Chapter 3

Linear Regression

Linear models are frequently used statistical tools that all biologists should know. They describe and quantify relationships between variables and are widely employed to predict the value of a dependent variable (or response variable, *Y*) based on the values of one or more independent variables (or predictor variables, *X*). A linear model is an equation where the relationship between the dependent variable and the independent variables is linear in the parameters (though not necessarily in the variables themselves), allowing us to predict the dependent variable from the predictors. In statistics, models are mathematical representations or descriptions of real-world processes or systems. They offer idealised and simplified representations of reality and capture the essential features and relationships we find interesting.

Regression analysis is a statistical technique used to estimate the parameters of the model that best describes the relationship between a dependent variable and one or more independent variables. The primary goal of regression analysis is to fit the model to the observed data and offer insights into the strength and nature of the relationships between variables.

One of the simplest forms of linear models is the **simple linear model**, which is the topic of this chapter. A simple linear model estimates model parameters through the process of simple linear regression (SLR). SLR involves a single independent variable and is often applied when the independent variable is hypothesised to causally influence the dependent variable. However, a causal relationship is not a strict requirement. The primary goal of SLR may simply be to derive a formula (model) that predicts the values of the dependent variable based on the independent variable, regardless of whether a causal relationship exists between them.

SLR serves as a foundational regression technique that extends to more complex forms, including **polynomial regression** (Chapter 4), **multiple linear regression (MLR)** (Chapter 5), and **generalised linear models (GLMs)** (Chapter 6). Polynomial regression includes polynomial terms (higher powers of the independent variable, like X^2, X^3, etc.) to model curvilinear relationships, while MLR involves multiple independent variables to describe more complex relationships where the dependent variable is influenced by several predictors simultaneously. GLMs further extend these concepts to handle various types of dependent variables (besides responses drawn from the normal distribution) and relationships (e.g. logistic).

In cases where prediction is not the primary objective, and causation is neither expected nor implied, but one variable exhibits a systematic change with another, **correlation analysis** (Chapter 2)

is a more appropriate technique.

The terminology surrounding linear models and linear regression can sometimes be confusing because we often use terms like 'linear model,' 'linear regression,' and 'least squares regression' interchangeably. But 'linear model' is a broader term that encompasses various types of linear relationships, including simple linear models, multiple linear models, polynomial models, and GLMs. In this section, you will learn about simple linear models and regression analysis, which will provide you with the foundational knowledge to understand more complex linear models and regression techniques.

3.1 Simple Linear Regression

Linear models help us answer questions like:

- How does body mass change with age in a particular species?
- Does the number of offspring depend on the amount of food available?
- How does a species' geographic distribution change with temperature?

By assuming a linear relationship between variables, these models provide a clear and interpretable way to quantify and predict biological outcomes. For example, should a linear model describe the relationship between body mass (g) and age (years), we can predict the body mass of a particular species of fish would increase by 230 g for every additional year of age up to the age of five years (however, please see the von Bertalanffy model in Chapter 7.6).

The simple linear model is given by:

$$Y_i = \beta \cdot X_i + \alpha + \epsilon \tag{3.1}$$

Where:

- Y_i is the i-th measurement of the dependent variable,
- X_i is the i-th measurement of the independent variable,
- α is the intercept (the value of Y when $X = 0$),
- β is the slope (the change in Y for a one-unit change in X), and
- ϵ is the error term (residual; see box 'The residuals, ϵ_i').

The residuals, ϵ_i

In most regression models, such as linear regressions and those discussed in Chapter 7, we assume that the residuals are *independent and identically distributed (i.i.d.)*. This implies that each residual ϵ_i is drawn from the same probability distribution and that they are mutually independent. When the residuals follow a normal distribution, this can be expressed as $\epsilon_i \sim N(0, \sigma^2)$, where:

- ϵ_i represents the residual for the i-th observation,
- $N(0, \sigma^2)$ denotes a normal distribution with a mean of 0 and a variance of σ^2.

The requirement of a zero mean for residuals implies that, on average, the model's predictions neither systematically overestimate nor underestimate the true values. The constant variance assumption ensures that the spread or dispersion of residuals around the mean remains consistent across all levels of the predictor variables. This ensures that the model's accuracy is uniform across the range of data.

The requirement for independence indicates that the residual for any given observation is not influenced by or correlated with the residuals of other observations. It also means that the residual for an observation does not depend on the order in which the observations were collected (i.e. no serial correlation or auto-correlation). Independence ensures that each data point contributes unique information to the model and prevents any systematic patterns from influencing the estimates of the model's parameters.
Violation of any of these assumptions could lead to biased or inefficient parameter estimates.

3.2 Nature of the Data

The experimenter must ensure the following key requirements for a simple linear regression:

1. **Causality:** There should be a theoretical or philosophical basis for expecting a causal relationship, where the independent variable (X) influences or determines the dependent variable (Y).[1] It is assumed that changes in X cause changes in Y.

2. **Independence of Observations:**

    - The observations or measured values of Y must be independent of each other. For each value of X, there should be only one corresponding value of Y, or if there are replicate Y values, they must be statistically independent and not influence each other.
    - The observations of Y must also be independently across the range of X values. This means that the value of Y at one point should not influence the value of Y at another point.[2]

3. **Independent Variable Scale:** The independent variable (X) should be measured on a continuous scale, such as integers, real numbers, intervals, or ratios.

4. **Dependent Variable Scale:** Similarly, the dependent variable (Y) should also be measured on a continuous scale, such as integers, real numbers, intervals, or ratios.[3]

What if my data are not continuous?

- If the independent variable is ordinal, use *ordinal regression*.
- If the dependent variable is ordinal, use *ordinal (logistic) regression*.

What if I have more than one independent variable?

- Use *multiple linear regression*.

Additional assumptions and requirements are discussed next in Section 3.3.

[1] The independent and dependent variables are also called the predictor and response variables, respectively. The predictor is often under the experimenter's control (in which case it is a fixed effects model), while the response is the variable predicted to respond in the manner hypothesised.

[2] If Y not independent across the range of X, use a different type of regression model, such as a linear mixed-effects model.

[3] The dependent variable can also be ordinal, but this is less common. If this is the case, use *ordinal (logistic) regression instead.

3.3 Assumptions

The following assumptions are made when performing a simple linear regression; 1-3 must be tested *after* fitting the linear model:

1. **Normality:** For each value of X, there is a corresponding normal distribution of Y values. Each value of Y is randomly sampled from this normal distribution.
2. **Homoscedasticity:** The variances of the Y distributions corresponding to each X value should be approximately equal.
3. **Linearity:** There exists a linear relationship between the variables Y and X.
4. **Measurement Error:** It is assumed that the measurements of X are obtained without error. However, in practical scenarios, this is rarely the case. Therefore, we assume any measurement error in X to be negligible.

See Section 3.8 for more information about how to proceed when assumptions 1-3 are violated.

3.4 Outliers and Their Impact on Simple Linear Regression

In simple linear regression, outliers can have significant detrimental effects on the analysis and the reliability of the results. Outliers are data points that deviate substantially from the overall pattern or trend observed in the data, and their presence can lead to biased parameter estimates, inflated standard errors, distorted confidence and prediction intervals, violation of assumptions, and masking of underlying patterns.

