	4	0	27	Left
0	0	0	134	11	Right
0	0	1	3	31	Left
0	0	0	0	16	Left

)

```

R> # A tibble: 1 x 4
R>       n mean_snails var_snails max_snails
R>   <int>      <dbl>      <dbl>      <dbl>
R> 1  120         NA         NA         NA

ggplot(mobile, aes(x = Location_m, y = Snails, colour = Side)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Location on shore gradient (m)",
    y = "Snail count"
  )

```

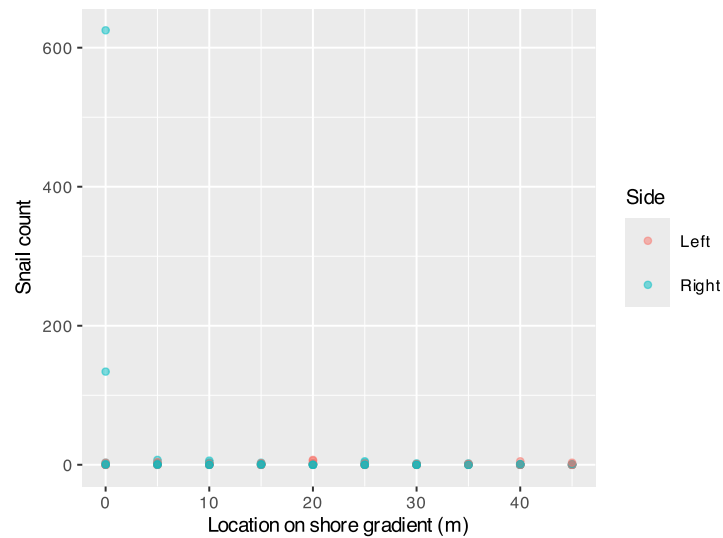


FIGURE 20.3. Snail counts across the shore-position gradient.

The counts in Figure 20.3 are non-negative integers and include many zeros together with some very large values. That already suggests we need to be cautious about the simplest Poisson assumptions.

20.8.3 State the Model Question and Hypotheses

The question is whether snail abundance changes along the shore gradient and between the two sides of the shore platform.

$$H_0 : \beta_{\text{Location}} = 0$$

$$H_a : \beta_{\text{Location}} \neq 0$$

20.8.4 Fit the Model

```
fit_pois <- glm(Snails ~ Location_m + Side,
               family = poisson,
               data = mobile)

fit_qpois <- glm(Snails ~ Location_m + Side,
                family = quasipoisson,
                data = mobile)
```

```
summary(fit_pois)
```

```
R>
R> Call:
R> glm(formula = Snails ~ Location_m + Side, family = poisson, data = mobile)
R>
R> Coefficients:
R>             Estimate Std. Error z value Pr(>|z|)
R> (Intercept)  1.801789    0.141224   12.76  <2e-16 ***
R> Location_m   -0.249638    0.009117  -27.38  <2e-16 ***
R> SideRight    2.757775    0.144401   19.10  <2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> (Dispersion parameter for poisson family taken to be 1)
R>
R> Null deviance: 6117.7 on 112 degrees of freedom
R> Residual deviance: 2921.5 on 110 degrees of freedom
R> (7 observations deleted due to missingness)
R> AIC: 3043.1
R>
R> Number of Fisher Scoring iterations: 8
```

```
summary(fit_qpois)
```

```
R>
R> Call:
R> glm(formula = Snails ~ Location_m + Side, family = quasipoisson,
R>      data = mobile)
R>
R> Coefficients:
R>             Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  1.8018      6.2655    0.288  0.774
R> Location_m   -0.2496      0.4045   -0.617  0.538
R> SideRight    2.7578      6.4065    0.430  0.668
R>
R> (Dispersion parameter for quasipoisson family taken to be 1968.325)
R>
R> Null deviance: 6117.7 on 112 degrees of freedom
R> Residual deviance: 2921.5 on 110 degrees of freedom
R> (7 observations deleted due to missingness)
R> AIC: NA
R>
R> Number of Fisher Scoring iterations: 8
```

20.8.5 Test Assumptions / Check Diagnostics

```
dispersion <- sum(residuals(fit_pois, type = "pearson")^2) / df.residual(fit_pois)
dispersion

R> [1] 1968.293

new_count <- expand.grid(
  Location_m = seq(min(mobile$Location_m), max(mobile$Location_m), length.out = 200),
  Side = levels(mobile$Side)
) >
as_tibble() >
mutate(mu = predict(fit_qpois, newdata = cur_data(), type = "response"))

ggplot(mobile, aes(x = Location_m, y = Snails, colour = Side)) +
```

```
geom_point(alpha = 0.35) +
geom_line(data = new_count, aes(y = mu), linewidth = 0.9) +
labs(
  x = "Location on shore gradient (m)",
  y = "Snail count"
)
```

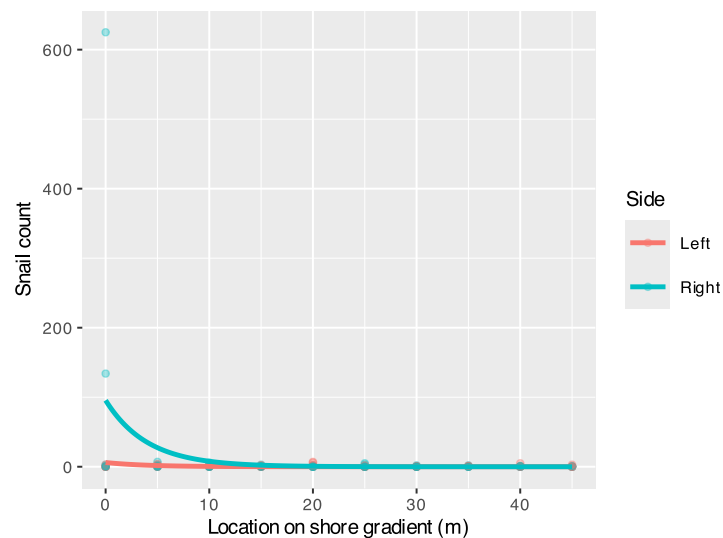


FIGURE 20.4. Observed snail counts and fitted quasi-Poisson means across the shore-position gradient.

The fitted curves in Figure 20.4 and the dispersion value of 1968.3 show that the simple Poisson model is much too optimistic. That is exactly the situation in which overdispersion must be acknowledged before interpreting the coefficients.

20.8.6 Interpret the Results

The Poisson model suggests that snail counts decline strongly along the shore gradient and differ between sides, but the overdispersion is so extreme that the nominal Poisson standard errors cannot be trusted.

The quasi-Poisson fit keeps the same mean structure but inflates the uncertainty appropriately. Under that more cautious model, the estimated decline with location and the side effect are no longer supported strongly. The biological conclusion therefore changes once overdispersion is taken seriously.

20.8.7 Reporting

WRITE-UP

Methods

Snail counts from rocky intertidal samples were analysed using a count-data GLM with location along the shore gradient and platform side as predictors. A Poisson model was fitted first, and overdispersion was then assessed using the Pearson residual dispersion statistic. Because overdispersion was extreme, a quasi-Poisson model was used for interpretation.

Results

The raw count data were highly variable, with the variance far exceeding the mean. The fitted Poisson model showed strong negative and side effects, but the dispersion statistic was approximately 1968.3, indicating severe overdispersion. Under the quasi-Poisson model, the same mean pattern remained but the standard errors were much larger, and neither predictor provided strong evidence for a clear effect ($p > 0.05$).

Discussion

The important lesson is not only that snail counts vary along the shore. It is that count-data inference can become misleading if overdispersion is ignored. In practice, the first biologically plausible pattern should always be checked against the adequacy of the assumed count distribution.

20.9 WHAT TO DO WHEN ASSUMPTIONS FAIL / ALTERNATIVES

- If a binary model shows dependence among observations, move to a generalised mixed model rather than forcing independence.
- If a count model is overdispersed, consider quasi-Poisson, negative binomial, or other more appropriate count models.
- If the response is a simple success proportion with no predictors, a proportion test may still be enough.
- If the link-scale coefficients are hard to interpret, translate the results back to probabilities or expected counts on the response scale.

20.10 SUMMARY

- GLMs extend regression to binary, proportional, and count responses.
- The family and link function must match the response structure.
- Simple proportion tests are part of the same broader modelling logic as binomial GLMs.

- In the acorn-barnacle example, the probability of presence declined across the shore gradient.
- In the snail-count example, severe overdispersion changed the inferential conclusion and showed why diagnostics matter in GLMs.

21

Generalised Additive Models

Flexible Smoother for Complex Relationships

IN THIS CHAPTER

- what a GAM is and why smooth terms are useful in ecology;
- how GAMs differ from polynomial and mechanistic nonlinear models;
- how to fit and inspect a Gaussian GAM with `mgcv::gam()`;
- how to interpret smooth terms, effective degrees of freedom, and model output;
- how to report a GAM in journal style without overclaiming mechanism.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Generalised additive models (GAMs) are often the most practical choice when biological relationships are clearly nonlinear, but the exact shape is unknown. Instead of forcing one global equation, GAMs estimate smooth functions directly from the data while retaining a clear regression structure.

This makes GAMs especially useful in ecology, where responses to environmental gradients are often curved, seasonal, and multi-scale. They are more flexible than low-order polynomials, but less explicitly mechanistic than a dedicated nonlinear process model.

21.1 KEY CONCEPTS

- **A GAM replaces straight-line terms with smooth functions** such as $s(x)$.
- **Smoothness is penalised** to avoid overfitting; wiggleness is not free.
- **Effective degrees of freedom (edf)** quantify the complexity of each smooth.
- **Inference shifts from individual coefficients to smooth-term significance and shape.**
- **Interpretation remains ecological:** the smooth tells you how the response changes across the predictor gradient.

21.2 WHEN THIS METHOD IS APPROPRIATE

Use a GAM when:

- the response-predictor relationship is nonlinear and not well captured by linear or quadratic forms;
- you do not have a single mechanistic process equation to impose;
- you need flexible trend estimation (for example seasonal or long-term environmental patterns);
- the sample size is sufficient to estimate smooths responsibly.

21.3 NATURE OF THE DATA AND ASSUMPTIONS

For Gaussian GAMs, assumptions are conceptually familiar:

1. independent observations;
2. approximately normal residuals;
3. reasonably constant residual variance;
4. appropriate smooth complexity (not too rigid, not too wiggly).

21.4 THE CORE EQUATIONS

For a Gaussian response, a GAM can be written as:

$$Y_i = \alpha + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_p(X_{ip}) + \epsilon_i \quad (21.1)$$

In Equation 21.1, each f_j is a smooth function estimated from the data rather than a straight-line coefficient multiplying the predictor directly. This is the key structural change relative to an ordinary linear model.

For introductory purposes, the important idea is that a GAM still has an additive regression structure. What changes is that the effect of a predictor is allowed to bend smoothly instead of being forced into a straight line or a fixed low-order polynomial.

21.5 R FUNCTIONS

The standard implementation is `mgcv::gam()`.

```
gam(y ~ s(x), data = df, method = "REML")
gam(y ~ s(x1) + s(x2), data = df, method = "REML")
gam(y ~ s(x, bs = "cc"), data = df, method = "REML") # cyclic smooth
```

`method = "REML"` is a widely-used default for smoothness selection.

21.6 EXAMPLE 1: SEA-TEMPERATURE STRUCTURE THROUGH TIME

21.6.1 Example Dataset

We use monthly sea surface temperature records from `data/BCB744/SACTN_day_1.csv` for Port Nolloth on the South African west coast. This is a realistic ecological time series where both long-term structure and within-year seasonality may matter.

```
temp_raw <- read_csv(here::here("data", "BCB744", "SACTN_day_1.csv"), show_col_types = FALSE)

temp_df <- temp_raw |>
  filter(site == "Port Nolloth", !is.na(temp)) |>
  mutate(
    date = as.Date(date),
    year = as.numeric(format(date, "%Y")),
    month = as.numeric(format(date, "%m")),
    t_index = as.numeric(date - min(date)) / 365.25
  )

gt(head(temp_df, 10) |> select(site, date, temp, year, month))
```

21.6.2 Do an Exploratory Data Analysis (EDA)

```
temp_df |>
  summarise(
    n = n(),
    start = min(date),
    end = max(date),
    mean_temp = mean(temp),
    sd_temp = sd(temp)
  )
```

```
R> # A tibble: 1 x 5
R>   n start      end      mean_temp sd_temp
R>   <int> <date>   <date>         <dbl>   <dbl>
R> 1  510 1973-07-01 2016-08-01         12.5    0.991
```

TABLE 21.1. A subset of the Port Nolloth temperature time series used in the GAM example.

site	date	temp	year	month
Port Nolloth	1277	11.722	1973	7
Port Nolloth	1308	11.534	1973	8
Port Nolloth	1339	10.879	1973	9
Port Nolloth	1369	11.786	1973	10
Port Nolloth	1400	12.308	1973	11
Port Nolloth	1430	12.340	1973	12
Port Nolloth	1461	11.538	1974	1
Port Nolloth	1492	12.105	1974	2
Port Nolloth	1520	11.971	1974	3
Port Nolloth	1551	12.462	1974	4

```
ggplot(temp_df, aes(x = date, y = temp)) +
  geom_line(alpha = 0.7) +
  labs(x = "Date", y = "Temperature (°C)")
```

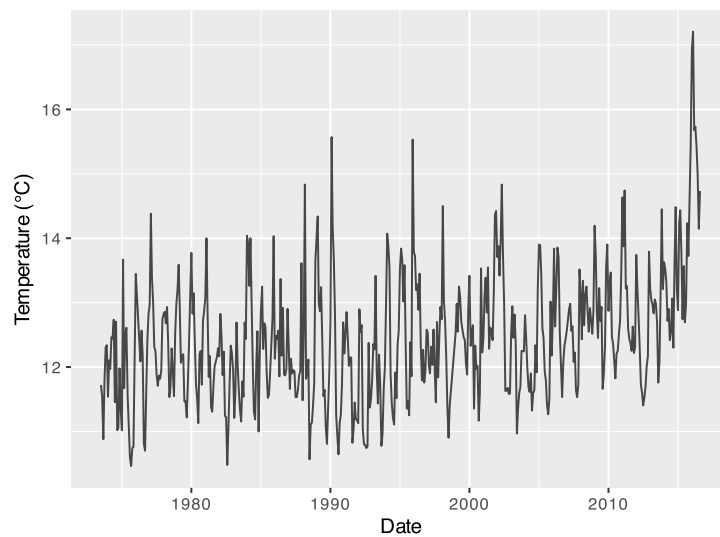


FIGURE 21.1. Monthly sea temperature at Port Nolloth.

```
ggplot(temp_df, aes(x = month, y = temp)) +
  geom_boxplot(fill = "grey80", colour = "grey30") +
  scale_x_continuous(breaks = 1:12) +
  labs(x = "Month", y = "Temperature (°C)")
```

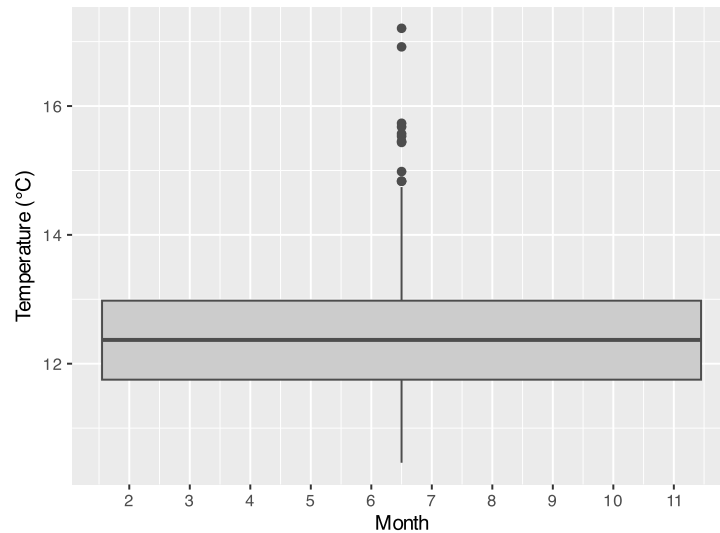


FIGURE 21.2. Seasonal pattern in monthly temperature (across years).