Specifically, they can greatly impact the estimation of the slope and intercept due to their influence on the process of minimising the sum of squared residuals. Their presence can increase the standard errors of the regression coefficients, making it harder to detect significant relationships between the independent and dependent variables. Furthermore, the inclusion of outliers in the dataset can distort the calculation of confidence and prediction intervals for individual observations, preventing accurate inference and prediction. Their presence may also lead to violations of the assumptions of linear regression, such as the normality of residuals and the constant variance of errors (homoscedasticity). Lastly, extreme outliers can mask underlying patterns or relationships in the data and hinder our ability to discern the true nature of the associations between variables.

3.5 R Function

The `lm()` function in R is used to fit linear models. It can be used to carry out simple linear regression, multiple linear regression, and more.

The general form of the function written in R is:

```
lm(formula, data, ...)
```

where `formula` is a symbolic description of the model to be fitted, and `data` is the data frame containing the variables. The `...` argument is used to pass additional arguments to the function (consult `?lm`). For example:

```
lm(y ~ x, data = df)                                                    ①
```

① You can read the statement `y ~ x` as "y is modelled as a function of x."

The above statement fits a simple linear regression model with `y` as the dependent variable and `x` as the independent variable. The data frame `df` contains the variables named `x` and `y`.

3.6 Example: The Penguin Dataset

The following example workflow uses the `penguin` dataset from the `palmerpenguins` package to demonstrate how to perform a simple linear regression in R. The data are in Table 3.1.

Although we can also do a correlation here, we will use a simple linear regression because we want to develop a predictive model that can be used to estimate the bill length of Adelie penguins based on their body mass—this is a permissible application of a simple linear regression even though the two variables are not assumed to be causally related.

Table 3.1: 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

3.6.1 Do an Exploratory Data Analysis (EDA)

```
dim(Adelie)
```

```
[1] 151   8
```

```
summary(Adelie)
```

```
     species          island   bill_length_mm  bill_depth_mm
 Adelie   :151   Biscoe   :44   Min.   :32.10   Min.   :15.50
 Chinstrap:  0   Dream    :56   1st Qu.:36.75   1st Qu.:17.50
 Gentoo   :  0   Torgersen:51   Median :38.80   Median :18.40
                                Mean   :38.79   Mean   :18.35
                                3rd Qu.:40.75   3rd Qu.:19.00
                                Max.   :46.00   Max.   :21.50
 flipper_length_mm  body_mass_g       sex           year
 Min.   :172       Min.   :2850   female:73   Min.   :2007
 1st Qu.:186       1st Qu.:3350   male  :73   1st Qu.:2007
 Median :190       Median :3700   NA's  : 5   Median :2008
 Mean   :190       Mean   :3701               Mean   :2008
 3rd Qu.:195       3rd Qu.:4000               3rd Qu.:2009
 Max.   :210       Max.   :4775               Max.   :2009
```

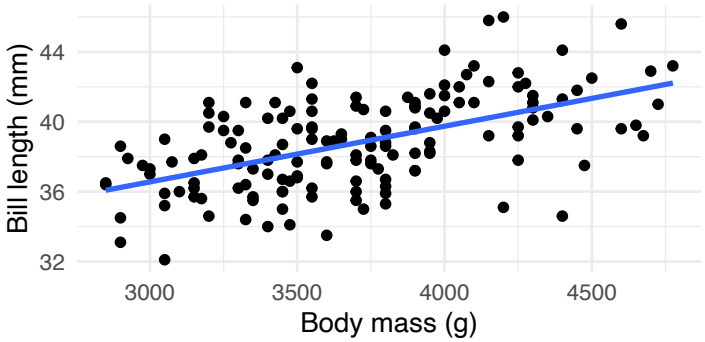

Figure 3.1: Scatter plot of the Palmer Station Adelie penguin data with a best fit line.

We see that the dataset contains 344 observations of 8 variables. We shall focus on the `body_mass_g` and `bill_length_mm` variables for this example. Importantly, the two variables are continuous, which seems to satisfy the requirements for a simple linear regression. We will also restrict this analysis to the `Adelie` penguins ($n = 152$). Is the relationship between the body mass and bill length of the penguins linear? Let's find out.

3.6.2 Create a Plot

Construct a scatter plot of the data and include a best fit straight line:

```
ggplot(Adelie,
       aes(x = body_mass_g, y = bill_length_mm)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Body mass (g)", y = "Bill length (mm)") +
  theme_minimal()
```

Although there is some scatter in the data (Figure 3.1), there appears to be a positive relationship between the body mass and bill length of the penguins. This relationship might be amenable for modelling with a linear relationship and we shall continue to explore this.

3.6.3 State the Hypothesis

- Null Hypothesis (H_0): there is no relationship between the body mass of the penguins and their bill length.
- Alternative Hypothesis (H_A): there is a relationship between the two variables.

This can be written as:

$$H_0 : \beta = 0 \tag{3.2}$$

As seen above, this hypothesis concerns the slope of the regression line, β. If the slope is zero, then there is no relationship between the two variables. Regression models also tests an hypothesis about the intercept, α, but this is less commonly reported.

3.6.4 Fit the Model

Since the assumptions of a linear regression can only be tested *after* fitting the model, we first fit the model and then test the assumptions.

```
mod1 <- lm(bill_length_mm ~ body_mass_g,
           data = Adelie)
summary(mod1)
```

```
Call:
lm(formula = bill_length_mm ~ body_mass_g, data = Adelie)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4208 -1.3690  0.1874  1.4825  5.6168

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.699e+01  1.483e+00  18.201  < 2e-16 ***
body_mass_g 3.188e-03  3.977e-04   8.015 2.95e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.234 on 149 degrees of freedom
Multiple R-squared:  0.3013,    Adjusted R-squared:  0.2966
F-statistic: 64.24 on 1 and 149 DF,  p-value: 2.955e-13
```

3.6.5 Test the Assumptions

Assumptions of normality, homoscedasticity, and linearity must be tested (Section 7.3).

We already noted that a linear model will probably be appropriate for the data (see Figure 3.1), so we proceed with the other assumptions.

To facilitate the production of the diagnostic plots, we will use the **broom** package's `augment()` function to add the residuals to the data within the original dataset (now appearing as the tidied dataset, `mod1_data`). This will allow us to create the diagnostic plots more easily, and later we can also use it to look for the presence of outliers (Section 3.6.6).

```
library(broom)

mod1_data <- augment(mod1)
```

Normality

I first check the normality assumption using one of several options (Options 1-3). Here I use the Shapiro-Wilk test, a Residual Q-Q plot, and a histogram of the residuals.