The data in Figure 21.1 and Figure 21.2 show clear nonlinearity through time and a strong seasonal cycle. A straight-line model in time is therefore likely to be inadequate.

21.6.3 State the Model Question and Hypotheses

Can sea temperature at Port Nolloth be explained by a flexible long-term trend plus a seasonal smooth cycle?

For smooth terms in GAMs, hypotheses are usually phrased as:

$$H_0 : f_j(X) = 0$$

$$H_a : f_j(X) \neq 0$$

for each smooth term f_j . In practice, we inspect smooth-term significance and shape together.

21.6.4 Fit the Model

We fit a baseline linear model and then a GAM with:

- a smooth long-term trend in continuous time ($s(t\_index)$), and
- a cyclic smooth for month ($s(month, bs = "cc")$) so December and January join naturally.

```
mod_lm <- lm(temp ~ t_index + factor(month), data = temp_df)
```

```
mod_gam <- gam(
  temp ~ s(t_index, k = 20) + s(month, bs = "cc", k = 12),
  data = temp_df,
  method = "REML"
)
```

```
summary(mod_lm)
```

```
R>
R> Call:
R> lm(formula = temp ~ t_index + factor(month), data = temp_df)
R>
R> Residuals:
R>      Min       1Q   Median       3Q      Max
R> -1.7770 -0.5217 -0.0540  0.4275  3.3436
R>
R> Coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)   12.370393   0.131398   94.144 < 2e-16 ***
R> t_index         0.028353   0.002738   10.354 < 2e-16 ***
R> factor(month)2    0.342733   0.166124    2.063 0.039621 *
R> factor(month)3    0.029053   0.167118    0.174 0.862058
R> factor(month)4   -0.170806   0.167121   -1.022 0.307254
R> factor(month)5   -0.415614   0.166126   -2.502 0.012677 *
R> factor(month)6   -0.722765   0.166128   -4.351 1.65e-05 ***
R> factor(month)7   -1.117334   0.166129   -6.726 4.82e-11 ***
R> factor(month)8   -1.206711   0.165178   -7.306 1.11e-12 ***
R> factor(month)9   -1.259495   0.167114   -7.537 2.30e-13 ***
R> factor(month)10  -1.020191   0.168138   -6.068 2.58e-09 ***
R> factor(month)11  -0.617730   0.167111   -3.697 0.000243 ***
R> factor(month)12  -0.133916   0.167111   -0.801 0.423306
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 0.7703 on 497 degrees of freedom
R> Multiple R-squared:  0.4098, Adjusted R-squared:  0.3955
R> F-statistic: 28.76 on 12 and 497 DF, p-value: < 2.2e-16
```

```
summary(mod_gam)
```

```
R>
R> Family: gaussian
R> Link function: identity
R>
R> Formula:
R> temp ~ s(t_index, k = 20) + s(month, bs = "cc", k = 12)
R>
R> Parametric coefficients:
R>              Estimate Std. Error t value Pr(>|t|)
R> (Intercept)  12.45935    0.02862   435.3 <2e-16 ***
```

```

R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Approximate significance of smooth terms:
R>      edf Ref.df      F p-value
R> s(t_index) 15.88  17.85 20.14 <2e-16 ***
R> s(month)    5.61  10.00 32.98 <2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> R-sq.(adj) =  0.574   Deviance explained = 59.2%
R> -REML = 542.09   Scale est. = 0.41773    n = 510
AIC(mod_lm, mod_gam)

```

```

par(mfrow = c(2, 2))
gam.check(mod_gam)

```

```

R>      df      AIC
R> mod_lm 14.00000 1195.935
R> mod_gam 24.37806 1027.884

```

21.6.5 Test Assumptions / Check Diagnostics

```

R>
R> Method: REML   Optimizer: outer newton
R> full convergence after 7 iterations.
R> Gradient range [-3.460542e-09,3.422151e-12]
R> (score 542.0866 & scale 0.4177274).
R> Hessian positive definite, eigenvalue range [2.683803,254.2529].
R> Model rank =  30 / 30
R>
R> Basis dimension (k) checking results. Low p-value (k-index<1) may
R> indicate that k is too low, especially if edf is close to k'.
R>
R>      k'   edf k-index p-value
R> s(t_index) 19.00 15.88    0.63 <2e-16 ***
R> s(month)   10.00  5.61    1.06    0.93
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
par(mfrow = c(1, 1))

```

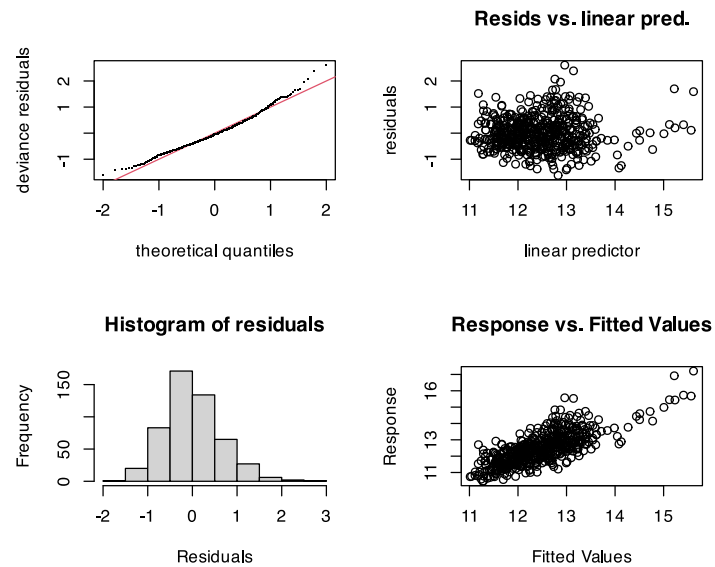


FIGURE 21.3. Standard mgcv diagnostic checks for the fitted GAM.

```
concurvity(mod_gam, full = TRUE)
```

```
R>           para  s(t_index)  s(month)
R> worst    3.927862e-24 0.0057595704 0.005759570
R> observed 3.927862e-24 0.0016454850 0.001803124
R> estimate 3.927862e-24 0.0007068493 0.001149347
```

`gam.check()` and the plots in Figure 21.3 help assess residual behaviour and whether the chosen basis dimensions (k) are adequate.

```
par(mfrow = c(1, 2))
plot(mod_gam, shade = TRUE, pages = 1)
par(mfrow = c(1, 1))
```

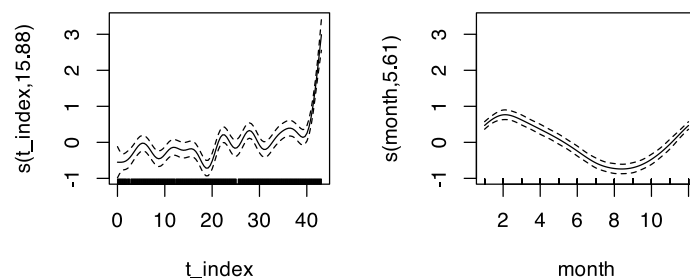



FIGURE 21.4. Estimated smooth terms for the GAM: long-term trend and cyclic monthly seasonality.

```
temp_df <- temp_df > mutate(gam_fit = fitted(mod_gam))
ggplot(temp_df, aes(x = date)) +
  geom_point(aes(y = temp), alpha = 0.35, size = 0.8) +
  geom_line(aes(y = gam_fit), colour = "red", linewidth = 0.3) +
  labs(x = "Date", y = "Temperature (°C)")
```

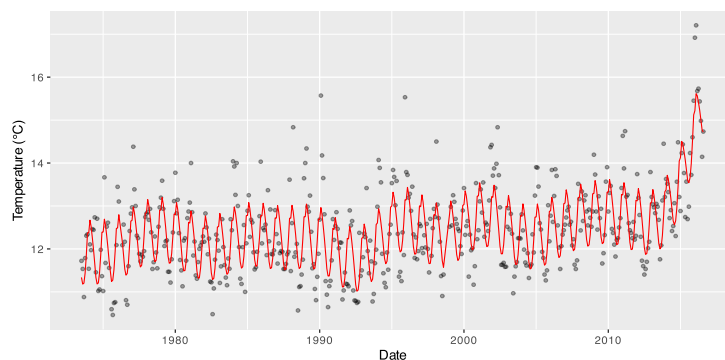


FIGURE 21.5. Observed temperature and fitted GAM values through time.

The smooths in Figure 21.4 show that t_index captures gradual long-term structure that a single linear slope cannot represent. The cyclic month smooth captures recurring within-year seasonality without forcing identical month effects in each year. The fitted values in Figure 21.5 show how those smooth components reconstruct the observed pattern through time.

The effective degrees of freedom (edf) indicate complexity: edf values close to 1 imply near-linearity; higher values indicate more curvature.

21.6.7 Reporting

WRITE-UP

Methods

Monthly sea surface temperature observations for Port Nolloth were analysed using a Gaussian GAM fitted with `mgcv::gam()` in R. Temperature was modelled as a function of a smooth long-term time index (`s(t_index)`) and a cyclic seasonal smooth for month (`s(month, bs = "cc")`), with smoothness selected by REML. A linear model with month as a factor and a linear time term was used as a baseline comparator.

Results

The GAM provided a better fit than the baseline linear model (lower AIC), indicating that nonlinear structure was important. The long-term smooth term was non-zero and captured gradual multi-year variability, while the cyclic month smooth described strong seasonal temperature cycling. Together, these terms reproduced the major temporal structure in observed temperatures without requiring a fixed parametric curve.

Discussion

For this ecological time series, a GAM was appropriate because both long-term and seasonal effects were clearly nonlinear. The model is best interpreted as a flexible description of temporal structure, not as a mechanistic oceanographic process model. Where mechanism is the primary goal, process-based nonlinear models should be considered alongside GAMs.

21.7 WHAT TO DO WHEN ASSUMPTIONS FAIL / ALTERNATIVES

- If residuals show strong autocorrelation, move to GAMM frameworks (e.g., correlation structures or random effects).
- If response variance is clearly non-Gaussian, use an appropriate family (e.g., Poisson, negative binomial, binomial).
- If smooths are implausibly wiggly, reduce basis size (k) and inspect diagnostics carefully.

21.8 SUMMARY

- GAMs are additive smooth regression models that handle complex nonlinear ecological relationships.
- They are often superior to low-order polynomials when shape is unknown.

- Interpretation focuses on smooth shapes and ecological plausibility rather than mechanistic parameters.
- Diagnostics and smoothness control are essential for responsible use.

22

Nonlinear Regression

When Straight Lines Are Not Enough

IN THIS CHAPTER

- Why some biological relationships are mechanistically nonlinear
- How nonlinear regression differs from polynomial and additive models
- The logic of named nonlinear functions
- A worked Michaelis-Menten example
- How nonlinear mixed-effects models extend mechanistic curves to repeated-measures data
- What to check when fitting nonlinear models

Some relationships are curved because the biology is curved. Growth saturates, enzyme kinetics asymptote, seasonal cycles oscillate, and responses may level off, bend, or change slope across the range of the predictor. In such cases, a straight-line model is often too rigid.

In this chapter, I focus specifically on mechanistic nonlinear regression: models where the response is described by a named nonlinear function whose parameters have scientific meaning. Examples include saturating uptake curves, logistic growth, seasonal sine curves, and other biologically motivated responses.

Polynomial regression and GAMs are covered in their own chapters because they answer a different modelling need. Here the emphasis is narrower and stronger: what to do when the biology suggests the form of the curve itself.

22.1 KEY CONCEPTS

- **Nonlinearity** often reflects real biological structure rather than unwanted curvature.
- **Flexible regression** can model patterns that a straight line misses.
- **Mechanistic nonlinear models are appropriate when the curve has a biologically motivated form.**
- **Greater flexibility increases the risk of overfitting** if used carelessly.
- **Model choice should follow biological form**, not only convenience or visual preference.

22.2 WHEN THIS METHOD IS APPROPRIATE

You should move beyond a simple linear model when:

- the fitted residuals show clear curvature;
- the biology suggests an asymptote, threshold, saturation, or oscillation;
- the response changes differently across the predictor range;
- the question concerns parts of the response distribution other than the mean.

The diagnostic chapters earlier in the sequence already suggested this possibility. This chapter is about what to do next.

22.3 NATURE OF THE DATA AND ASSUMPTIONS

Mechanistic nonlinear regression is still regression, so the familiar concerns about independence, residual spread, and overall model adequacy do not disappear. In many introductory applications, the response variable is continuous and the residuals are assumed to be approximately normally distributed with constant variance.

The key distinction is that the mean structure is explicitly nonlinear and tied to a named function. These models are usually fitted by nonlinear least squares with `nls()`, or by mixed-effects extensions when grouped or repeated-measures data are present.

The practical point is that you choose a non-linear model because the straight-line model is inadequate and because the curved alternative matches the biological process or the inferential question more closely. And it looks pretty and complicated.

22.4 THE CORE EQUATION

The general form of a nonlinear regression is:

$$Y_i = f(X_i, \theta) + \epsilon_i \quad (22.1)$$

In Equation 22.1, $f(\cdot)$ is a biologically chosen nonlinear function and θ represents the parameter set that defines its shape. This chapter then uses a Michaelis-Menten curve as the main worked example of that general idea.

22.5 R FUNCTIONS

Some of the most useful R functions are:

- `nls()` for nonlinear least-squares models;
- `nlme::nlme()` when nonlinear relationships occur in grouped or repeated-measures data.

Examples:

```
lm(response ~ poly(x, degree = 2), data = df)
nls(y ~ Vmax * x / (Km + x), data = df,
    start = list(Vmax = 1, Km = 1))
nlme::nlme(y ~ Vmax * x / (Km + x), data = df, ...)
```

One practical difference from ordinary regression is that `nls()` usually needs sensible starting values. Unlike `lm()`, it uses an iterative search for the best-fitting parameter values, and poor starting values can prevent convergence.

22.6 EXAMPLE 1: WHY USE A MECHANISTIC NONLINEAR MODEL?

The first decision is whether the biology gives you a reason to prefer a specific nonlinear function rather than a descriptive curve.

Figure 22.1 shows three broad situations:

- A polynomial model can describe modest curvature, but it is mainly descriptive.
- A GAM is very useful when the relationship is clearly curved but its exact form is unknown or not informative.
- A mechanistic nonlinear model is most attractive when the biology suggests a specific function, because its parameters can often be interpreted directly.

That last point is important. If we know the process should saturate, as in nutrient uptake kinetics, a mechanistic model tells us more than a flexible smoother. It gives us interpretable parameters such as an asymptote or a half-saturation constant.

22.7 EXAMPLE 2: ALGAL NUTRIENT UPTAKE KINETICS

We can measure algal nutrient uptake rates using a multiple flask experiment. We prepare a series of flasks, each containing a different initial concentration of the

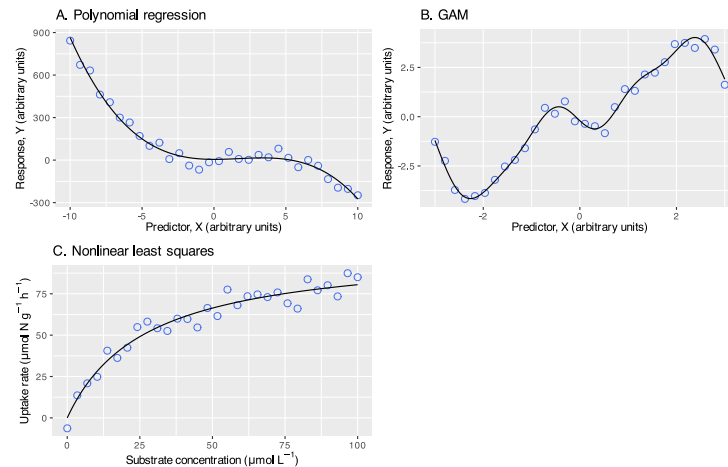


FIGURE 22.1. Examples of curved models fitted to simulated data. A polynomial can capture modest curvature, a GAM can adapt flexibly to the data, and a mechanistic nonlinear model can represent a known saturating biological response.

substrate nutrient, and then estimate nutrient uptake rate over a fixed time interval. The result is a set of uptake rates, V ($\mu\text{mol N g}^{-1} \text{h}^{-1}$), paired with substrate concentrations, $[S]$ ($\mu\text{mol L}^{-1}$).