Option 1: Perform the Shapiro-Wilk test on the residuals. The Shapiro-Wilk test is useful for detecting departures from normality in small sample sizes. The hypothesis is:

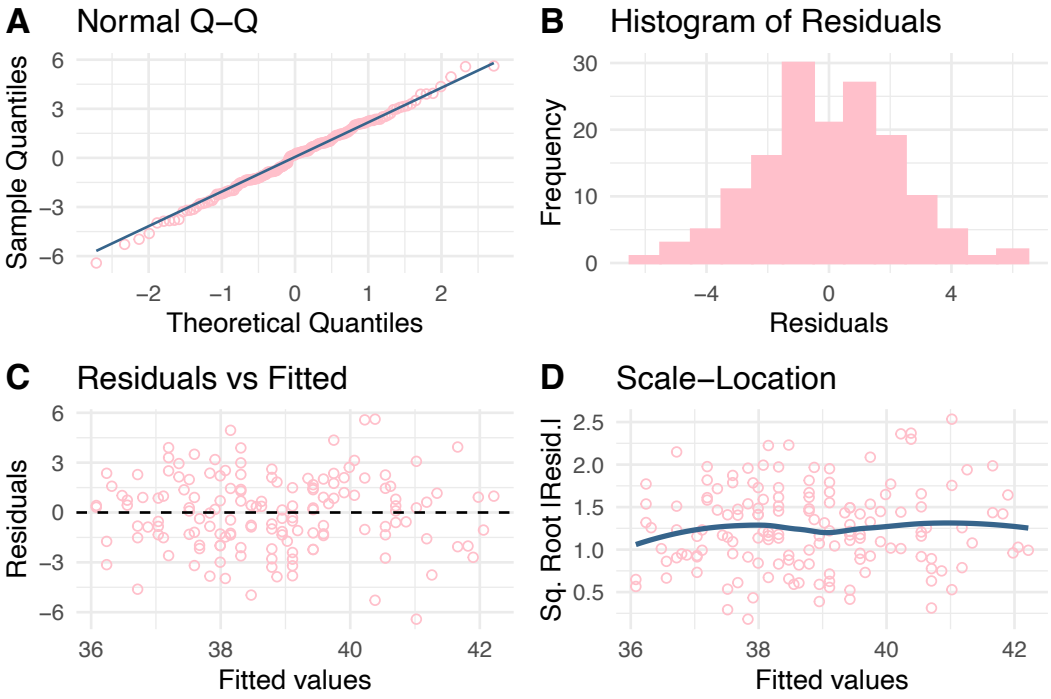

Figure 3.2: Diagnostics plots the linear regression, `mod1`, for assumption testing.

- H_0: the residuals are normally distributed.
- H_A: the residuals are not normally distributed.

```
shapiro.test(residuals(mod1))
```

```
    Shapiro-Wilk normality test

data:  residuals(mod1)
W = 0.99613, p-value = 0.9637
```

The *p*-value is greater than 0.05, so I reject the alternative hypothesis. I conclude that the residuals are normally distributed.

Option 2: Create a Residual Q-Q plot to visually assess the normality of the residuals:

The residuals are plotted against a theoretical normal distribution. The residuals fall along the line without major deviations, therefore the residuals are normally distributed (Figure 3.2 A).

Option 3: Create a histogram of the residuals to visually assess the normality of the residuals:

The histogram of the residuals appears to be normally distributed (Figure 3.2 B).

Homoscedasticity

I now examine the homoscedasticity assumption. The residuals should be approximately equal

across all values of the independent variable. There are several options.

Option 1: I will use the Breusch-Pagan test to test for homoscedasticity.

The Breusch-Pagan test is used to assess the presence of heteroscedasticity (non-constant variance) in the residuals of a regression model.

The hypothesis is:

- H_0: the residuals are homoscedastic.
- H_A: the residuals are heteroscedastic.

```
library(lmtest)
bptest(mod1)
```

```
    studentized Breusch-Pagan test

data:  mod1
BP = 1.6677, df = 1, p-value = 0.1966
```

The *p*-value is greater than 0.05, so I reject the alternative hypothesis. I conclude that the residuals are homoscedastic.

Option 2: Create a plot of the residuals against the fitted values to visually assess homoscedasticity:

The residuals are scattered evenly around zero from short through to long bill lengths, indicating that the residuals have constant variance (Figure 3.2 C).

Option 3: Create a plot of the standardised residuals against the independent variable to visually assess homoscedasticity:

The residuals are scattered evenly around zero from low through to high bill lenghts, indicating that the residuals have constant variance (Figure 3.2 D).

Other tests for homoscedasticity include the Goldfeld-Quandt (`lmtest::gqtest`) test, Levene's test (`car::leveneTest`), and others.

3.6.6 Check for outliers

How do we identify outliers in linear regression analysis? There are several approaches (see Figure 3.3):

1. **Difference in Fits (DFFITS):** DFFITS is a measure of the impact of each observation on the predicted values (fitted values) of the model. It quantifies how much the predicted values would change if an observation were removed from the analysis. DFFITS values > Threshold $= 2\sqrt{\frac{p}{n}}$ indicate observations that have a substantial impact on the predicted values and may be influential or outliers. Here, *p* is the number of parameters in the model (including the intercept, i.e. 2 in a simple linear regression) and *n* is the number of observations.
2. **Cook's Distance Plot:** Cook's distance is a measure of the influence of each observation on the estimated regression coefficients. The Cook's distance plot shows the Cook's distance values for each observation against the row numbers (or observation numbers). Points

with large Cook's distance values (typically greater than $\frac{4}{n}$) indicate observations that are potentially influential and may have a significant impact on the regression results.

3. **Residuals vs Leverage Plot:** This plot displays the standardised residuals against the leverage values (hat values) for each observation. Leverage values measure the influence of an observation on the fitted values (predicted values) of the model. The plot helps identify outliers and influential observations. Points with high leverage (typically greater than 2-3 times the average leverage) and large residuals are considered influential observations that may warrant further investigation or potential removal from the analysis.

4. **Cook's Distance vs Lev./(1-Lev.) Plot:** This plot combines information from Cook's distance and leverage values. The *x*-axis represents the leverage values divided by (1 minus the leverage values), which is a transformation that spreads out the points for better visualisation. The *y*-axis shows the Cook's distance values. This plot helps identify influential observations by considering both their impact on the regression coefficients (Cook's distance) and their influence on the fitted values (leverage). Points in the top-right corner of the plot indicate observations that are potentially influential and may require further examination or removal.

```
cooksd_thresh <- 4 / nrow(mod1_data)                                    ①
dffits_threshold <- 2 * sqrt(2 / nrow(Adelie))                          ②

mod1_data <- mod1_data %>%
  mutate(index = row_number(),
         leverage = hatvalues(mod1),
         dffits = dffits(mod1),
         colour = ifelse(.cooksd > cooksd_thresh, "black", "pink"))
```

① Calculate thresholds for Cook's distance.
② Calculate the threshold for DFFITS.

Once we have found them (Figure 3.4), what do we do with outliers? There are a few strategies:

1. **Remove them:** If the outliers are due to data entry errors or other issues, it may be appropriate to remove them from the analysis. However, this should be done with caution, as outliers may be functionally important in the dataset if they represent rare, extreme events.

2. **Robust regression methods:** When there is certainty that the outliers are part of the observed response and represent extreme but rare occurrences, robust regression techniques such as M-estimation or least trimmed squares, which are less sensitive to the presence of outliers, could be used.

3. **Transformation of variables:** Applying appropriate transformations (e.g., logarithmic, square root) to the variables can sometimes reduce the impact of outliers.

3.6.7 Interpret the Results

Now that we have tested the assumptions, we can interpret the results of the model fitted in Section 3.6.4. The slope of the regression line is 0.003188 mm/g, with a standard error of $\pm$ 0.0003977. The *p*-value is less than 0.001, so we reject the null hypothesis that the slope is zero. We conclude that there is a significant relationship between the body mass of the penguins and their bill length.

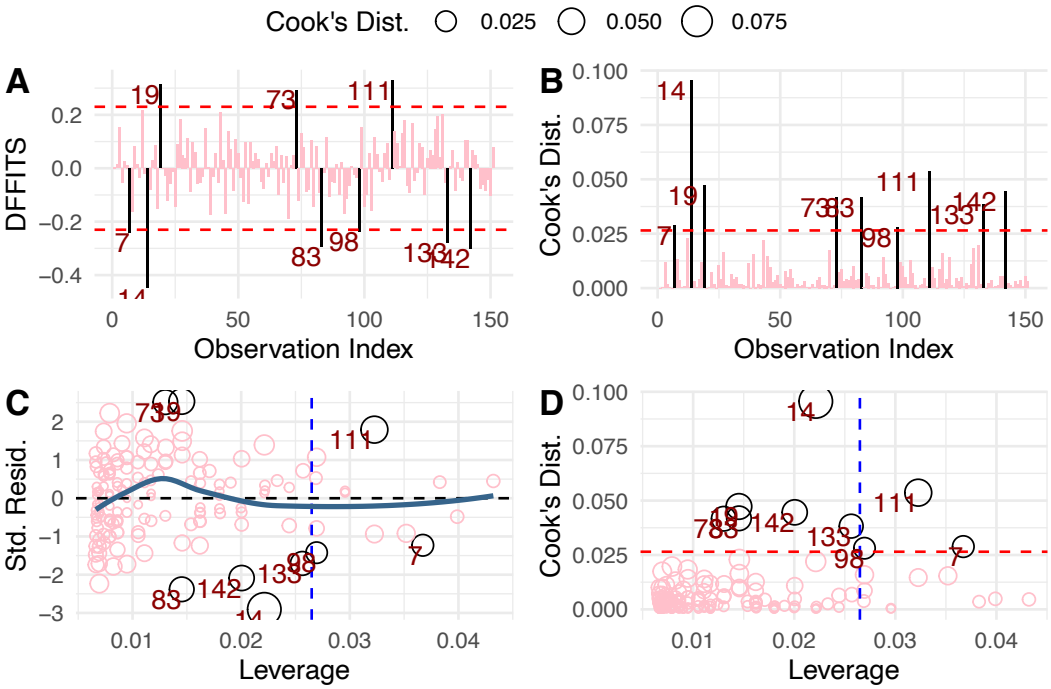

Figure 3.3: Diagnostic plots for visual inspection of outliers in the pernguin data. A) Difference in Fits (DFFITS) for `mod1`. B) Cook's distance. C) Residuals vs. leverage. D) Cook's distance vs. Lev./(1-Lev.). Outliers are identified beyond the Cook's distance threshold ($4/n$) and are plotted in black and their row numbers in dark red. The vertical dashed blue lines in C) and D) are positioned at 2 times the average leverage. The horizontal red dashed lines in B) and D) are located at the Cook's distance threshold. A) to C) are custom **ggplot2** plots corresponding to `plot(mod1, which = c(4, 5, 6))`.

The fit of the model is given by the multiple R^2 value, which is 0.3013. This means that 30.13% of the variation in bill length can be explained by body mass. The remaining ~70% is due to other factors not included in the model. The intercept of the model is 26.99 mm, with a standard error of ± 0.0003977. The intercept is the value of the dependent variable when the independent variable is zero. In this case, it is the bill length of a penguin with a body mass of zero grams, which is not a meaningful value.

The significance of the overall fit of the model can be assessed using an analysis of variance (ANOVA) test. The p-value is less than 0.001, so we reject the null hypothesis that the model does not explain a significant amount of the variation in the data against an F-value of 64.25 on 1 and 149 degrees of freedom. We conclude that the model is a good fit for the data.

3.6.8 Reporting

I provide example Methods, Results, and Discussion sections in a format more-or-less suited for inclusion in a scientific manuscript. Feel free to use it as a template and edit it as necessary to describe your study.

Figure 3.4: Plot of the linear regression resulting from `mod1` with the outliers identified using Cook's distance highlighted.

Methods

Study data

The data analysed in this study were derived from the Palmer Penguins dataset, a comprehensive collection of measurements from three penguin species (Adelie, Chinstrap, and Gentoo) collected in the Palmer Archipelago, Antarctica. The dataset includes variables species, island, bill length, bill depth, flipper length, body mass, and sex of the penguins. This dataset has been made publicly available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Statistical analysis

The primary objective of our statistical analysis was to investigate the relationship between the penguins' body mass and bill length. For this purpose, we employed a simple linear regression model to quantify the extent to which the independent variable predicts bill length.

We fitted a simple linear regression model using the `lm()` function in R version 4.4.0 (R Core Team, 2024). The model included bill length as the dependent variable, and body mass as continuous predictor. We ensured all assumptions for linear regression were assessed including linearity, independence, homoscedasticity, and normality of residuals.

After fitting the model, diagnostic plots were generated using the `plot()` function in R to visually assess the residuals for any patterns indicating potential violations of regression assumptions. Additionally, the Shapiro-Wilk test was conducted to confirm the normality of the residuals. The presence of heteroscedasticity was evaluated using the Breusch-Pagan test.

The adequacy of the model fit was judged based on the coefficient of determination (R^2), which provided insight into the variance in body mass explained by the predictors. The significance of the regression coefficients was determined using t-tests, and the overall model fit was evaluated by an F-test.

Results

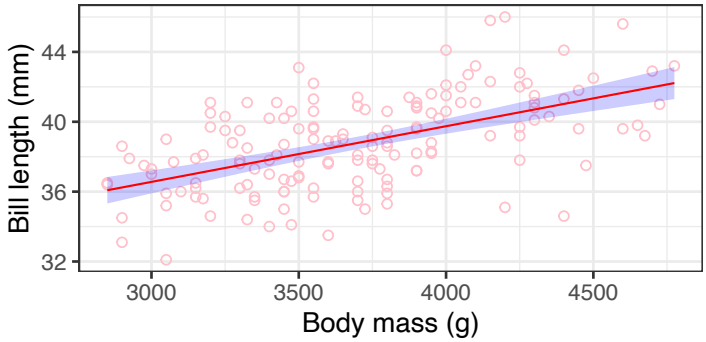

Figure 3.5: Plot of bill length as a function of body mass for Adelie penguins sampled at the Palmer Station. The straight line indicates the best fit regression line and the blue shading is the 95% confidence interval.

The regression coefficient for bill length with respect to body mass was estimated to be approximately 3.2×10^{-3} mm/g $\pm$ 3.977×10^{-4} (mean slope $\pm$ SE) ($p < 0.001$, $t = 8.015$), indicating a significant dependence of bill length on body mass (Figure 3.5).

The multiple R^2 value of the model was 0.3013, suggesting that approximately 30.13% of the variability in bill length can be accounted for by changes in body mass. This indicates that while bill length variation is notably influenced by body mass, about 69.87% of the variation is attributable to other factors not included in the model.

The overall fit of the model, assessed by an ANOVA, strongly supported the model's validity ($F = 64.25$, $p < 0.001$, d.f. = $1, 149$) and confirms that a linear model provides adequate support for predicting penguin bill length from body mass.

Discussion

In conclusion, the statistical analysis confirms a significant relationship between body mass and bill length in penguins. Although the model explains a substantial portion of the variation, future studies should consider additional variables that could account for the remaining variability in bill length. This would enhance our understanding of the morphological adaptations of penguins in their natural habitat.