If you need a reminder of the biological and experimental context behind these data, refer back to the BDC223 nutrient-uptake material in [Lecture 8a: Nutrient Uptake](#), [Lecture 8b: Uptake Kinetics – Michaelis-Menten](#), and [Lab 4: Uptake Kinetics – Michaelis-Menten](#). Those pages explain the experimental principles of the multiple-flask design in much more detail than is needed here.

Applied to algae, the Michaelis-Menten model assumes an irreversible uptake process that saturates at high substrate concentrations. It effectively quantifies key characteristics of the nutrient uptake system, including the maximum uptake rate and the alga's affinity for the nutrient:

$$V_i = \frac{V_{max} \cdot [S_i]}{K_m + [S_i]} + \epsilon_i \quad (22.2)$$

In Equation 22.2:

- V_i is the uptake rate at the i -th observation ($\mu\text{mol N g}^{-1} \text{h}^{-1}$);
- V_{max} is the maximum nutrient uptake rate achieved ($\mu\text{mol N g}^{-1} \text{h}^{-1}$);
- $[S_i]$ is the substrate concentration at the i -th observation ($\mu\text{mol L}^{-1}$);

- K_m is the substrate concentration at which uptake reaches half of V_{max} ($\mu\text{mol L}^{-1}$); and
- ϵ_i is the residual error.

The two parameters are biologically meaningful. V_{max} represents the maximum capacity of the alga to utilise the nutrient and is measured in $\mu\text{mol N g}^{-1} \text{h}^{-1}$. K_m describes the affinity of the uptake system for that nutrient and is measured in $\mu\text{mol L}^{-1}$. Lower values of K_m indicate higher affinity.

22.7.1 Do an Exploratory Data Analysis (EDA)

To demonstrate fitting a nonlinear model to V versus $[S]$ data from a multiple flask experiment, I simulate data across a range of substrate concentrations. The dataset consists of five replicate flasks for each of 13 substrate concentrations.

```
conc_vec <- c(0, 0.1, 0.5, 2, 5, 7.5, 10, 12.5, 15, 17.5, 20, 25, 30)
n_rep <- 5
Km_vec <- c(10)
Vmax_vec <- c(50)
Km_vec_sd <- c(1.2)
Vmax_vec_sd <- c(0.7)

mf_data <- generate_data(
  n_trt = 1,
  n_rep = n_rep,
  conc_vec = conc_vec,
  Km_vec = Km_vec,
  Vmax_vec = Vmax_vec,
  Km_vec_sd = Km_vec_sd,
  Vmax_vec_sd = Vmax_vec_sd
) >
select(rep, S, V)

mf_data >
slice(c(1:4, (n() - 3):n())) >
gt::gt() >
gt::fmt_number(columns = c(S, V), decimals = 2)
```

The plot in Figure 22.2 already suggests that a straight line is not the right model. Uptake increases quickly away from low substrate concentrations and then begins to level off when the concentrations become higher. That is exactly the sort of pattern for which a Michaelis-Menten model was designed.

You will also often see greater spread in uptake rate at higher substrate concentrations. That happens because the high-concentration part of the curve approaches the enzyme-limited region. Once transport and assimilation systems are near saturation, small biological differences among flasks in tissue condition, enzyme capacity, light history, or temperature response can translate into visibly different realised uptake rates. At low concentrations, by contrast, uptake is constrained more strongly by substrate availability itself, so the responses are often pulled into a narrower range.

TABLE 22.1. Preview of the simulated multiple-flask dataset (first and last four rows).

rep	S	V
1	0.00	0.00
2	0.00	0.00
3	0.00	0.00
4	0.00	0.00
2	30.00	35.12
3	30.00	39.07
4	30.00	37.78
5	30.00	36.73

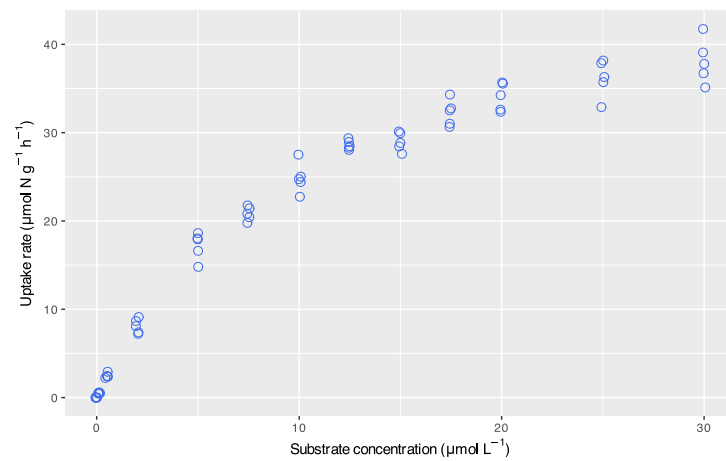


FIGURE 22.2. Plot of uptake rate, V , as a function of substrate concentration, $[S]$, for a multiple flask experiment. The relationship is clearly curved and tends towards saturation.

22.7.2 State the Model Question

At this point, two related questions arise.

The first is descriptive and mechanistic: can we estimate the uptake parameters V_{max} ($\mu\text{mol N g}^{-1} \text{ h}^{-1}$) and K_m ($\mu\text{mol L}^{-1}$) from these data?

The second is comparative: does a Michaelis-Menten model fit these data better than a simple linear model?

Those are not the same question. The first is about parameter estimation within a nonlinear biological model. The second is about choosing between two competing mean structures.

22.7.3 Fit the Model

The Michaelis-Menten model is fit with `nls()`:

```
nls_mod <- nls(
  V ~ mm_fun(S, Vmax, Km),
  data = mf_data,
  start = c(Vmax = 30, Km = 5)
)

summary(nls_mod)

R>
R> Formula: V ~ mm_fun(S, Vmax, Km)
R>
R> Parameters:
R>      Estimate Std. Error t value Pr(>|t|)
R> Vmax  51.4417    1.1770   43.70  <2e-16 ***
R> Km    10.5507     0.5975   17.66  <2e-16 ***
R> ---
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R>
R> Residual standard error: 1.312 on 63 degrees of freedom
R>
R> Number of iterations to convergence: 3
R> Achieved convergence tolerance: 8.864e-06
```

The starting values do not have to be exact, but they should be sensible. Here we know from the plot that the asymptote is probably somewhere around 40 to 50 $\mu\text{mol N g}^{-1} \text{ h}^{-1}$, and that the half-saturation constant is likely to be in the single digits of $\mu\text{mol L}^{-1}$.

22.7.4 Check Diagnostics

Because this is a least-squares nonlinear model, residual diagnostics are interpreted much as they are for ordinary regression.

```
mf_data <- mf_data %>%
  mutate(
    fitted = fitted(nls_mod),
```

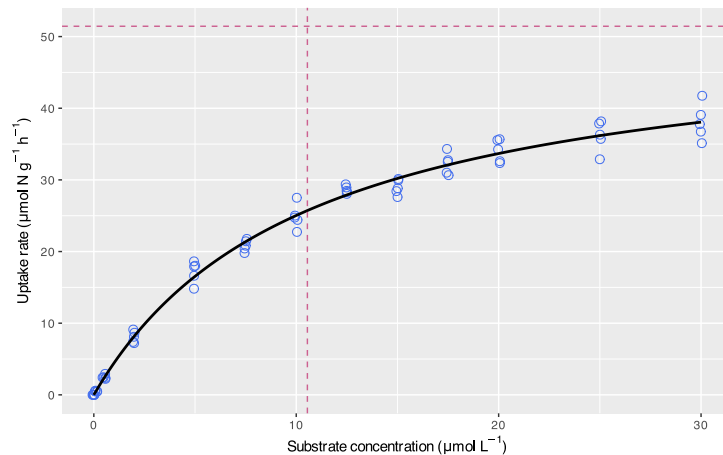


FIGURE 22.3. Michaelis-Menten model fitted to the uptake data. The dashed lines indicate the estimated K_m and V_{max} values.

```
resid = resid(nls_mod)
)
shapiro.test(mf_data$resid)

R>
R> Shapiro-Wilk normality test
R>
R> data: mf_data$resid
R> W = 0.97225, p-value = 0.1517
```

The diagnostics in Figure 22.4 do not suggest any serious problem. The residual spread is fairly even, and the histogram is approximately symmetric. As always, you should not rely on a single formal test, but the residual behaviour is consistent with a usable fit.

22.7.5 Interpret the Results

```
lm_mod <- lm(V ~ S, data = mf_data)
model_comp <- anova(lm_mod, nls_mod)
f_stat <- model_comp$F[2]

AIC(lm_mod, nls_mod)

R>      df      AIC
R> lm_mod  3 395.8491
R> nls_mod  3 223.6862
```

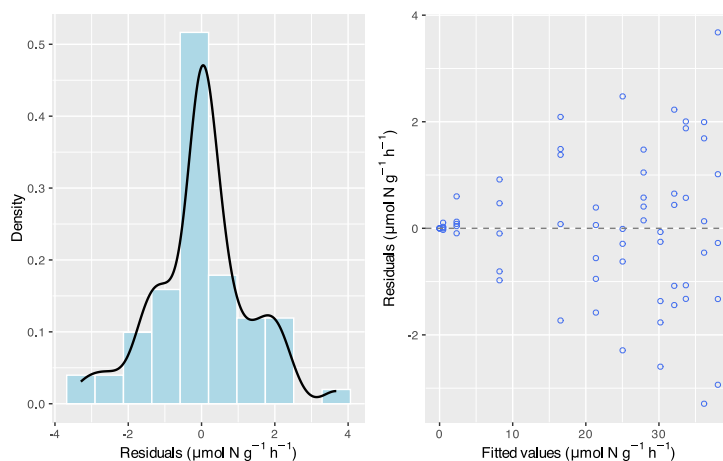


FIGURE 22.4. Basic diagnostics for the Michaelis-Menten fit. The residuals show no serious trend with the fitted values, and the distribution is approximately symmetric.

```
BIC(lm_mod, nls_mod)
```

```
R>      df      BIC
R> lm_mod 3 402.3723
R> nls_mod 3 230.2094
```

```
model_comp
```

```
R> Analysis of Variance Table
```

```
R>
```

```
R> Response: V
```

```
R>      Df Sum Sq Mean Sq F value    Pr(>F)
R> S      1 10274.8  10274.8   422.59 < 2.2e-16 ***
R> Residuals 63  1531.8     24.3
```

```
R> ---
```

```
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The fitted nonlinear model gives estimated values for V_{max} and K_m , and these are the biologically meaningful outputs of the analysis. The estimated V_{max} is the asymptotic uptake rate in $\mu\text{mol N g}^{-1} \text{h}^{-1}$, while K_m indicates the substrate concentration in $\mu\text{mol L}^{-1}$ at which the uptake rate reaches half of that asymptote.

The model comparison also matters. If the linear and Michaelis-Menten models are both fitted to the same data, the nonlinear model has markedly lower information criteria and a much smaller residual error. That supports the biological interpretation already suggested by the plot; that is, uptake increases with substrate concentration, but it does so in a saturating rather than indefinitely linear way.

22.7.6 Reporting

WRITE-UP

Methods

Nutrient uptake rate was modelled as a function of substrate concentration with a Michaelis-Menten nonlinear regression fitted by nonlinear least squares. The nonlinear model was compared with a simple linear regression using information criteria and nested model comparison to assess whether the saturating form provided a better description of the data.

Results

Nutrient uptake increased rapidly with substrate concentration at low concentrations and then approached an asymptote, consistent with Michaelis-Menten kinetics (Figure 22.3). The fitted nonlinear model estimated a maximum uptake rate of $51.44 \mu\text{mol N g}^{-1} \text{h}^{-1}$ and a half-saturation constant of $10.55 \mu\text{mol L}^{-1}$. Relative to a simple linear regression, the Michaelis-Menten model provided the better description of the data, with lower AIC and BIC values and a significant improvement in fit ($F = \text{NA}$, $p < 0.05$).

Discussion

The main biological point is that uptake is saturating rather than indefinitely linear over the observed concentration range. That matters because the fitted parameters now speak directly to uptake capacity and nutrient affinity.

22.7.7 What If We Add Treatments?

Experiments are seldom as simple as the one above. Consider an experiment designed to assess whether an experimental treatment, such as light intensity or seawater temperature, affects nutrient uptake kinetics. It is biologically plausible to expect that each treatment will result in unique $V_{\max}$ values ($\mu\text{mol N g}^{-1} \text{h}^{-1}$) and K_m values ($\mu\text{mol L}^{-1}$).

```
mf_data2 <- generate_data(
  n_trt = 3,
  n_rep = 5,
  conc_vec = conc_vec,
  Km_vec = c(10, 8, 6),
  Vmax_vec = c(50, 40, 22),
  Km_vec_sd = c(1.2, 1.1, 1.0),
  Vmax_vec_sd = c(0.7, 1.0, 0.8)
)
```

The treatment-specific curvature in Figure 22.5 makes the modelling problem more challenging and exciting. A separate `nls()` fit can be used for each treatment, or a nonlinear mixed-effects model can be used when dependence or grouping must be handled explicitly. That is the nonlinear counterpart of the fixed-effects and

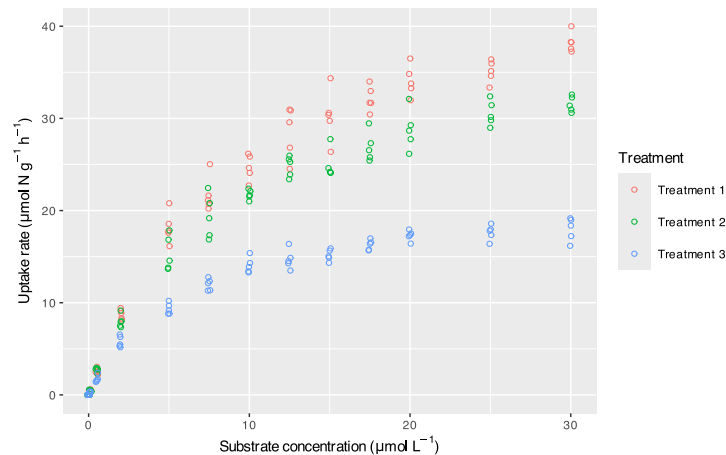


FIGURE 22.5. Simulated nutrient uptake curves for three treatments. The pattern suggests treatment-specific nonlinear responses rather than a single shared curve.

mixed-effects logic introduced in [Chapter 19](#). If the response is also non-normal, you are moving towards the broader territory covered in [Chapter 20](#).

The important point is that nonlinear modelling “draws a curve” *and* it chooses a mean structure that matches the process and then extending that structure carefully when treatments, dependence, or more complicated designs are present.

22.7.8 Repeated-Measures Extension

The next extension appears when the same experimental unit is measured repeatedly across the concentration gradient. For example, you might measure uptake repeatedly on the same thallus, plant, or mesocosm under several substrate concentrations. At that point, the observations within each unit are no longer independent, so a separate `nls()` fit per unit or a pooled `nls()` fit is no longer enough.

The nonlinear mixed-effects version of the model keeps the Michaelis-Menten mean structure but lets one or more parameters vary among experimental units. A common starting point is to allow each unit to have its own $V_{\max}$ ($\mu\text{mol N g}^{-1} \text{h}^{-1}$) around a population-level mean:

```
nlsme::nlsme(
  V ~ mm_fun(S, Vmax, Km),
  data = uptake_long,
  fixed = Vmax + Km ~ treatment,
  random = Vmax ~ 1 | individual,
  start = c(...)
)
```

)

In that formulation, the fixed effects describe how treatment shifts the average kinetic parameters, while the random effect acknowledges that repeated observations from the same individual share the same underlying uptake curve. A full worked example needs a dataset with true repeated measurements on the same biological units and enough observations per unit to estimate the random-effects structure sensibly.

22.8 EXAMPLE 3: REPEATED-MEASURES MICHAELIS-MENTEN KINETICS

The previous examples treated each flask as an independent unit observed once at each starting concentration. A perturbation experiment is different. The same flask is sampled repeatedly through time as nutrient concentration declines, so the observations from that flask are correlated. This is exactly the kind of dependence problem introduced in [Chapter 19](#), except that the mean structure is now nonlinear rather than linear.

For the BDC223 treatment of this experimental design, including depletion curves and the distinction between perturbation and multiple-flask methods, see [Lecture 8a: Nutrient Uptake](#), [Lecture 8b: Uptake Kinetics – Michaelis-Menten](#), and [Lab 4: Uptake Kinetics – Michaelis-Menten](#). Here we pick up at the point where those experiments become a modelling problem.

22.8.1 Example Dataset

These data come from a perturbation experiment on the red seaweed *Gracilaria* sp. Flasks were enriched to approximately 55–60 μM nitrate and sampled repeatedly over roughly 2.5 hours. Uptake rates were measured under three levels of water movement (low, med, high), with three replicate flasks per treatment.