3.7 Confidence and Prediction Intervals

Confidence intervals estimate the range within which the true mean of the dependent variable (Y) is likely to fall for a given value of the independent variable (X). In other words, if you were to repeat your experiment many times and calculate the mean response at a specific X value each time, the confidence interval would contain the true population mean a certain percentage of the time (e.g., 95%). Therefore, a 95% confidence interval means you can be 95% confident that the interval contains the true mean response for the population at that particular X value. It's about the average, not individual data points.

Prediction intervals, on the other hand, provide a range of Y values that are likely to contain a single new observation of the dependent variable for a given value of the independent variable

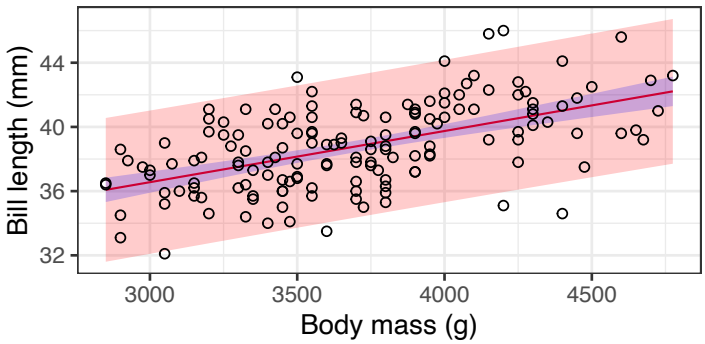

Figure 3.6: Plot of pernguin data with the confidence interval (blue) and prediction interval (pink) around the fitted values.

X. These intervals account for the variability around individual observations and are generally wider than confidence intervals because they include both the variability of the estimated mean response and the variability of individual observations around that mean. Continuing with the Adelie penguin data, the confidence and prediction intervals are shown in Figure 3.6.

```
# Predict values with confidence intervals
pred_conf <- as.data.frame(predict(mod1,
                                   newdata = Adelie,
                                   interval = "confidence"))

# Predict values with prediction intervals
pred_pred <- as.data.frame(predict(mod1,
                                   newdata = Adelie,
                                   interval = "prediction"))

# Add body mass to the data frame
results <- cbind(Adelie, pred_conf, pred_pred[,2:3])

# Rename columns for clarity
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_bw()
```

Confidence and prediction intervals are relevant for understanding the uncertainty associated

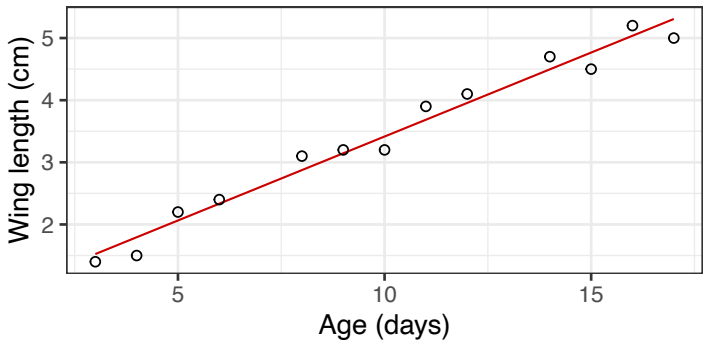

Figure 3.7: Scatter plot of the sparrow dataset with a best fit line.

with a linear regression model's predictions. While confidence intervals focus on quantifying the uncertainty around the estimated mean response, prediction intervals comprehensively assess the variability that can be expected for individual observations. We can use both when interpreting the results of a linear regression analysis.

Confidence intervals are useful when the primary interest lies in making inferences about the mean response at specific values of the independent variable(s). For instance, in a study examining the relationship between soil nutrient levels and plant biomass, confidence intervals can help determine the range of mean biomass that can be expected for a given level of soil nutrients. This information may be valuable for crop management practices, such as designing fertilisation strategies or assessing the impact of nutrient depletion on plant productivity.

Prediction intervals, on the other hand, are more relevant when the goal is to predict the value of an individual observation or to assess the range of values that future observations might take. For example, in a study investigating the relationship between ambient temperature and the growth rate of a species of fish, prediction intervals provide a range of growth rates that an individual fish might exhibit based on the observed temperature. This information is invaluable in aquaculture, for instance, where predicting individual growth patterns can inform decisions about optimal stocking densities or feed management strategies.

The relative widths of confidence and prediction intervals can provide insights into the variability in the data. If the prediction intervals are substantially wider than the confidence intervals, it may indicate a high level of variability in individual observations around the mean response, which could suggest the presence of influential factors or sources of variation that are not accounted for by the current model, such as microhabitat differences or genetic variation within the studied population.

3.8 What Do I Do When Some Assumptions Fail?

3.8.1 Failing Assumptions of Normality and Homoscedasticity

I will use the sparrow data from Zar (1999) to demonstrate what to do when the assumptions of normality and homoscedasticity are violated. I will fit a linear model to the data and then check the assumptions.

Figure 3.7 is a scatter plot of the sparrow data with a best fit line. At first glance, the linear model seems to almost perfectly describe the relationship of wing length on age. I will fit a linear model to the data and then check the assumptions.

```
mod2 <- lm(wing ~ age, data = sparrows)
summary(mod2)
```

```
Call:
lm(formula = wing ~ age, data = sparrows)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30699 -0.21538  0.06553  0.16324  0.22507

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.71309    0.14790   4.821 0.000535 ***
age          0.27023    0.01349  20.027 5.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared:  0.9733,    Adjusted R-squared:  0.9709
F-statistic: 401.1 on 1 and 11 DF,  p-value: 5.267e-10
```

Check the assumption of normality of residuals using the Shapiro-Wilk test, a histogram, and a residual Q-Q plot.

```
shapiro.test(residuals(mod2))
```

```
    Shapiro-Wilk normality test

data:  residuals(mod2)
W = 0.84542, p-value = 0.02487
```

The *p*-value for the Shapiro-Wilk test is < 0.05, indicating that the residuals are not normally distributed. The histogram and Q-Q plot of the residuals also show that the residuals are not normally distributed (Figure 3.8 and Figure 3.9). In the Residual Q-Q plot, the points deviate from the straight line, indicating non-normality—note the S-shaped curvature to the data.

```
hist(residuals(mod2))
```

```
plot(mod2, which = 2)
```

It is enough to know that the normality assumption is not met – I cannot proceed with a simple linear regression. However, let us for completeness also look at the homoscedasticity assumption. I will use the Breusch-Pagan test to check for homoscedasticity, followed by a plot of residuals against fitted values.