```
mm_data <- read.csv(here::here("data", "BCB744", "flow_rates.txt"), sep = ",") >
mutate(
  flask = factor(flask),
  trt = factor(trt, levels = c("low", "med", "high"))
)

head(mm_data, 8) >
gt::gt() >
gt::fmt_number(columns = c(V, S), decimals432 2)
```


TABLE 22.2. Preview of the perturbation dataset (first eight rows).

flask	trt	V	S
a1	low	10.80	60.20
a2	low	10.00	61.10
a3	low	14.10	60.80
a1	low	7.30	59.40
a2	low	15.40	59.40
a3	low	9.50	59.40
a1	low	12.50	56.70
a2	low	10.50	56.20

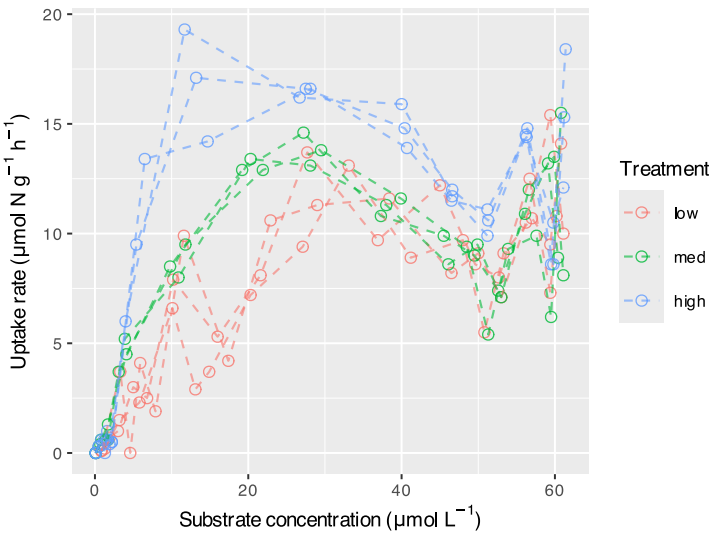


FIGURE 22.6. Repeated-measures uptake data from the perturbation experiment. Each dashed line traces one flask through time as substrate concentration declines.

22.8.2 Do an Exploratory Data Analysis (EDA)

The relationship in Figure 22.6 is again saturating, so a Michaelis-Menten mean structure is still sensible. The important difference is that each dashed trajectory comes from the same flask measured repeatedly. That means the residuals are likely to be correlated within flask, and the flasks are nested within treatment.

22.8.3 State the Model Question

The biological question is whether the Michaelis-Menten parameters differ among water-movement treatments once the repeated-measures structure of the perturbation experiment is taken into account. Here that means asking whether treatment shifts V_{max} ($\mu\text{mol N g}^{-1} \text{ h}^{-1}$) and/or K_m ($\mu\text{mol L}^{-1}$).

H_0 : V_{max} and K_m are the same across treatments after accounting for flask-level dependence

H_a : At least one of V_{max} or K_m differs among treatments after accounting for flask-level dependence

22.8.4 Fit the Model

We compare two nonlinear mixed-effects models. The first is a **global** model in which all treatments share the same Michaelis-Menten parameters. The second allows V_{max} and K_m to vary by treatment while retaining the same repeated-measures structure.

```
num_levels <- nlevels(mm_data$trt)

start_global <- c(
  Vmax = max(mm_data$V),
  Km = median(mm_data$S)
)

start_trt <- list(
  fixed = c(
    Vmax = rep(max(mm_data$V), num_levels),
    Km = rep(median(mm_data$S), num_levels)
  )
)

mm_global <- nlme(
  V ~ mm_fun(S, Vmax, Km),
  data = mm_data,
  fixed = Vmax + Km ~ 1,
  random = Vmax ~ 1 | trt/flask,
  groups = ~ trt/flask,
  correlation = corAR1(form = ~ 1 | trt/flask),
  start = start_global,
  method = "ML"
```

```

)

mm_trt <- nlme(
  V ~ mm_fun(S, Vmax, Km),
  data = mm_data,
  fixed = list(Vmax ~ trt, Km ~ trt),
  random = Vmax ~ 1 | trt/flask,
  groups = ~ trt/flask,
  correlation = corAR1(form = ~ 1 | trt/flask),
  start = start_trt,
  method = "ML"
)

mm_comp <- anova(mm_global, mm_trt)

mm_comp

R>      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
R> mm_global      1 6 656.0306 673.3274 -322.0153
R> mm_trt        2 10 648.1068 676.9348 -314.0534 1 vs 2 15.92387 0.0031

summary(mm_trt)

R> Nonlinear mixed-effects model fit by maximum likelihood
R> Model: V ~ mm_fun(S, Vmax, Km)
R> Data: mm_data
R>      AIC      BIC    logLik
R> 648.1068 676.9348 -314.0534
R>
R> Random effects:
R> Formula: Vmax ~ 1 | trt
R>      Vmax.(Intercept)
R> StdDev:      0.0001334694
R>
R> Formula: Vmax ~ 1 | flask %in% trt
R>      Vmax.(Intercept) Residual
R> StdDev:      2.572043e-05 2.641716
R>
R> Correlation Structure: AR(1)
R> Formula: ~1 | trt/flask
R> Parameter estimate(s):
R>      Phi
R> 0.1539479
R> Fixed effects: list(Vmax ~ trt, Km ~ trt)
R>      Value Std.Error DF   t-value p-value
R> Vmax.(Intercept) 13.727753 2.011411 118 6.824938 0.0000
R> Vmax.trtmed      -1.917486 2.239487 118 -0.856217 0.3936
R> Vmax.trthigh      1.623586 2.258494 118 0.718880 0.4736
R> Km.(Intercept)   16.847853 7.413292 118 2.272655 0.0249
R> Km.trtmed        -11.917481 7.783803 118 -1.531062 0.1284
R> Km.trthigh       -11.480850 7.623149 118 -1.506051 0.1347
R> Correlation:
R>      Vm.(I) Vmx.trtm Vmx.trth Km.(I) Km.trtm
R> Vmax.trtmed -0.898
R> Vmax.trthigh -0.891 0.800
R> Km.(Intercept) 0.923 -0.829 -0.822
R> Km.trtmed     -0.879 0.892 0.783 -0.952
R> Km.trthigh    -0.898 0.806 0.877 -0.972 0.926
R>
R> Standardized Within-Group Residuals:
R>      Min      Q1      Med      Q3      Max
R> -2.0770735 -0.7687046 -0.2436554 0.4550068 3.3221378

```

```
R>
R> Number of Observations: 132
R> Number of Groups:
R>          trt flask %in% trt
R>          3          9
```

Three parts of the specification are important:

- `fixed = ... ~ trt` allows the Michaelis-Menten parameters to differ among treatments.
- `random = Vmax ~ 1 | trt/flask` allows replicate flasks within treatments to vary in baseline uptake capacity.
- `correlation = corAR1( ... )` acknowledges that repeated measurements within the same flask are serially correlated.

This is the nonlinear counterpart of the grouped-dependence ideas from Chapter 19. The model still estimates a mechanistic curve, but it now does so in a way that respects repeated measurement on the same experimental unit while keeping the kinetic parameters in their usual units of $\mu\text{mol N g}^{-1} \text{h}^{-1}$ ($V_{\max}$) and $\mu\text{mol L}^{-1}$ (K_m).

22.8.5 Check Diagnostics

```
mm_data <- mm_data %>%
  mutate(
    fitted_nlme = fitted(mm_trt),
    resid_nlme = resid(mm_trt, type = "pearson")
  )

shapiro.test(mm_data$resid_nlme)
```

```
R>
R> Shapiro-Wilk normality test
R>
R> data: mm_data$resid_nlme
R> W = 0.97547, p-value = 0.01719
```

The diagnostics in Figure 22.7 show whether the fitted nonlinear mean structure and the residual assumptions are at least broadly compatible with the data.

22.8.6 Interpret the Results

```
coef_mm <- fixed.effects(mm_trt)

Vmax_by_trt <- c(
  low = coef_mm["Vmax.(Intercept)"],
  med = coef_mm["Vmax.(Intercept)"] + coef_mm["Vmax.trtmed"],
  high = coef_mm["Vmax.(Intercept)"] + coef_mm["Vmax.trthigh"]
)

Km_by_trt <- c(
  low = coef_mm["Km.(Intercept)"],
  med = coef_mm["Km.(Intercept)"] + coef_mm["Km.trtmed"],
  high = coef_mm["Km.(Intercept)"] + coef_mm["Km.trthigh"]
)
```

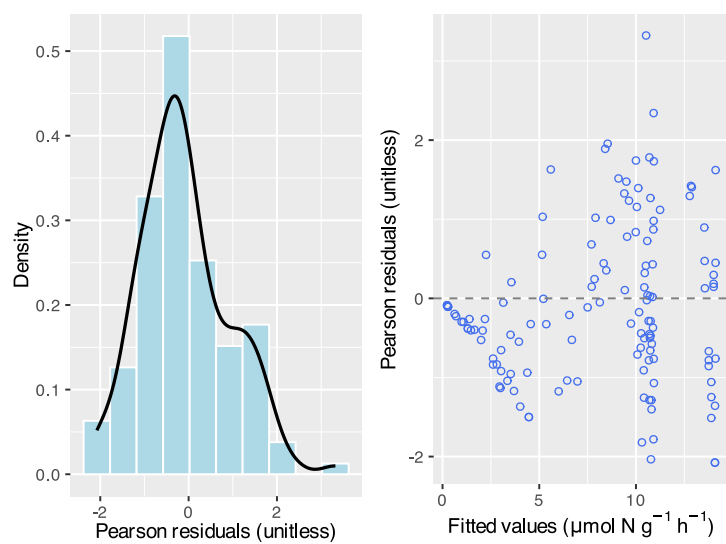


FIGURE 22.7. Basic diagnostics for the nonlinear mixed-effects model. The residual pattern is acceptable for a teaching example, with no strong trend against the fitted values.

```
)
param_tbl <- tibble(
  trt = factor(c("low", "med", "high"), levels = c("low", "med", "high")),
  Vmax = round(unname(Vmax_by_trt), 2),
  Km = round(unname(Km_by_trt), 2)
)
```

```
param_tbl
```

```
R> # A tibble: 3 x 3
R>   trt    Vmax    Km
R>   <fct> <dbl> <dbl>
R> 1 low    13.7  16.8
R> 2 med    11.8   4.93
R> 3 high   15.4   5.37
```

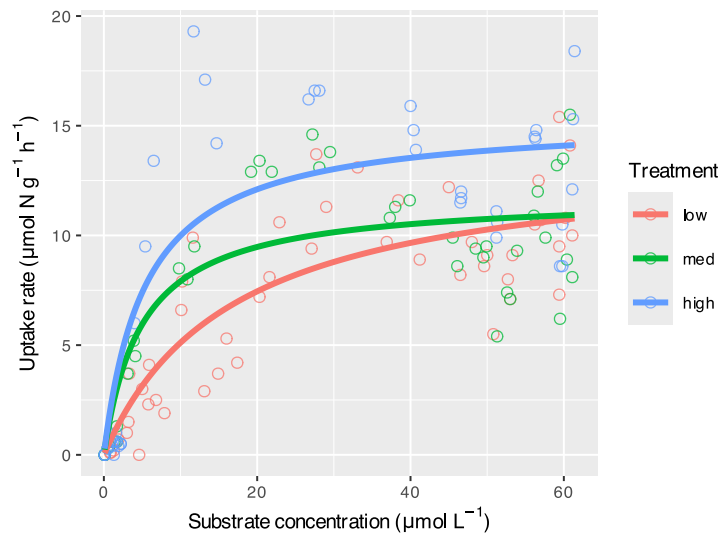


FIGURE 22.8. Treatment-specific nonlinear mixed-effects fits for the perturbation experiment.

The likelihood-ratio test comparing the shared-parameter model with the treatment-specific model shows that allowing V_{max} and K_m to vary by treatment improves fit strongly ($p = 0.0031$). That is the main inferential result. The treatment-specific parameter estimates suggest that the low-flow treatment has the highest fitted K_m in $\mu\text{mol L}^{-1}$, while the medium and high treatments reach similar uptake behaviour at lower substrate concentrations. The fitted V_{max} values, measured in $\mu\text{mol N g}^{-1} \text{h}^{-1}$, differ less strongly than the K_m estimates in this example.

In nonlinear mixed-effects work, the global model comparison is often more informative than reading each coefficient one-by-one as though they were ordinary linear-model terms. The treatment effect here is a shift in the shape of the kinetic curve under a repeated-measures design.

22.8.7 Reporting

WRITE-UP

Methods

Nitrate-uptake kinetics were analysed with a Michaelis-Menten nonlinear mixed-effects model fitted to repeated observations from perturbation flasks. Treatment was included as a fixed effect on the Michaelis-Menten parameters (V_{max} and K_m), replicate flasks were treated as random effects nested within treatment, and a first-order autoregressive correlation structure was used to account for serial dependence among repeated measurements within each flask. A global model with shared kinetic parameters across treatments was compared with a treatment-specific model using a likelihood-ratio test.

Results

Allowing Michaelis-Menten parameters to vary among treatments improved model fit relative to the shared-parameter model (likelihood-ratio test: $L.Ratio = 15.92$, $p = 0.0031$). The fitted treatment-specific curves differed mainly in the half-saturation constant, with the low-flow treatment showing the highest fitted K_m and the medium- and high-flow treatments reaching high uptake rates at lower substrate concentrations (Figure 22.8). Estimated parameter values were low flow: $V_{max} = 13.73 \mu\text{mol N g}^{-1} \text{h}^{-1}$ and $K_m = 16.85 \mu\text{mol L}^{-1}$; medium flow: $V_{max} = 11.81 \mu\text{mol N g}^{-1} \text{h}^{-1}$ and $K_m = 4.93 \mu\text{mol L}^{-1}$; high flow: $V_{max} = 15.35 \mu\text{mol N g}^{-1} \text{h}^{-1}$ and $K_m = 5.37 \mu\text{mol L}^{-1}$.

Discussion

The biological point is that repeated-measures uptake data can support treatment differences in kinetic behaviour, but only if the dependence structure is modelled explicitly. Ignoring the repeated sampling of the same flasks would treat correlated measurements as independent and would make the inferential result less trustworthy.

22.9 CHOOSING AMONG FLEXIBLE MODELS

There are a few practical considerations to keep in mind when choosing a mechanistic nonlinear model. Sometimes different curved models can provide similar fits to the same data, but they have very different implications for interpretation. The reason to prefer a mechanistic nonlinear model is precisely that its parameters mean something biologically. If that interpretability is weak or artificial, a descriptive alternative such as a polynomial regression or GAM may be more honest.

22.10 PRACTICAL CAUTION

Flexibility is valuable, but it comes with a cost.

- More flexible models are often harder to interpret.
- They can overfit small datasets.
- They may fit noise rather than structure if the biological logic is weak.
- Nonlinear least-squares models can fail to converge if the functional form or starting values are poor.

This is why flexible regression should be a response to a clear diagnostic or biological need, not just a default preference for curved lines.

22.11 SUMMARY

- Not all biological relationships are well described by straight lines.
- Mechanistic nonlinear models serve a distinct purpose within the broader family of curved-response models.
- Model choice should follow the biology and the inferential goal.
- Mechanistic nonlinear models are especially valuable when their parameters have clear biological meaning.
- Greater flexibility can improve fit, but it also increases interpretive demands and the risk of overfitting.

In the next chapter, I extend the regression sequence in a different direction by asking what happens when the scientific target is another part of the response distribution rather than the mean alone.

23

Quantile Regression

Modelling the Whole Response Distribution

IN THIS CHAPTER

- why modelling only the mean can hide important ecological structure;
- what quantile regression estimates and how it differs from ordinary least squares;
- when quantile regression is appropriate, and what its assumptions are;
- how to fit, compare, and visualise quantile regressions in R with `quantreg::rq()`;
- how to interpret lower, median, and upper quantile slopes biologically;
- how to report a quantile regression analysis in journal style.

TASKS TO COMPLETE IN THIS CHAPTER

- None

Most regression chapters so far have focused on the **conditional mean** of the response: ordinary least squares regression describes the average value of Y expected at each value of X .

Focussing on the mean may not always be enough. In ecology, the most interesting pattern may lie at the **edges** of the data cloud rather than through its centre. A predictor may place a strong upper limit on performance without explaining why some individuals or populations fall well below that limit. In such cases, a mean regression line can flatten over a broad wedge of points and miss the biological implication.

Instead of modelling only the mean, **quantile regression** models one or more conditional quantiles of the response distribution. This allows us to ask whether a predictor has the same effect on the lower, middle, and upper parts of the response.

This may be useful when:

- variance changes across the predictor gradient;
- upper or lower boundaries are biologically important;
- outliers make mean-based summaries unstable;
- the ecological question is about constraints, limits, or heterogeneous responses rather than only about the average trend.

23.1 KEY CONCEPTS

- **Quantile regression models conditional quantiles**, not the conditional mean.
- **Quantiles** divide a distribution into cumulative proportions such as the 10th, 50th, or 90th percentile.
- **The median is the 0.5 quantile**, so median regression is a special case of quantile regression.
- **Tau (τ)** denotes the target quantile and ranges from 0 to 1.
- **Inference can differ across quantiles**, meaning that the same predictor may be weak at the lower boundary but strong at the upper boundary.
- **Quantile regression is robust to heteroscedasticity** and does not require normally distributed residuals in the way ordinary least squares does.

23.2 WHEN THIS METHOD IS APPROPRIATE

Quantile regression is useful when:

- the response spread changes along the predictor gradient;
- the data form a wedge, funnel, or boundary pattern;
- upper or lower limits are biologically meaningful;
- you want a robust model of the median rather than the mean;
- the same predictor may affect slow-growing, average, and high-performing organisms differently.

Quantile regression is not a substitute for every nonlinear or hierarchical model. If the relationship is curved, use polynomial or additive extensions where needed. If the data are clustered by site, year, or individual, the dependence structure is still important and should not be ignored just because the model is quantile-based.

23.3 NATURE OF THE DATA AND ASSUMPTIONS

Quantile regression is more flexible than ordinary least squares, but it is not assumption-free.

The main requirements are:

1. independent observations;
2. a response variable that is continuous or at least ordered on a meaningful scale;
3. a predictor-response relationship that is appropriately specified for the quantile model being fitted;
4. enough data across the predictor range to estimate the chosen quantiles sensibly.

Unlike ordinary least squares, quantile regression does **not** assume normally distributed residuals or constant variance. In fact, it is most useful exactly when spread changes with the predictor.

What we need to consider is model form. If the relationship is curved, then a straight-line quantile model may be inadequate. It is also important to check that fitted quantile lines do not behave implausibly or cross in the range of interest.

23.4 THE CORE EQUATIONS

For a simple linear quantile regression, the conditional quantile of the response can be written as:

$$Q_Y(\tau | X) = \beta_0(\tau) + \beta_1(\tau)X \quad (23.1)$$

In Equation 23.1, $Q_Y(\tau | X)$ is the τ -th conditional quantile of the response, $\beta_0(\tau)$ is the intercept for that quantile, and $\beta_1(\tau)$ is the slope for that quantile. The principle idea is that the coefficients are allowed to change with τ .

Ordinary least squares chooses coefficients by minimising squared residuals. Quantile regression instead minimises an asymmetrically weighted sum of absolute residuals:

$$\min_{\beta} \sum_{i=1}^n \rho_{\tau}(y_i - x_i^T \beta) \quad (23.2)$$

where the check-loss function is:

$$\rho_{\tau}(u) = u [\tau - I(u < 0)] \quad (23.3)$$

You do not need to memorise Equation 23.2 and Equation 23.3, but they explain why quantile regression behaves differently from mean regression. Residuals above and below the fitted line are weighted differently depending on the target quantile.

For the median, positive and negative residuals are weighted equally; for the 90th quantile, underprediction is penalised more strongly than overprediction.

23.5 R FUNCTIONS

The main function is `quantreg::rq()`.

```
rq(y ~ x, tau = 0.5, data = df)           # median regression
rq(y ~ x, tau = 0.9, data = df)           # upper quantile
rq(y ~ x, tau = c(0.1, 0.5, 0.9), data = df) # several quantiles at once
summary(model, se = "nid")                 # coefficient summaries
anova(model_multi_tau)                     # compare slopes across quantiles
```

The syntax is similar to `lm()`, which makes quantile regression easy to adopt once you are comfortable with ordinary regression.

23.6 EXAMPLE 1: BLADE-LENGTH ENVELOPES IN LAMINARIA PALLIDA

23.6.1 Example Dataset

We use the `laminaria.csv` dataset from the course data folder. These data come from a morphometric sampling campaign on the kelp *Laminaria pallida* at 13 sites along the Cape Peninsula, spanning the west coast and False Bay regions. The sampled plants were large adult kelps, and several blade and stipe measurements were recorded for each individual.

Here we focus on the relationship between `stipe_length` and `blade_length`.

The biological question is simply, **does increasing stipe length shift the whole distribution of blade length in the same way, or is the effect stronger for the upper envelope of blade length than for the lower envelope?**

```
laminaria <- read_csv(here::here("data", "BCB744", "laminaria.csv"),
  show_col_types = FALSE) >
select(region, site, blade_length, stipe_length) >
filter(is.finite(blade_length), is.finite(stipe_length))

gt(head(laminaria, 10))
```

23.6.2 Do an Exploratory Data Analysis (EDA)

```
laminaria > summarise(
  n = n(),
  n_sites = n_distinct(site),
  min_stipe = min(stipe_length),
  max_stipe = max(stipe_length),
  mean_stipe = mean(stipe_length),
  mean_blade = mean(blade_length),
  sd_blade = sd(blade_length))
```

TABLE 23.1. A subset of the `laminaria.csv` data used in the quantile-regression example.

region	site	blade_length	stipe_length
WC	Kommetjie	160	120
WC	Kommetjie	120	149
WC	Kommetjie	110	97
WC	Kommetjie	159	167
WC	Kommetjie	149	146
WC	Kommetjie	107	161
WC	Kommetjie	104	110
WC	Kommetjie	111	136
WC	Kommetjie	178	176
FB	Bordjiesif North	145	82

)

```
R> # A tibble: 1 x 7
R>       n n_sites min_stipe max_stipe mean_stipe mean_blade sd_blade
R>   <int> <int>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
R> 1   140     13      34      224     111.     139.     23.9

ggplot(laminaria, aes(x = stipe_length, y = blade_length)) +
  geom_point(alpha = 0.65) +
  labs(
    x = "Stipe length (cm)",
    y = "Blade length (cm)"
  )
```

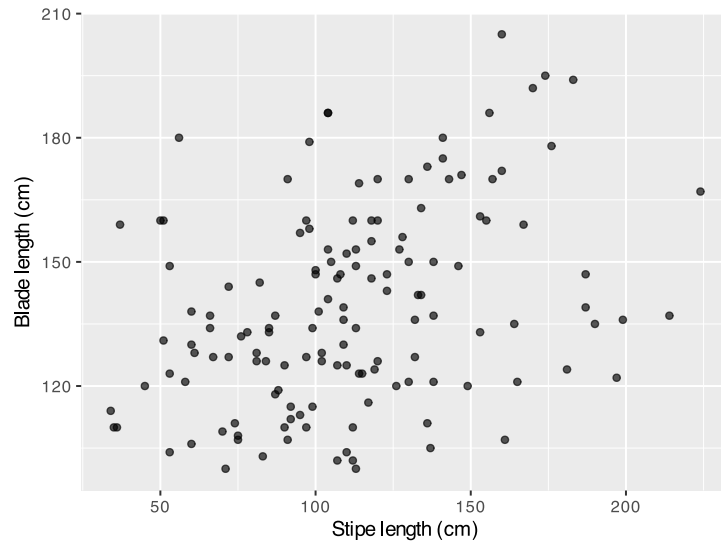


FIGURE 23.1. Blade length as a function of stipe length in adult *Laminaria pallida* collected across 13 sites.

The scatterplot in Figure 23.1 does not suggest a single, uniform linear effect. Blade length generally increases with stipe length, but the spread also changes across the predictor range. In particular, the upper edge of the point cloud steepens more strongly than the lower edge. That is exactly the kind of pattern for which quantile regression is useful.

23.6.3 State the Model Question and Hypotheses

We want to know whether stipe length affects the lower, median, and upper parts of the blade-length distribution differently.

For any fitted quantile, a useful hypothesis pair is:

$$H_0 : \beta_1(\tau) = 0$$

$$H_a : \beta_1(\tau) \neq 0$$

This asks whether stipe length has any association with blade length at the chosen quantile.

Because the chapter is about comparing quantiles, we also care about whether the slopes differ among quantiles:

$$H_0 : \beta_1(0.1) = \beta_1(0.5) = \beta_1(0.9)$$

$$H_a : \text{not all quantile slopes are equal}$$

23.6.4 Fit the Models

We fit one ordinary least squares regression for comparison, and then three quantile regressions for the 10th, 50th, and 90th quantiles.

```
fit_ols <- lm(blade_length ~ stipe_length, data = laminaria)

fit_rq10 <- rq(blade_length ~ stipe_length, tau = 0.1, data = laminaria)
fit_rq50 <- rq(blade_length ~ stipe_length, tau = 0.5, data = laminaria)
fit_rq90 <- rq(blade_length ~ stipe_length, tau = 0.9, data = laminaria)

fit_rq_all <- rq(blade_length ~ stipe_length,
  tau = c(0.1, 0.5, 0.9),
  data = laminaria)
```

The coefficient summaries are:

```
coef_tbl <- bind_rows(
  as.data.frame(summary(fit_ols)$coefficients) >
  rownames_to_column("term") >
  mutate(model = "OLS mean"),
  as.data.frame(summary(fit_rq10, se = "nid")$coefficients) >
  rownames_to_column("term") >
  mutate(model = "tau = 0.1"),
  as.data.frame(summary(fit_rq50, se = "nid")$coefficients) >
  rownames_to_column("term") >
  mutate(model = "tau = 0.5"),
  as.data.frame(summary(fit_rq90, se = "nid")$coefficients) >
  rownames_to_column("term") >
  mutate(model = "tau = 0.9")
) >
rename(
  estimate = 2,
  std_error = 3,
  t_value = 4,
  p_value = 5
) >
mutate(across(c(estimate, std_error, t_value), round, 3))

gt(coef_tbl)
```

To compare the quantile slopes directly, we test whether the fitted slopes are equal across $\tau = 0.1, 0.5$, and 0.9 :

```
anova(fit_rq_all)

R> Quantile Regression Analysis of Deviance Table
R>
R> Model: blade_length ~ stipe_length
R> Joint Test of Equality of Slopes: tau in 0.1 0.5 0.9
R>
R>   Df Resid Df F value Pr(>F)
R> 1 2      418 4.3359 0.01368 *
R> ---
```

TABLE 23.2. Coefficient summaries for the ordinary least squares fit and the 0.1, 0.5, and 0.9 quantile-regression fits.

term	estimate	std_error	t_value	p_value	model	Value
(Intercept)	114.967	5.699	20.172	5.492740e-43	OLS mean	NA
stipe_length	0.217	0.048	4.482	1.537502e-05	OLS mean	NA
(Intercept)	NA	5.948	16.537	0.000000e+00	tau = 0.1	98.3600000
stipe_length	NA	0.059	2.023	4.504584e-02	tau = 0.1	0.1200000
(Intercept)	NA	6.458	17.023	0.000000e+00	tau = 0.5	109.9299363
stipe_length	NA	0.073	3.488	6.523367e-04	tau = 0.5	0.2547771
(Intercept)	NA	10.231	12.770	0.000000e+00	tau = 0.9	130.6538462
stipe_length	NA	0.055	6.242	4.979948e-09	tau = 0.9	0.3461538

```
R> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

23.6.5 Test Assumptions / Check Diagnostics

Quantile regression does not require homoscedasticity or normal residuals, so diagnostics focus more on whether the fitted lines are sensible and whether the model form is adequate.

The first check is visual: do the fitted quantile lines follow the structure of the data, and do they avoid pathological crossing?

```
line_df <- tibble(
  model = c("Mean (OLS)", "tau = 0.1", "tau = 0.5", "tau = 0.9"),
  intercept = c(coef(fit_ols)[1], coef(fit_rq10)[1], coef(fit_rq50)[1],
    coef(fit_rq90)[1]),
  slope = c(coef(fit_ols)[2], coef(fit_rq10)[2], coef(fit_rq50)[2], coef(fit_rq90)[2])
)

ggplot(laminaria, aes(x = stipe_length, y = blade_length)) +
  geom_point(alpha = 0.45) +
  geom_abline(
    data = line_df,
    aes(intercept = intercept, slope = slope, colour = model, linetype = model),
    linewidth = 0.8
  ) +
  scale_colour_manual(
    values = c(
      "Mean (OLS)" = "black",
      "tau = 0.1" = "dodgerblue4",
      "tau = 0.5" = "firebrick3",
      "tau = 0.9" = "darkgreen"
    ),
    name = "Model"
  ) +
  scale_linetype_manual(
    values = c(
      "Mean (OLS)" = "dashed",
```



```

    "tau = 0.1" = "solid",
    "tau = 0.5" = "solid",
    "tau = 0.9" = "solid"
  ),
  name = "Model"
) +
labs(
  x = "Stipe length (cm)",
  y = "Blade length (cm)"
)

```

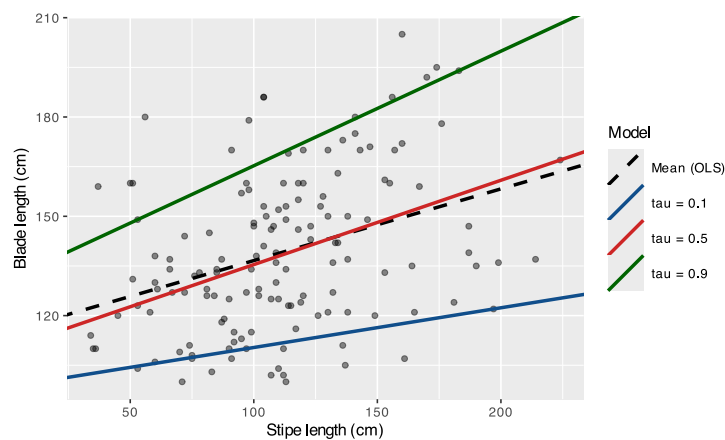


FIGURE 23.2. Ordinary least squares and quantile-regression fits for the 10th, 50th, and 90th quantiles.