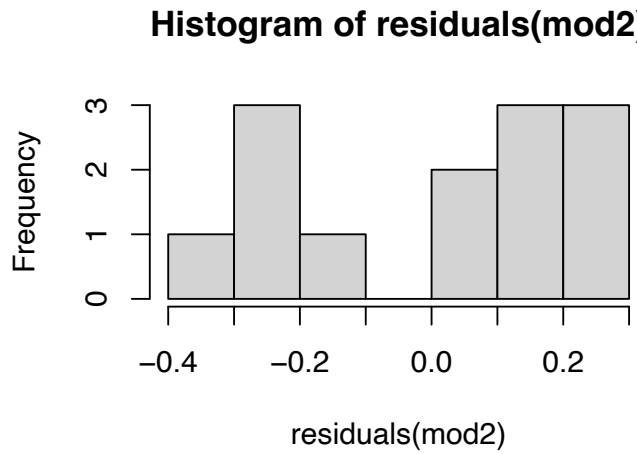

Figure 3.8: A histogram of the residual of the linear regression, `mod2`.

```
bptest(mod2)
```

```
    studentized Breusch-Pagan test

data:  mod2
BP = 1.6349, df = 1, p-value = 0.201
```

The *p*-value for the Breusch-Pagan test is > 0.05, indicating that the residuals are homoscedastic. The plot of residuals against fitted values shows gives a slightly different impression (Figure 3.10).

```
plot(mod2, which = 1)
```

The assumptions of normality and homoscedasticity are violated (it is sufficient that one or the other fails, not both). As already noted, I cannot proceed with the linear model. I will need to consider alternative models or transformations to address these issues.

When the assumptions of normality and homoscedasticity are violated, I have some options—these broadly group into transforming the data and using a non-parametric test.

Transforming the data can sometimes help attain normality and homoscedasticity. Common transformations include the logarithmic, square root, and inverse transformations. However, be cautious when interpreting the results of transformed data, as the transformed coefficients may not be directly interpretable.

I will show the Theil-Sen estimator (also known as Sen's slope estimator) as a robust non-parametric replacement for a simple linear model. It calculates the median of the slopes of all pairs of sample points to determine the overall slope of the line.

```
library(mblm)

mod3 <- mblm(wing ~ age, data = sparrows)
```

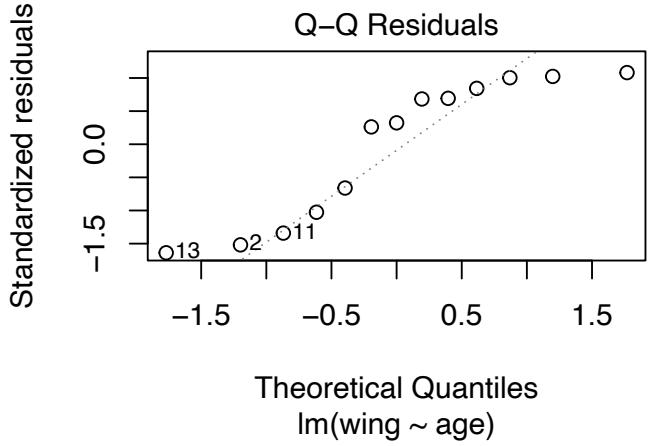

Figure 3.9: A Residual Q-Q plot of the linear regression, `mod2`.

```
summary(mod3)
```

```
Call:
mblm(formula = wing ~ age, dataframe = sparrows)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44524 -0.31190 -0.00714  0.06905  0.14048

Coefficients:
            Estimate     MAD V value Pr(>|V|)
(Intercept)  0.75000 0.18532      91 0.000244 ***
age          0.27619 0.00956      91 0.000244 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.244 on 11 degrees of freedom
```

The interpretation of the Theil-Sen estimator is similar to the simple linear regression. The Theil-Sen estimator provides a robust estimate of the slope of the relationship between age and wing length. The slope of the line is 0.28 ($\pm$ 0.19 mean absolute deviation) (V value = 91, $p < 0.001$), indicating that for each additional day of age, the wing length increases by 0.28 cm. The intercept of the line is 0.75, indicating that the wing length is ~0.8 cm when the age is 0 days.

3.8.2 My Data Do Not Display a Linear Response

In simple linear regression, the dependent variable Y is expected to exhibit a straight-line relationship with the independent variable X. However, several factors can cause deviations from a linear pattern.

Statistical assumptions underlying linear regression can affect the appearance of a linear response.

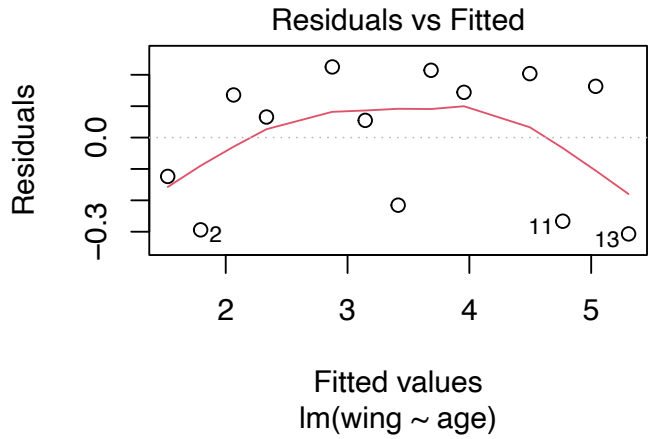

Figure 3.10: A plot of residuals against fitted values for the linear regression, `mod2`.

The normality assumption is important but primarily pertains to the residuals rather than the *Y* vs. *X* plot. A scatterplot of *Y* vs. *X* might deviate from a linear pattern due to the non-normality of the residuals or heteroscedasticity, where the variability of the residuals changes with the level of *X*. Addressing these issues and then reassessing the linearity of the relationship is a logical first step. Refer to Section 3.8 for more details on how to proceed.

Outliers in the data can significantly impact the regression line, leading to misleading results (Section 3.6.6). Measurement errors in the independent variable can also lead to biased and inconsistent estimations, which may require revisiting the data collection process to address systemic problems. Variable bias, where excluding relevant variables distorts the observed relationship, could also explain seemingly nonlinear responses. Considering multiple predictor variables in a regression model (Chapter 5) might be more appropriate in such situations.

It's important to note that simple linear regression might not be suitable for all scenarios. For instance, the dependent variable *Y* might inherently follow a different probability distribution, such as a Poisson or a binomial distribution, rather than a normal distribution. This is particularly relevant in count data or binary outcome scenarios. In such cases, other types of models like Poisson regression or logistic regression, accommodated by generalised linear models (GLM; Chapter 6), would be more appropriate.

Lastly, if the data do not exhibit a linear relationship even after addressing these issues, the relationship between the variables may really be nonlinear. This can occur when the underlying functional relationship between *X* and *Y* is better described by exponential, logarithmic, or other more complex mechanistic responses. In such cases, nonlinear regression (Chapter 7) or generalised additive models (GAM; Chapter 9) might be necessary to describe the relationship between the variables accurately.

Chapter 5

Multiple Linear Regression

In Section 3.1 we have seen how to model the relationship between two variables using simple linear regression (SLR). However, in ecosystems, the relationship between the response variable and the explanatory variables is more complex and in many cases cannot be adequately captured by a single driver (i.e. influential or predictor variable). In such cases, multiple linear regression (MLR) can be used to model the relationship between the response variable and multiple explanatory variables.

5.1 Multiple Linear Regression

Multiple linear regression helps us answer questions such as:

- How do various environmental factors influence the population size of a species? Factors like average temperature, precipitation levels, and habitat area can be used to predict the population size of a species in a given region. Which of these factors are most important in determining the population size?
- What are the determinants of plant growth in different ecosystems? Variables such as soil nutrient content, water availability, and light exposure can help predict the growth rate of plants in various ecosystems. How do these factors interact to influence plant growth?
- How do genetic and environmental factors affect the spread of a disease in a population? The incidence of a disease might depend on factors like genetic susceptibility, exposure to pathogens, and environmental conditions (e.g., humidity and temperature). What is the relative importance of these factors in determining the spread of the disease?