The second useful check is whether the fitted slope itself changes smoothly across a wider range of quantiles.

```

tau_grid <- seq(0.1, 0.9, by = 0.1)

slope_grid <- map_dfr(
  tau_grid,
  \(tau_now) {
    fit_now <- rq(blade_length ~ stipe_length, tau = tau_now, data = laminaria)
    sum_now <- summary(fit_now, se = "nid")$coefficients
    tibble(
      tau = tau_now,
      slope = sum_now["stipe_length", 1],
      se = sum_now["stipe_length", 2]
    )
  }
) >
mutate(
  lo = slope - 1.96 * se,
  hi = slope + 1.96 * se
)

```

```
ggplot(slope_grid, aes(x = tau, y = slope)) +
  geom_ribbon(aes(ymin = lo, ymax = hi), alpha = 0.2) +
  geom_line(linewidth = 0.8) +
  geom_point(size = 1.5) +
  labs(
    x = expression(tau),
    y = "Estimated slope"
  )
)
```

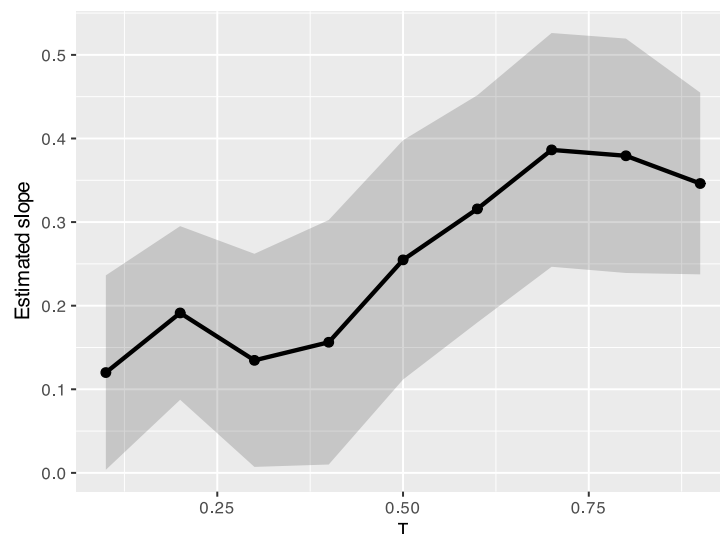


FIGURE 23.3. Estimated slope of blade length on stipe length across quantiles from 0.1 to 0.9.

In this example the fitted lines do not cross in the observed data range, and the slope increases steadily towards the upper quantiles. That is consistent with a biologically meaningful quantile effect rather than a numerical artefact.

23.6.6 Interpret the Results

The ordinary least squares fit gives the average change in blade length with stipe length. In these data that mean slope is positive, but it understates what is happening at the upper edge of the distribution.

The quantile regressions show that the relationship is not constant across the response distribution. This is clearest in Figure 23.2 and is reinforced by the slope-by-quantile summary in Figure 23.3:

- at $\tau = 0.1$, the fitted slope is about 0.12;
- at $\tau = 0.5$, the fitted slope is about 0.255;
- at $\tau = 0.9$, the fitted slope is about 0.346.

This means that longer stipes are associated with modest increases in blade length near the lower boundary of the data, but much larger increases near the upper boundary. In other words, stipe length is a much stronger predictor of the **longest** blades than of the **shortest** blades.

The joint quantile comparison must also be considered. The quantile ANOVA indicates that the slopes are not all the same across $\tau = 0.1, 0.5$, and 0.9 ($F = 4.34$, $p < 0.05$), and gives formal statistical support to what the figure already suggests visually.

Biologically, the result reveals that increasing stipe length appears to raise the upper potential for blade elongation more strongly than it shifts the lower end of the distribution. That is the kind of edge-case ecological pattern that mean ordinary least squares regression can hide; it is visible directly in Figure 23.2.

23.6.7 Reporting

WRITE-UP

Methods

Blade length was analysed as a function of stipe length in adult *Laminaria pallida* sampled across 13 sites on the Cape Peninsula. To assess whether stipe length affected only the mean blade length or different parts of the blade-length distribution, we fitted an ordinary least squares regression and quantile regressions for the 10th, 50th, and 90th quantiles using `rq()` from the **quantreg** package in R. We then tested whether the fitted slopes differed among quantiles.

Results

Blade length increased with stipe length in the mean model, but the fitted quantile slopes showed that this relationship was not constant across the response distribution (see Figure 23.2 and Figure 23.3). The estimated slope was weakest at the 10th quantile (0.12), intermediate at the median (0.255), and strongest at the 90th quantile (0.346). A joint test across the fitted quantiles indicated that the slopes were not all equal ($F = 4.34$, $p < 0.05$). Thus, increasing stipe length was associated with a much stronger increase in upper-bound blade length than in lower-bound blade length.

Discussion

The quantile regression reveals a biologically important pattern that would be muted in a mean-only analysis. Stipe length appears to influence the upper potential for blade development more strongly than the lower end of the blade-length distribution. This suggests that while longer stipes may permit greater blade elongation, shorter

blades can still occur across much of the stipe-length range, presumably because other unmeasured factors also constrain morphology.

23.7 WHAT TO DO WHEN ASSUMPTIONS FAIL / ALTERNATIVES

- If the fitted straight quantile lines are clearly inadequate, extend the model with polynomial terms or quantile smoothers rather than forcing a linear form.
- If quantile lines cross implausibly, reduce model complexity and inspect whether sparse data at extreme predictor values are driving the problem.
- If observations are not independent because of repeated measures, sites, or years, then the dependence structure must be addressed explicitly; ordinary `rq()` is not enough.
- If the scientific question is still about the conditional mean, ordinary least squares remains simpler and easier to interpret.

23.8 SUMMARY

- Quantile regression models the response distribution more completely than mean regression.
- It is especially useful for wedge-shaped ecological data and for questions about limits or constraints.
- It does not require constant variance or normally distributed residuals, but it still requires sensible model specification and independent observations.
- In the *Laminaria* example, stipe length had a much stronger effect on the upper quantiles of blade length than on the lower quantiles.
- The main biological gain is not mathematical novelty, but a clearer view of how predictors shape the whole distribution of ecological responses.

24

Prediction and Explanation

Choosing Models for Different Scientific Goals

IN THIS CHAPTER

- How explanatory and predictive models differ
- Why modelling purpose should guide model choice
- When interpretation matters more than raw predictive accuracy
- Why out-of-sample thinking matters for prediction
- What this distinction means in practice

TASKS TO COMPLETE IN THIS CHAPTER

- None

Not all models are built for the same purpose. Some are built to explain mechanisms, isolate effects, and support biologically interpretable statements. Others are built to predict future outcomes as accurately as possible. These goals overlap, but they are not identical.

That distinction matters because a model that is good for explanation is not always the one that is best for prediction. The trade-off becomes especially important once models become high-dimensional, predictors become strongly overlapping, or out-of-sample performance becomes the main concern.

This distinction has been present in the background of many earlier chapters, even when I did not always name it explicitly. When I insisted on clear biological interpretation in the regression chapters, I was leaning toward explanation. When I

introduced cross-validation and model comparison, I was already moving toward prediction. Here I make that contrast explicit so that model quality can be judged against purpose rather than by one generic standard of “best fit.”

24.1 KEY CONCEPTS

- **Explanation** and **prediction** are related but different modelling goals.
- **Interpretability** matters most when explanation is the main goal.
- **Out-of-sample performance** matters most when prediction is the main goal.
- **Model choice should be justified by purpose**, not only by convention.

24.2 EXPLANATION VERSUS PREDICTION

An explanatory model typically emphasises:

- biologically interpretable coefficients;
- careful model specification;
- defensible causal or process-based language;
- explicit treatment of confounders and assumptions.

A predictive model typically emphasises:

- accurate performance on new data;
- stability under collinearity or large predictor sets;
- lower prediction error rather than coefficient-by-coefficient interpretation.

Neither goal is superior in the abstract. The important thing is to know which goal you are pursuing.

24.3 CROSS-VALIDATION AND OUT-OF-SAMPLE THINKING

If prediction is the main goal, the model should be evaluated primarily on data that were not used to fit it. Cross-validation is one of the main ways to do this.

The basic idea is:

1. split the data into training and validation parts;
2. fit the model on the training data;
3. evaluate predictive performance on held-out data;
4. repeat this across several splits.

This kind of out-of-sample thinking becomes especially important when models are built for forecasting, classification, or any situation where performance on new data matters more than close interpretation of each coefficient.

24.4 A SHORT PAIRED EXAMPLE

The same dataset can motivate different models depending on the scientific goal. Suppose we measure algal biomass together with temperature, nutrient concentration, wave exposure, and turbidity. If the biological question is whether biomass depends on temperature differently under different nutrient conditions, an explanatory model should keep that interaction central. If the goal is simply to forecast biomass at new sites, we may tolerate a more flexible and less interpretable model if it predicts better.

```
set.seed(42)
eco_dat <- tibble(
  temp = runif(180, 8, 24),
  nutrient = runif(180, 0.5, 4.5),
  wave = runif(180, 0, 8),
  turbidity = runif(180, 0, 5)
) >
mutate(
  biomass = 6 + 0.8 * temp + 2.1 * nutrient - 0.45 * wave +
    0.32 * temp * nutrient - 0.05 * temp^2 + 0.35 * turbidity +
    rnorm(n(), sd = 3.2)
)

set.seed(42)
train_id <- sample(seq_len(nrow(eco_dat)), size = round(0.75 * nrow(eco_dat)))
train_dat <- eco_dat[train_id, ]
test_dat <- eco_dat[-train_id, ]

mod_expl <- lm(biomass ~ temp * nutrient + wave, data = train_dat)
mod_pred <- lm(
  biomass ~ temp * nutrient + wave + turbidity +
    I(temp^2) + I(nutrient^2) + temp:wave + nutrient:wave,
  data = train_dat
)

rmse <- function(obs, pred) sqrt(mean((obs - pred)^2))

comparison_tbl <- tibble(
  model_goal = c("Explanation", "Prediction"),
  model = c(
    "biomass ~ temp * nutrient + wave",
    "biomass ~ temp * nutrient + wave + turbidity + I(temp^2) + I(nutrient^2) + temp:wave
    + nutrient:wave"
  ),
  adj_r2_training = c(
    summary(mod_expl)$adj.r.squared,
    summary(mod_pred)$adj.r.squared
  ),
  rmse_test = c(
    rmse(test_dat$biomass, predict(mod_expl, newdata = test_dat)),
    rmse(test_dat$biomass, predict(mod_pred, newdata = test_dat))
  )
) >
mutate(across(c(adj_r2_training, rmse_test), \(x) round(x, 3)))

comparison_tbl

R> # A tibble: 2 x 4
R>   model_goal model adj_r2_training rmse_test
```

```

R>   <chr>           <chr>                                <dbl>   <dbl>
R> 1 Explanation biomass ~ temp * nutrient + wave          0.873   3.27
R> 2 Prediction  biomass ~ temp * nutrient + wave +~       0.876   2.97

```

The explanatory model is smaller because it stays close to the biological question: does the temperature effect depend on nutrient concentration, after accounting for wave exposure? Its coefficients can be discussed directly. The predictive model is judged differently. It adds curvature and extra interactions because the goal is lower error on new data, not a cleaner causal statement.

The same response variable and the same predictors do not force the same model. The modelling goal determines what counts as a good model.

24.5 WHAT THIS MEANS IN PRACTICE

If you are building a model to explain ecological structure or process, you will usually prioritise:

- coefficients that can be interpreted directly;
- predictors chosen for biological reasons rather than only for predictive convenience;
- explicit discussion of assumptions, confounding, and design limitations;
- model statements that remain close to the original scientific question.

If you are building a model to predict accurately, you will usually prioritise:

- out-of-sample predictive performance;
- stable behaviour when predictors overlap or become numerous;
- procedures that reduce overfitting;
- honest evaluation on data not used in fitting.

These priorities often overlap, but they are not identical. A model that is excellent for explanation may be suboptimal for forecasting. A model that predicts well may be harder to interpret mechanistically.

24.6 PRACTICAL GUIDANCE

If your goal is primarily **explanation**:

- prefer models whose coefficients can be interpreted clearly;
- keep the biological argument central;
- avoid adding complexity unless it strengthens the scientific question.

If your goal is primarily **prediction**:

- evaluate models out of sample;
- accept that some loss of interpretability may be worthwhile;
- consider methods that shrink coefficients or otherwise stabilise model performance.

24.7 WHY THIS DISTINCTION MATTERS FOR THE REST OF THE COURSE

Once the modelling goal is named clearly, several earlier chapter choices become easier to interpret. Collinearity is especially damaging for explanation because it destabilises biological attribution. Overfitting is especially damaging for prediction because it reduces performance on new data. Diagnostics also change meaning slightly: some are about whether the model is interpretable and defensible, others about whether it generalises honestly.

The distinction is therefore not philosophical decoration. It tells you what kind of error you most want to avoid, what model complexity is acceptable, and which summaries deserve priority when you report the result.

24.8 SUMMARY

- Explanation and prediction are overlapping but distinct modelling goals.
- Ordinary regression is often strongest for explanation when the model is interpretable and well specified.
- Cross-validation is central when predictive performance is the goal.
- The next chapter takes up regularisation directly and shows how ridge, lasso, and elastic net help when prediction and coefficient stability matter most.

In this chapter, I make the modelling goal explicit. In the next chapter, I turn to regularisation, and in the final chapter I close the sequence with reproducible analytical workflow.

25

Regularisation

Ridge, Lasso, Elastic Net, and Cross-Validation

IN THIS CHAPTER

- Why regularisation is useful in high-dimensional or collinear settings
- The logic of ridge, lasso, and elastic net
- How coefficient shrinkage changes model behaviour
- Why cross-validation is needed
- How regularisation relates to prediction and explanation

TASKS TO COMPLETE IN THIS CHAPTER

- None

Regularisation techniques are useful when ordinary multiple regression starts to struggle under the weight of many predictors, overlapping predictors, or a modelling goal that leans more toward prediction than explanation. In that setting, ordinary least squares can produce unstable coefficients, poor generalisation, and models that appear stronger in the sample than they really are.

Regularisation addresses this by shrinking coefficients towards zero. In some cases, this simply stabilises them. In others, it also removes weak predictors from the model entirely. This makes regularisation relevant when we want to reduce overfitting, manage multicollinearity, or build models that predict more reliably on new data.

In this chapter, I follow directly from the previous one. There I distinguished explanation from prediction. Here I develop one of the main statistical responses to that distinction.

25.1 KEY CONCEPTS

- **Regularisation** shrinks coefficients to stabilise the model.
- **Ridge regression** shrinks all coefficients continuously towards zero.
- **Lasso regression** can shrink some coefficients exactly to zero.
- **Elastic net** combines ridge and lasso behaviour.
- **Cross-validation** is used to tune the amount of shrinkage.
- **The interpretation** of coefficients is changed by regularisation, which improves model stability, prediction, and objective variable reduction.

25.2 WHEN THIS METHOD IS APPROPRIATE

You should consider regularisation when:

- the predictor set is large relative to the amount of data;
- several predictors are strongly correlated;
- ordinary multiple regression produces unstable coefficients;
- prediction on new data matters more than exact coefficient interpretation;
- you want a data-driven complement to the more theory-driven model selection discussed in [Chapter 14](#) and the collinearity material in [Chapter 16](#).

Regularisation is not a magic correction for weak scientific questions or poor study design. It is still your responsibility to define sensible predictors and to understand the biology of the system. These methods work best when they extend ecological reasoning rather than replace it.

25.3 WHY REGULARISATION MATTERS

Regularisation addresses several common modelling problems.

Variable selection becomes difficult when many candidate predictors are available and only some are genuinely useful. Traditional selection procedures often rely on stepwise inclusion or exclusion, or on statistics such as VIF. Regularisation offers an alternative data-driven route.

Overfitting occurs when the model begins to fit noise together with the underlying biological signal. Such a model often performs well on the observed data but poorly on new observations.

Multicollinearity inflates standard errors and destabilises coefficients when predictors overlap. Regularisation reduces this instability by shrinking coefficients, which usually improves model behaviour even if it introduces some bias.

Regularisation aims to produce a model that is more stable, more generalisable, and more useful for the modelling goal at hand, rather than to recover perfectly unbiased coefficients.

25.4 THE CORE EQUATIONS

Regularisation methods all start from the ordinary least-squares objective and then add a penalty term. The key teaching point is therefore not to memorise three unrelated formulas, but to see how each method modifies the same underlying fitting problem.

25.4.1 Ridge Regression

Ridge regression adds a penalty proportional to the squared size of the coefficients. Large coefficients are penalised more heavily, so the fitted model shrinks them towards zero without setting them exactly to zero:

$$\min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (25.1)$$

The practical effect is that all predictors remain in the model, but the most unstable coefficients are tamed. Ridge is therefore especially useful when multicollinearity is the main problem and you do not necessarily want automatic variable removal.