Multiple linear regression extends the simple linear regression model to include several independent variables. The model is expressed as:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i \tag{5.1}$$

Where:

- Y_i is the response variable for the i-th observation,
- $X_{i1}, X_{i2}, \ldots, X_{ik}$ are the k predictor variables for the i-th observation,
- α is the intercept,

- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the k predictor variables, and
- ϵ_i is the error term for the i-th observation (the residuals).

When including a categorical variable in a multiple linear regression model, dummy (indicator) variables are used to represent the different levels of the categorical variable. Let's assume we have a categorical variable C with three levels: C_1, C_2, and C_3. We can represent this categorical variable using two dummy variables:

- D_1: Equals 1 if $C = C_2$, 0 otherwise.
- D_2: Equals 1 if $C = C_3$, 0 otherwise.

C_1 is considered the reference category and does not get a dummy variable. This way, we avoid multicollinearity (see Section 5.6.4). R's `lm()` function will automatically convert the categorical variables to dummy variables (sometimes called treatment coding). The first level of the alphabetically sorted categorical variable is taken as the reference level. See Section 5.5 for more information about how to include categorical variables in a multiple linear regression model. At the end of the chapter you'll find alternative ways to assess categorical variables in a multiple linear regression model (Section 5.9).

Assume we also have k continuous predictors $X_1, X_2, \ldots, X_k$. The multiple linear regression model with these predictors and the categorical variable can be expressed as:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i \tag{5.2}$$

Where:

- Y_i is the dependent variable for observation i.
- α is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the continuous independent variables $X_{i1}, X_{i2}, \ldots, X_{ik}$.
- D_{i1} and D_{i2} are the dummy variables for the categorical predictor C.
- γ_1 and γ_2 are the coefficients for the dummy variables, representing the effect of levels C_2 and C_3 relative to the reference level C_1.
- ϵ_i is the error term for observation i.

5.2 Nature of the Data

You are referred to the discussion in simple linear regression (Section 3.1). The only added consideration is that the data should be multivariate, i.e., it should contain more than one predictor variable. The predictor variables are generally continuous, but there may also be categorical variables.

5.3 Assumptions

Basically, this is as already discussed in simple linear regression (Section 3.1)—in multiple linear regression, the same assumptions apply to the response relative to each of the predictor variables. In Section 5.6.7 I will assess the assumptions in an example dataset. An additional consideration is that the predictors must not be highly correlated with each other (multicollinearity) (see Section 5.6.4).

5.4 Outliers

Again, this is as discussed in simple linear regression (Section 3.1). In multiple linear regression, the same considerations apply to the response relative to each of the predictor variables.

5.5 R Function

The `lm()` function in R is used to fit a multiple linear regression model. The syntax is similar to that of the `lm()` function used for simple linear regression, but with multiple predictor variables. The function takes the basic form:

```
lm(formula, data)
```

For a multiple linear regression with only continuous predictor variables (as in Equation 5.1), the formula is:

```
lm(response ~ predictor1 + predictor2 + ... + predictorN,
   data = dataset)
```

Interaction effects are implemented by including the product of two variables in the formula. For example, to include an interaction between `predictor1` and `predictor2`, we can use:

```
lm(response ~ predictor1 * predictor2, data = dataset)
```

When we have both continuous and categorical predictor variables (Equation 5.2), the formula is:

```
lm(response ~ continuous_predictor1 + continuous_predictor2 + ...
   + continuous_predictorN + factor(categorical_predictor1) +
     factor(categorical_predictor2) + ...
   + factor(categorical_predictorM),
   data = dataset)
```

5.6 Example 1: The Seaweed Dataset

Load some data produced in the analysis by Smit et al. (2017). Please refer to the chapter Deep Dive into Gradients on Tangled Bank for the data description.

This dataset is suitable for a multiple linear regression because it has continuous response variables ($\beta_{sør}$, β_{sim}, and β_{sne}, the Sørenesen dissimilarity, the turnover component of β-diversity, and the nestedness-resultant component of β-diversity, respectively), continuous predictor variables (the mean climatological temperature for August, the mean climatological temperature for the year, the temperature range for February and August, and the SD of February and August), and a categorical variable (the bioregional classification of the samples).

```
sw <- read.csv("data/spp_df2.csv")
rbind(head(sw, 3), tail(sw, 3))[,-1]
```

```
       dist  bio    augMean   febRange      febSD     augSD    annMean
1     0.000  BMP 0.00000000 0.00000000 0.00000000 0.0000000 0.00000000
2    51.138  BMP 0.05741369 0.09884404 0.16295271 0.3132800 0.01501846
3   104.443  BMP 0.15043904 0.34887754 0.09934163 0.4188239 0.02602247
968 102.649 ECTZ 0.41496099 0.11330069 0.24304493 0.7538546 0.52278161
969  49.912 ECTZ 0.17194242 0.05756093 0.18196664 0.3604341 0.24445006
970   0.000 ECTZ 0.00000000 0.00000000 0.00000000 0.0000000 0.00000000
              Y        Y1          Y2
1   0.000000000 0.0000000 0.000000000
2   0.003610108 0.0000000 0.003610108
3   0.003610108 0.0000000 0.003610108
968 0.198728140 0.1948882 0.003839961
969 0.069337442 0.0443038 0.025033645
970 0.000000000 0.0000000 0.000000000
```

We will do a multiple linear regression analysis to understand the relationship between some of the environmental variables and the seaweed species. Specifically, we will consider only the variables `augMean`, `febRange`, `febSD`, `augSD`, and `annMean` as predictors of the species composition as measured by $\beta_{sør}$ (`Y` in the data file).

The model, which we will call `full_mod1` below, can be stated formally as Equation 5.3:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon \tag{5.3}$$

Where:

- Y is the response variable, the mean Sørensen dissimilarity,
- the predictors X_1, X_2, X_3, X_4, and X_5 correspond to `augMean`, `febRange`, `febSD`, `augSD`, and `annMean`, respectively, and
- ϵ is the error term.

But before we jump into multiple linear regression, let's warm up by first fitting some simple linear regressions.