25.4.2 Lasso Regression

Lasso regression uses a penalty based on the absolute values of the coefficients:

$$\min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (25.2)$$

This has an important consequence: some coefficients can be shrunk all the way to zero.

That means lasso does two jobs at once. It shrinks the model, and it can also perform automatic variable selection. When you have many candidate predictors and suspect that some contribute little to predictive performance, lasso can be attractive.

25.4.3 Elastic Net

Elastic net combines the ridge and lasso penalties:

$$\min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \left[(1 - \alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right] \right\} \quad (25.3)$$

Equations Equation 25.1, Equation 25.2, and Equation 25.3 all use λ to control the strength of shrinkage, while the elastic-net parameter α controls the balance between ridge-like and lasso-like behaviour.

In practice, elastic net is often a strong default when you are unsure whether ridge or lasso is the more appropriate starting point.

25.5 CROSS-VALIDATION

The amount of shrinkage is controlled by a tuning parameter, usually written as λ . When $\lambda = 0$, the fitted model behaves like ordinary least squares. As λ increases, the penalty grows stronger and coefficients are shrunk more aggressively.

The problem is that we do not know the best value of λ in advance. This is where cross-validation becomes central.

In k -fold cross-validation, the data are split into k subsets. The model is trained repeatedly on $k - 1$ folds and evaluated on the held-out fold. This gives an estimate of how well the model performs away from the data used to fit it. We then choose the tuning parameter that gives the best average predictive performance.

Cross-validation therefore helps us avoid selecting a model that is optimised only for the present sample.

25.6 R FUNCTIONS

In R, the usual introductory function for regularised regression is `glmnet::cv.glmnet()`:

```
glmnet::cv.glmnet(x, y, alpha = 0) # ridge
glmnet::cv.glmnet(x, y, alpha = 1) # lasso
glmnet::cv.glmnet(x, y, alpha = 0.5) # elastic net
```

The important practical detail is that `glmnet` expects a predictor matrix rather than the formula interface used by `lm()`.

25.7 EXAMPLE 1: RIDGE REGRESSION WITH THE SEAWEED DATA

The seaweed dataset should now be familiar. We will use the climatic predictors `annMean`, `augMean`, `augSD`, `febSD`, and `febRange` to predict Y .

25.7.1 Prepare the Data

The response is centred, and the predictors are supplied as a matrix:

```
y <- sw >
  select(Y) >
  scale(center = TRUE, scale = FALSE) >
  as.matrix()

X <- sw >
  select(-X, -dist, -bio, -Y, -Y1, -Y2) >
  as.matrix()
```

25.7.2 Fit the Cross-Validated Ridge Model

```
set.seed(123)
ridge_cv <- cv.glmnet(
  X, y,
  alpha = 0,
  lambda = lambdas_to_try,
  standardize = TRUE,
  nfolds = 10
)
```

25.7.3 Inspect the Cross-Validation Curve

The two vertical lines in Figure 25.1 identify the λ that minimises cross-validated error (`lambda.min`) and the larger, more conservative value within one standard error of that minimum (`lambda.1se`). The latter is often chosen when a simpler, more stable model is preferred.

25.7.4 Coefficient Paths and the Meaning of Shrinkage

Cross-validation tells us which value of λ is attractive for prediction, but it does not by itself show what the penalty is *doing* to the model. A coefficient-path plot fills that gap. It tracks each fitted coefficient as λ changes from very small values (weak penalty) to large values (strong penalty).

For ridge regression, the usual pattern is that all coefficients move smoothly towards zero, but none are dropped completely. For lasso, some coefficients are driven exactly to zero as the penalty strengthens. This is the visual expression of the conceptual difference between the two methods.

In Figure 25.2, each coloured trace is one predictor coefficient. Moving from left to right corresponds to stronger shrinkage. The lines contract towards zero, but ridge keeps every predictor in the model.

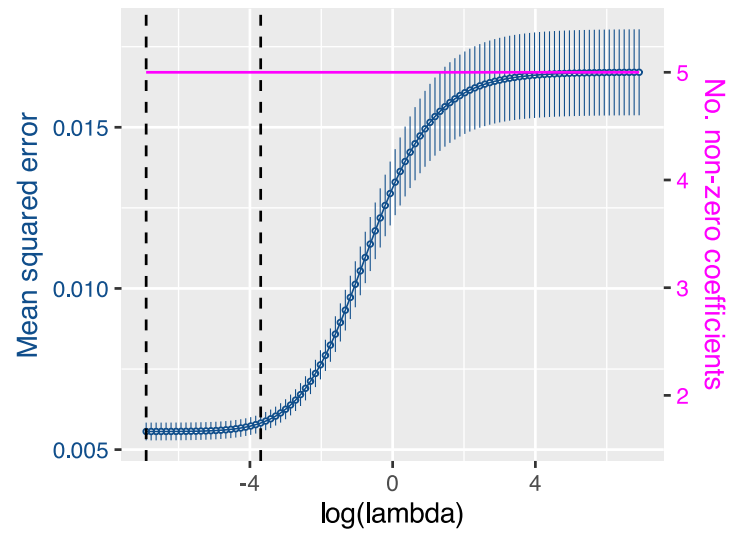


FIGURE 25.1. Cross-validation statistics for ridge regression applied to the seaweed data.

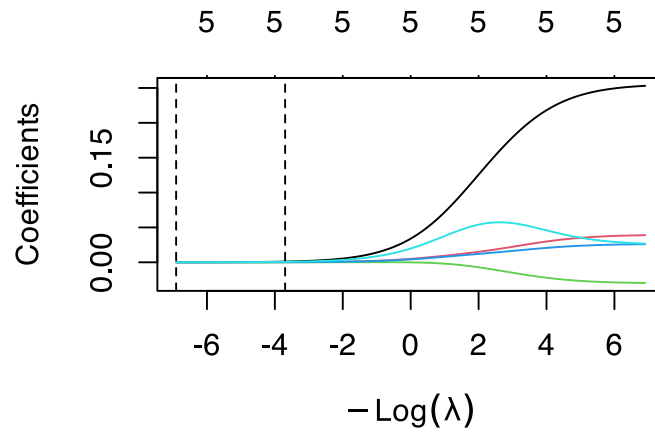


FIGURE 25.2. Coefficient paths for ridge regression. As the penalty increases, all coefficients are shrunk towards zero, but none are removed completely.

The vertical reference lines show where the cross-validation procedure places `lambda.min` and `lambda.1se`. Reading the plot this way helps connect model selection to coefficient behaviour. At `lambda.min`, shrinkage is weaker because the model is allowed to keep more of the original coefficient size. At `lambda.1se`, shrinkage is stronger, so the fitted model is more conservative even though its estimated prediction error is still close to the minimum.

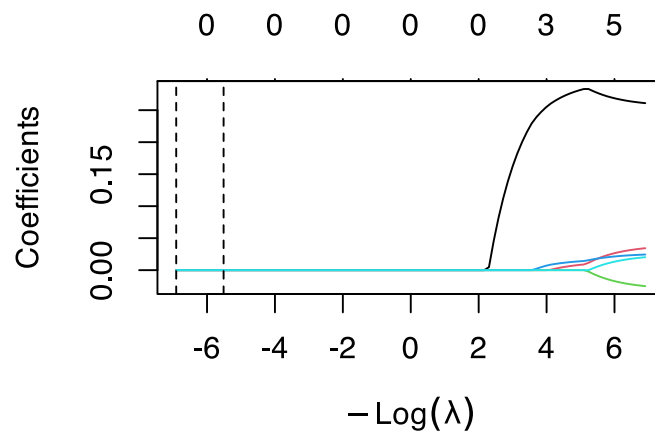


FIGURE 25.3. Coefficient paths for lasso regression. As the penalty increases, some coefficients are shrunk all the way to zero.

In Figure 25.3, some traces hit the horizontal axis and stay there. Those are the predictors that lasso has excluded. This is why coefficient-path plots are so useful pedagogically: they let you see which variables remain influential across a range of penalties and which ones disappear quickly once shrinkage becomes strong.

25.7.5 Extract the Fitted Model

```
ridge_model <- glmnet(
  X, y,
  alpha = 0,
  lambda = ridge_cv$lambda.min,
  standardize = TRUE
)
ridge_pred <- predict(ridge_model, X)
```

```

ridge_rsqu <- cor(y, ridge_pred) ^ 2
coef(ridge_model)

R> 6 x 1 sparse Matrix of class "dgCMatrix"
R>          s0
R> (Intercept) -0.12916991
R> augMean      0.25319971
R> febRange      0.03884481
R> febSD        -0.02964465
R> augSD         0.02599145
R> annMean       0.02672789

ridge_model_1se <- glmnet(
  X, y,
  alpha = 0,
  lambda = ridge_cv$lambda.1se,
  standardize = TRUE
)

ridge_pred_1se <- predict(ridge_model_1se, X)
ridge_rsqu_1se <- cor(y, ridge_pred_1se) ^ 2
ridge_coef_norm <- sqrt(sum(as.matrix(coef(ridge_model))[-1, 1] ^ 2))
ridge_coef_norm_1se <- sqrt(sum(as.matrix(coef(ridge_model_1se))[-1, 1] ^ 2))

```

25.7.6 Interpret the Ridge Results

Ridge keeps all predictors in the model, but shrinks them towards zero. This means the coefficients remain interpretable in the broad sense, but their absolute values are biased by design. Ridge is therefore most useful when the main aim is stable prediction or improved behaviour under collinearity, rather than precise coefficient interpretation.

In this seaweed example, the regularised model still retains all climatic predictors, but it reduces their instability and produces a model that is better suited to predictive use than an ordinary unpenalised fit under the same overlap among predictors.

25.7.7 $\lambda_{\min}$ Versus λ_{1se}

The choice between $\lambda_{\min}$ and λ_{1se} is often more important in practice than the choice between two nearby values of R^2 . These two tuning parameters answer slightly different modelling preferences:

- $\lambda_{\min}$ gives the value of λ with the lowest estimated cross-validated error.
- λ_{1se} gives the largest value of λ whose error is still within one standard error of the minimum.

That means λ_{1se} accepts a tiny loss in apparent predictive performance in exchange for stronger shrinkage and usually a more stable model. In ridge regression, that means smaller coefficients. In lasso or elastic net, it often means fewer non-zero predictors as well.

```

ridge_compare <- tibble(
  lambda_choice = c("lambda.min", "lambda.1se"),
  lambda_value = c(ridge_cv$lambda.min, ridge_cv$lambda.1se),
  training_rsqa = c(ridge_rsqa, ridge_rsqa_1se),
  non_zero = c(
    sum(as.matrix(coef(ridge_model))[-1, 1] != 0),
    sum(as.matrix(coef(ridge_model_1se))[-1, 1] != 0)
  ),
  coefficient_norm = c(ridge_coef_norm, ridge_coef_norm_1se)
) >
mutate(
  lambda_value = signif(lambda_value, 3),
  training_rsqa = round(training_rsqa, 3),
  coefficient_norm = round(coefficient_norm, 3)
)

ridge_compare

```

```

R> # A tibble: 2 x 5
R>   lambda_choice lambda_value training_rsqa non_zero coefficient_norm
R>   <chr>          <dbl>          <dbl>    <int>          <dbl>
R> 1 lambda.min      0.001          0.671      5          0.261
R> 2 lambda.1se      0.0248         0.664      5          0.218

```

In the seaweed example, ridge keeps all predictors at both tuning values, so the practical difference is not variable inclusion but coefficient size. The `lambda.1se` fit has the smaller coefficient norm, which means the predictor effects have been pulled closer to zero overall. That is the usual ridge trade-off: a small increase in penalty often changes the model by stabilising coefficients rather than by changing which predictors appear in the equation.

25.7.8 Reporting

WRITE-UP

Methods

Ridge regression was fitted to the seaweed climate data using 10-fold cross-validation to select the optimal penalty parameter. The predictors were entered as a matrix and the response was centred prior to fitting.

Results

The final ridge model retained all five climatic predictors and explained a substantial proportion of the variation in the response (R^2 approximately 0.67). The regularised fit reduced coefficient instability while preserving the full predictor set.

Discussion

The point of ridge regression is not exact coefficient interpretation. In Discussion, the emphasis would be that shrinkage improved model stability and predictive behaviour under overlapping climatic predictors.

25.8 EXAMPLE 2: LASSO REGRESSION

Lasso uses the same data structure and the same cross-validation logic. The key difference is $\alpha = 1$.

```
lasso_model <- glmnet(
  X, y,
  alpha = 1,
  lambda = lasso_cv$lambda.min,
  standardize = TRUE
)

lasso_model_1se <- glmnet(
  X, y,
  alpha = 1,
  lambda = lasso_cv$lambda.1se,
  standardize = TRUE
)

lasso_pred <- predict(lasso_model, X)
lasso_rsq <- cor(y, lasso_pred) ^ 2
lasso_pred_1se <- predict(lasso_model_1se, X)
lasso_rsq_1se <- cor(y, lasso_pred_1se) ^ 2
```

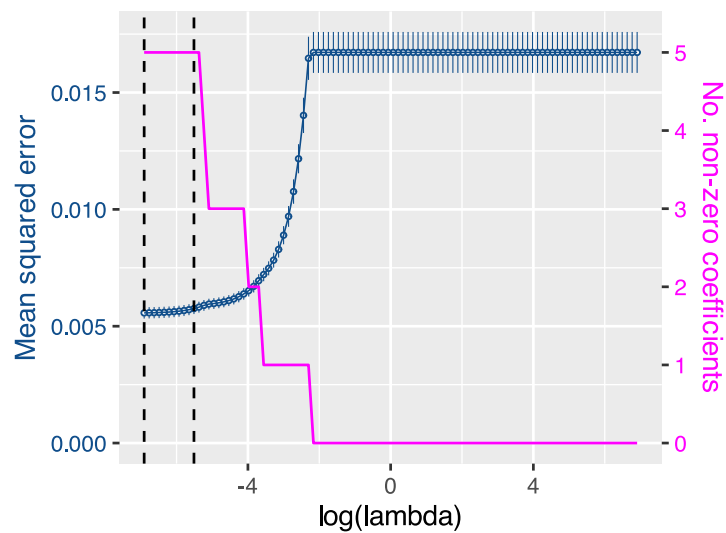


FIGURE 25.4. Cross-validation statistics for lasso regression applied to the seaweed data.

```
coef(lasso_model)
R> 6 x 1 sparse Matrix of class "dgCMatrix"
R>          s0
```

```
R> (Intercept) -0.12886019
R> augMean      0.26097296
R> febRange     0.03431981
R> febSD        -0.02497532
R> augSD        0.02441380
R> annMean      0.02021480
```

The important feature of lasso is that some coefficients can be set exactly to zero. This makes lasso useful when you want the model itself to carry out a degree of variable selection.

In the seaweed example, the chosen penalty reduces the effective model complexity by shrinking weaker coefficients more strongly. The resulting model is easier to simplify than the ridge model because some predictors can be removed altogether.

```
lmin_terms <- rownames(coef(lasso_model))[-1][as.matrix(coef(lasso_model))[-1, 1] != 0]
l1se_terms <- rownames(coef(lasso_model_1se))[-1][as.matrix(coef(lasso_model_1se))[-1, 1]
!= 0]

lasso_compare <- tibble(
  lambda_choice = c("lambda.min", "lambda.1se"),
  lambda_value = c(lasso_cv$lambda.min, lasso_cv$lambda.1se),
  training_rsqa = c(lasso_rsqa, lasso_rsqa_1se),
  non_zero = c(
    sum(as.matrix(coef(lasso_model))[-1, 1] != 0),
    sum(as.matrix(coef(lasso_model_1se))[-1, 1] != 0)
  ),
  retained_predictors = c(
    paste(lmin_terms, collapse = ", "),
    paste(l1se_terms, collapse = ", ")
  )
) >
mutate(
  lambda_value = signif(lambda_value, 3),
  training_rsqa = round(training_rsqa, 3)
)

lasso_compare

R> # A tibble: 2 x 5
R>   lambda_choice lambda_value training_rsqa non_zero retained_predictors
R>   <chr>          <dbl>          <dbl>    <int> <chr>
R> 1 lambda.min      0.001            0.67         5 augMean, febRange, febS~
R> 2 lambda.1se      0.00404          0.661         5 augMean, febRange, febS~
```

For lasso, the `lambda.min` versus `lambda.1se` distinction is easiest to explain in terms of retained predictors. `lambda.min` usually gives the lowest estimated prediction error, while `lambda.1se` often keeps fewer predictors and gives a leaner model. If the two fits perform similarly, `lambda.1se` is often preferable when explanation, communication, or field interpretation matters. If prediction is the dominant aim and the larger active set clearly performs better under resampling, `lambda.min` may be the better choice.

25.8.1 Reporting

WRITE-UP**Methods**

Lasso regression was fitted to the seaweed climate data, with the amount of shrinkage selected by 10-fold cross-validation. The purpose was to assess whether coefficient shrinkage combined with variable removal could produce a more compact predictive model.

Results

The regularised fit reduced model complexity by shrinking weak coefficients strongly and, where appropriate, setting some coefficients to zero. The final model explained substantial variation in the response (R^2 approximately 0.67) while providing a more compact predictor set than ridge regression.

Discussion

Lasso offers a bridge between prediction and variable selection. The simplified model is useful when the goal is not only stability, but also a more compact predictor set.

25.9 EXAMPLE 3: ELASTIC NET REGRESSION

Elastic net introduces a second tuning parameter, α , which controls the balance between ridge and lasso behaviour.

```
cv_results <- lapply(alphas_to_try, function(a) {
  cv.glmnet(
    X, y,
    alpha = a,
    lambda = lambdas_to_try,
    standardize = TRUE,
    nfolds = 10
  )
})

best_result <- which.min(sapply(cv_results, function(x) min(x$cvm)))
best_alpha <- alphas_to_try[best_result]
best_lambda <- cv_results[[best_result]]$lambda.min

elastic_model <- glmnet(
  X, y,
  alpha = best_alpha,
  lambda = best_lambda,
  standardize = TRUE
)

elastic_pred <- predict(elastic_model, X)
elastic_rsqa <- cor(y, elastic_pred) ^ 2

coef(elastic_model)
```

```
R> 6 x 1 sparse Matrix of class "dgCMatrix"
R>      s0
R> (Intercept) -0.12910260
R> augMean      0.25467960
R> febRange     0.03797258
R> febSD        -0.02874393
```

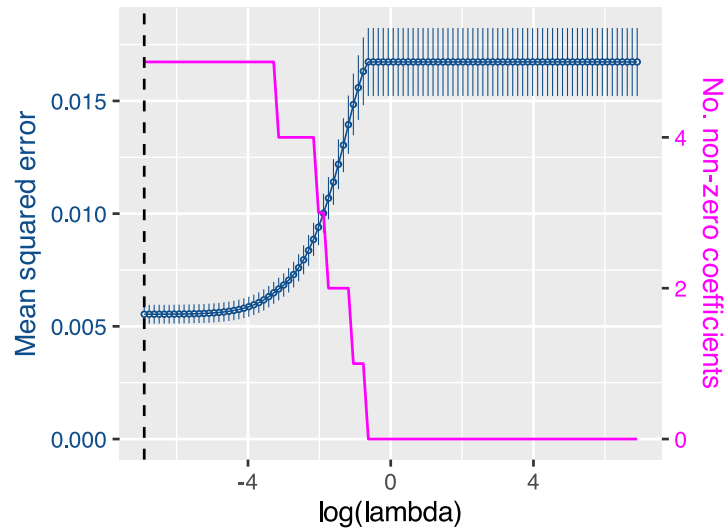


FIGURE 25.5. Cross-validation statistics for elastic net regression applied to the seaweed data.

```
R> augSD      0.02568390
R> annMean    0.02547694
```

Elastic net is often useful when predictors occur in correlated groups. Lasso may select one predictor and drop the others. Ridge keeps them all. Elastic net often provides a more balanced compromise.

25.9.1 Reporting

WRITE-UP

Methods

Elastic net regression was used to model the seaweed response, with both the mixing parameter and the penalty term selected by cross-validation. This allowed the model to combine ridge-like shrinkage with lasso-like variable reduction.

Results

The optimal model combined coefficient shrinkage with variable reduction ($\alpha = 0.2$), explained substantial variation in the response (R^2 approximately 0.67), and provided a stable compromise between ridge and lasso behaviour.

Discussion

The value of elastic net is pragmatic balance: it handles correlated predictors more flexibly than lasso alone while still allowing the model to simplify where the data support that.

25.10 THEORY-DRIVEN AND DATA-DRIVEN VARIABLE SELECTION

The choice between theory-driven and data-driven variable selection should not be treated as a fight with a single winner. In practice, the strongest ecological modelling often combines both.

Theory-driven selection is central to the scientific method. It uses prior ecological reasoning to define a defensible set of candidate predictors. This keeps the model close to mechanism and strengthens interpretation.

Data-driven methods, including regularisation, can then help assess which predictors contribute most strongly to predictive performance, where redundancy lies, and how strongly coefficients need to be stabilised. They are especially useful in high-dimensional settings or when the predictors are strongly overlapping.

The danger is to let automated variable selection replace ecological thinking. Regularisation can help refine the model, but it cannot tell you what the scientific question ought to be.

25.11 SUMMARY

- Regularisation is useful when predictors are many, overlapping, or likely to produce unstable ordinary regression coefficients.
- Ridge shrinks all coefficients, lasso can remove some, and elastic net blends both behaviours.
- Cross-validation is central to selecting the amount of shrinkage.
- These methods are usually most useful when prediction, stability, and model simplification matter more than exact coefficient interpretation.
- Regularisation should complement, not replace, ecological reasoning and theory-driven model building.

The final chapter now turns from model choice to the workflow required to make the whole analysis transparent and reproducible.

Part V

Reproducibility

The final part brings the workflow back into view. A defensible analysis is not complete when the model has been fitted. It must still be documented, regenerated, checked, and communicated in a way that leaves a visible path from source data to reported conclusion.

This closing chapter therefore treats reproducibility as part of statistical practice rather than as a separate technical concern. The same care applied earlier to design, assumptions, and model choice must also be applied to how the analysis is stored, rendered, and revised.

This is the point at which the whole course closes its loop: question, design, data, model, interpretation, and transparent reporting all have to remain joined.

26

Reproducible Workflow

From Analysis to Transparent Reporting

IN THIS CHAPTER

- why reproducibility is part of statistical practice rather than an optional extra;
- how a Quarto project links data, code, figures, tables, and narrative;
- what a practical reproducible workflow looks like in this project;
- how to generate a small report component directly from source data and code;
- how to write up results in a way that remains traceable to the analysis.

TASKS TO COMPLETE IN THIS CHAPTER

- None

A statistically correct analysis remains incomplete until everyone can see how it was produced. Reproducibility is the workflow that keeps data, code, tables, figures, and written interpretation linked from the beginning rather than cobbled together at the end, as one would do if MS Word is our writing tool of preference.

This final chapter therefore closes the loop opened at the start of the course. Earlier chapters focused on questions, design, assumptions, inference, and models. Here I ask whether the entire analytical chain remains visible and regenerable. If it does not, then even a technically correct analysis becomes harder to trust, revise, and communicate.

In practice, reproducibility means that:

- the data source can be identified;

- the analysis steps can be rerun from code;
- figures and tables are generated from source, not edited by hand;
- the final report remains connected to the analysis that produced it.

This is implicit in maintaining scientific credibility. Quarto gives us this ability.

26.1 KEY CONCEPTS

- **Reproducibility** means the analysis can be rerun from source.
- **Transparency** means analytical decisions are visible and documented.
- **Traceability** means every reported result has a path back to data and code.
- **Literate analysis** means code, output, and prose are kept close together.
- **Project structure matters** because disorder is one of the main causes of irreproducible work.

26.2 WHEN THIS METHOD IS APPROPRIATE

In this chapter, I take a different approach because I focus less on a single statistical test and more on the workflow habits that we must keep in mind **all the time** when:

- exploring data;
- fitting models;
- producing figures and tables;
- writing interpretations;
- revising a report after feedback.

In the earlier chapters, I showed what to analyse and how to do it, but here I focus on how to keep that analysis reproducible from beginning to end.

26.3 NATURE OF THE DATA AND ASSUMPTIONS

Reproducibility has practical assumptions of its own:

1. data files should live in stable, known locations;
2. analysis steps should be saved in code rather than performed only interactively;
3. outputs should be regenerated rather than edited manually;
4. the report (or scientific publication, even) should be linked directly to the analysis that created it.

If any of those fail, reproducibility begins to collapse, even if the statistical model itself is fine.

TABLE 26.1. A subset of the `laminaria.csv` dataset used in the reproducible-workflow example.

region	site	Ind	blade_weight	blade_length	blade_thickness	stipe_mass	stipe_length	stipe_diameter
WC	Kommetjie	2	1.90	160	2.00	1.50	120	56.0
WC	Kommetjie	3	1.50	120	1.40	2.25	149	68.5
WC	Kommetjie	4	0.55	110	1.50	1.15	97	69.0
WC	Kommetjie	5	1.00	159	1.50	2.60	167	60.0
WC	Kommetjie	6	2.30	149	2.00	NA	146	73.0
WC	Kommetjie	7	1.60	107	1.75	2.90	161	63.0
WC	Kommetjie	8	0.65	104	2.00	0.75	110	51.0
WC	Kommetjie	10	0.95	111	1.25	1.60	136	56.0
WC	Kommetjie	11	2.30	178	2.50	4.20	176	76.0
FB	Bordjiesif North	1	1.75	145	1.00	0.75	82	40.0

26.4 TOOLS AND PRACTICE

Reproducibility relies mainly on:

- a stable project structure;
- data stored in known subdirectories such as `data/BCB744/`;
- Quarto source files (`.qmd`);
- R code embedded directly in those source files;
- rendered outputs in `_site/` and cached or frozen outputs in `_freeze/`.

The practical habit to build is, if a figure, table, or result appears in the report, there should be a clear path back to the code and data that generated it.

26.5 EXAMPLE 1: A REPRODUCIBLE QUARTO WORKFLOW BUILT FROM PROJECT FILES

26.5.1 Example Dataset

We use the `laminaria.csv` dataset once more because it allows us to demonstrate a complete mini-workflow from source data to rendered output inside the project itself.

```
kelp <- read_csv(here::here("data", "BCB744", "laminaria.csv"),
  show_col_types = FALSE)
gt(head(kelp, 10))
```

TABLE 26.2. Grouped summary statistics for the Laminaria dataset, generated directly from code in the chapter source.

region	n	mean_stipe	mean_blade	sd_blade
FB	100	96.15	135.420	21.81843
WC	40	149.05	148.175	26.56476

```
kelp_summary <- kelp %>%
  group_by(region) %>%
  summarise(
    n = n(),
    mean_stipe = mean(stipe_length),
    mean_blade = mean(blade_length),
    sd_blade = sd(blade_length),
    .groups = "drop"
  )

gt(kelp_summary)

ggplot(kelp, aes(x = stipe_length, y = blade_length, colour = region)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "Stipe length (cm)",
    y = "Blade length (cm)"
  )
```

26.5.2 Do an Exploratory Data Analysis (EDA)

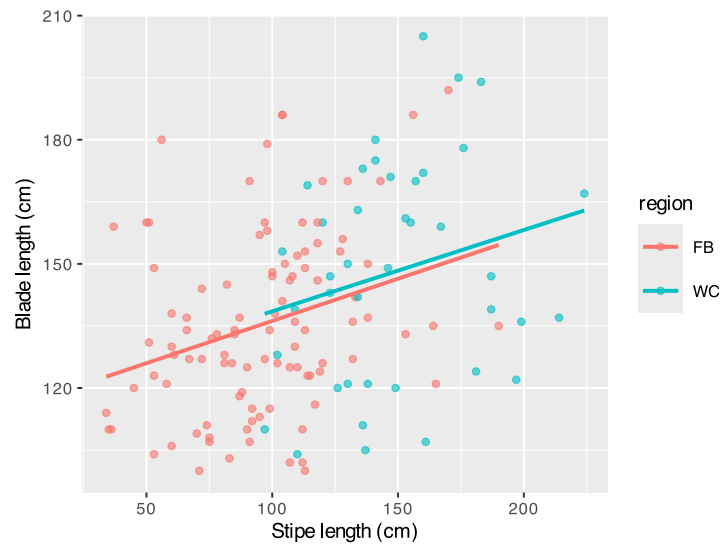


FIGURE 26.1. A reproducibly generated figure showing blade length as a function of stipe length in the Laminaria dataset.

Even this small example already illustrates that the table and figure are not separate objects made by hand. The figure in Figure 26.1 is generated directly from the dataset by code embedded in the document.

Similarly, you can also write the Introduction, Methods, Discussion, and Conclusion, *i.e.*, all the textual material that comprises the report or paper. I don't show those here, but the principle is the same. All of it is contained in the *same* Quarto document that accomplishes and reports the analysis.

26.5.3 State the Workflow Question

The workflow question is not a hypothesis test or even a research question, although it certainly guides the logical steps needed to complete the analysis and write-up. It is:

Can the reported table and figure be regenerated directly from the project data and source code without manual reconstruction?

In a good workflow, the answer should always be yes. In fact, an excellent workflow may accommodate the entire report or article, as I have already pointed out.

TABLE 26.3. Minimal project components needed to regenerate the chapter from source.

component	path
data source	data/BCB744/laminaria.csv
chapter source	BCB744/basic_stats/26-reproducible-workflow.qmd
rendered output	_site/BCB744/basic_stats/26-reproducible-workflow.html
shared styling	styles/styles.css

26.5.4 *Generate the Outputs*

In a Quarto document, the analysis, output, and prose remain connected because the code that generates the result is part of the source document itself.

The practical sequence is:

1. import the data from a stable relative path;
2. generate the summary table and figure from code;
3. write the interpretation next to the code that produced the output;
4. render the document with Quarto.

Steps 1-3 may also be accompanied by justifications of reasoned decisions or our acknowledgement of any assumptions made. So, a reproducible workflow can double as a research notebook of sorts. It can be read by you in the future, or shared with colleagues.

The render step is a command such as:

```
quarto render BCB744/basic_stats/26-reproducible-workflow.qmd
```

That one command rebuilds the chapter from source. If the data or code change, the outputs change with them. In fact, it created the document (website page) you are reading right now.

26.5.5 *Check the Workflow*

The most useful diagnostic questions in a reproducible workflow are:

- are the data paths explicit and stable;
- can the document be rendered from source without manual intervention;
- are the figures and tables generated in code rather than edited after export;
- can another person inspect the source and understand what was done.

For this project, a minimal reproducible structure looks like:

The purpose of these components, woven together in the Quarto file, is that the workflow remains legible from source to output.

26.5.6 *What This Means for Us*

The result of a reproducible workflow is the scientific conclusion and, as importantly, the fact that the conclusion, table, and figure remain traceable to the same source analysis.

In this mini-example, anyone with the project can:

- locate the source dataset;
 - inspect the code used to create the grouped summary and figure;
 - rerender the chapter;
 - verify that the reported output matches the source.
- That makes revision safer, collaboration easier, and error detection more likely.

26.6 COMMON FAILURES

The most common failures of reproducibility are usually workflow failures rather than advanced technical problems:

- doing the analysis interactively without saving code;
- editing figures by hand after export;
- keeping the final report separate from the analysis that generated it;
- changing data or exclusions without documenting those changes;
- using paths that only work on one computer and are not stable within the project.

26.7 SUMMARY

- Reproducibility links data, code, output, and interpretation.
- In a Quarto-based workflow, the report can be regenerated from source rather than rebuilt manually.
- A reproducible figure or table is more scientifically valuable than a hand-edited one with unclear provenance.
- Good workflow is therefore part of good statistics, not an optional final step.

The statistical workflow now comes full circle. A biological question leads to a design, the design produces data, the data are explored and modelled, the results are interpreted, and the whole chain is documented so that someone else can inspect and rerun it. Reproducibility is what keeps those pieces joined.

Graham MH (2003) Confronting multicollinearity in ecological multiple regression. *Ecology* 84:2809–2815.

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187–211.

Smit AJ, Bolton JJ, Anderson RJ (2017) Seaweeds in two oceans: Beta-diversity. *Frontiers in Marine Science* 4:404.