5.6.1 Simple Linear Models

For interest sake, let's fit simple linear models for each of the predictors against the response variable. Let's look at relationships between the continuous predictors and the response in the East Coast Transition Zone (`ECTZ`), ignoring the other bioregions for now. We will first fit the simple linear models and then create scatter plots of the response variable $\beta_{sør}$ against each of the predictor variables. To these plots, we will add a best fit (regression) lines.

```
sw_ectz <- sw |> filter(bio == "ECTZ")

predictors <- c("augMean", "febRange", "febSD", "augSD", "annMean")

# Fit models using purrr::map and store in a list
```

```
models <- map(predictors, ~ lm(as.formula(paste("Y ~", .x)),
                               data = sw_ectz))

names(models) <- predictors

model_summaries <- map(models, summary)
model_summaries
```

```
$augMean

Call:
lm(formula = as.formula(paste("Y ~", .x)), data = sw_ectz)

Residuals:
      Min        1Q    Median        3Q       Max
-0.180961 -0.059317 -0.008346  0.045695  0.192444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.060104   0.007359   8.168 1.01e-14 ***
augMean     0.346011   0.010899  31.748  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07721 on 287 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7776
F-statistic:  1008 on 1 and 287 DF,  p-value: < 2.2e-16


$febRange

Call:
lm(formula = as.formula(paste("Y ~", .x)), data = sw_ectz)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21744 -0.08311 -0.01543  0.07536  0.25699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.092722   0.009638   9.621   <2e-16 ***
febRange    0.181546   0.008897  20.405   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1048 on 287 degrees of freedom
Multiple R-squared:  0.592, Adjusted R-squared:  0.5905
F-statistic: 416.4 on 1 and 287 DF,  p-value: < 2.2e-16
```

```
$febSD

Call:
lm(formula = as.formula(paste("Y ~", .x)), data = sw_ectz)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24267 -0.10709 -0.02587  0.08888  0.39171

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12018    0.01168   10.29   <2e-16 ***
febSD        0.17166    0.01245   13.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1272 on 287 degrees of freedom
Multiple R-squared:  0.3985,    Adjusted R-squared:  0.3964
F-statistic: 190.1 on 1 and 287 DF,  p-value: < 2.2e-16


$augSD

Call:
lm(formula = as.formula(paste("Y ~", .x)), data = sw_ectz)

Residuals:
      Min        1Q    Median        3Q       Max
-0.307683 -0.111051 -0.003922  0.086322  0.308041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12781    0.01231   10.38   <2e-16 ***
augSD        0.08793    0.00720   12.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.133 on 287 degrees of freedom
Multiple R-squared:  0.3419,    Adjusted R-squared:  0.3396
F-statistic: 149.1 on 1 and 287 DF,  p-value: < 2.2e-16


$annMean

Call:
lm(formula = as.formula(paste("Y ~", .x)), data = sw_ectz)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.144251 -0.051607 -0.005023  0.045095  0.145173

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.053883   0.006309   8.541 7.94e-16 ***
annMean     0.332150   0.008667  38.325  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0663 on 287 degrees of freedom
Multiple R-squared:  0.8365,    Adjusted R-squared:  0.836
F-statistic:  1469 on 1 and 287 DF,  p-value: < 2.2e-16
```

The individual models show that, for each predictor, the estimate of the coefficients (for slope) and the test for the overall hypothesis are both significant ($p < 0.05$ in all cases; refer to the model output). All the predictor variables are therefore good predictors of the structure of seaweed species composition along.

```
# Create individual plots for each predictor
plts1 <- map(predictors, function(predictor) {
  ggplot(sw_ectz, aes_string(x = 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_bw()
})

# Name the list elements for easy reference
names(plts1) <- predictors

ggpubr::ggarrange(plotlist = plts1, ncol = 2,
                  nrow = 3, labels = "AUTO")
```

Figure 5.1 is a series of scatter plots showing the relationship between the response variable $\beta_{sør}$ and each of the predictor variables. The blue line represents the linear regression fitted to the data. We see that the relationship between the response variable and each of the predictors is positive and linear. Each of the models are significant, as indicated by the p-values in the model summaries. These simple models do not tell us how some predictors might act together to influence the response variable.

To consider combined effects and interactions between predictor variables, we must explore multiple linear regression models that include all the predictors. Multiple regression will give us a more integrated understanding of how various environmental variables jointly influence species composition along the coast. In doing so, we can control for confounding variables, improve model fit, deal with multicollinearity, test for interaction effects, and enhance predictive power.

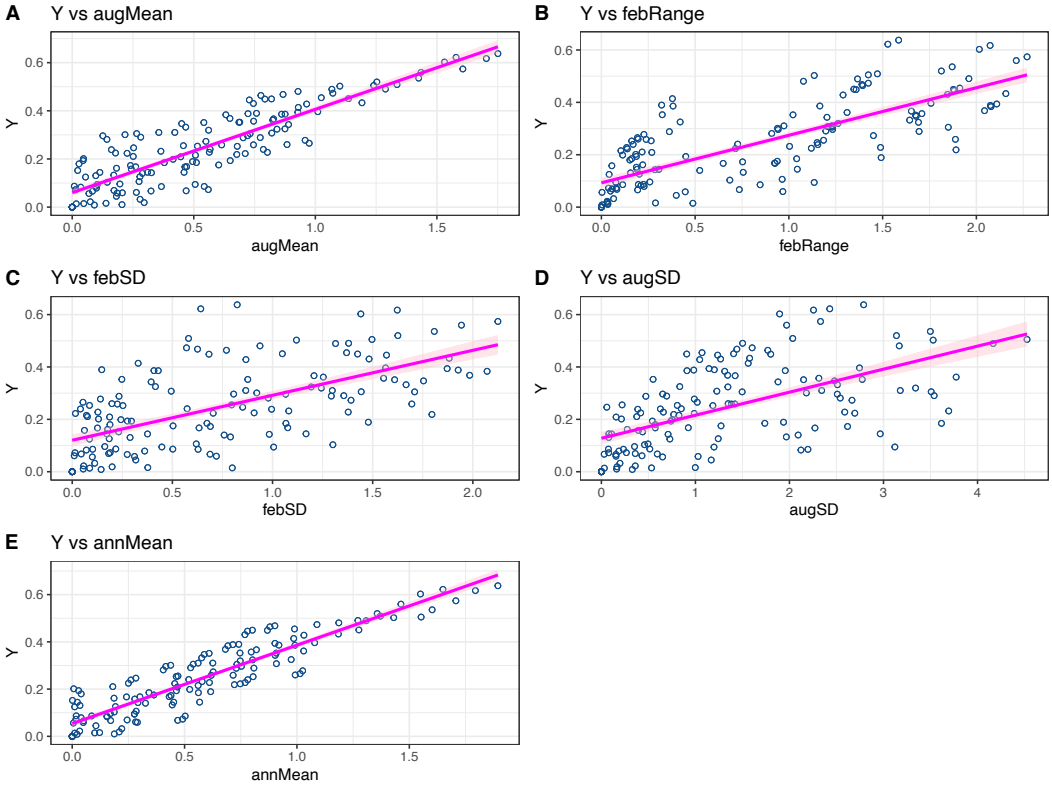

Figure 5.1: Individual simple linear regressions fitted to the variables `augMean`, `febRange`, `febSD`, `augSD`, and `annMean` as predictors of the seaweed species composition as measured by the Sørensen dissimilarity, `Y`.

We will fit this multiple regression model next.

5.6.2 State the Hypotheses for a Multiple Linear Regression

As with all inferential statistics, we need to consider the hypotheses when performing multiple linear regression.

The null hypothesis (H_0) states that there is no significant relationship between the Sørensen diversity index and any of the the climatological variables entered into the model, implying that the coefficients for all predictors are equal to zero. The alternative hypothesis (H_A), on the other hand, states that there is a significant relationship between the Sørensen diversity index and the climatological variables, positing that at least one of the coefficients is not equal to zero.

The hypotheses can be divided into two kinds: those dealing with the main effects and the one assessing the overall model stated in Equation 5.3.

Main effects hypotheses

The main effects hypotheses test, for each predictor, X_i, if the predictor has a significant effect on

the response variable Y.

H_0: There is no linear relationship between the environmental variables (`augMean`, `febRange`, `febSD`, `augSD`, and `annMean`) and the community composition as measured by $\beta_{sør}$ (in `Y`). Formally, for each predictor variable X_i:

- $H_0 : \beta_i = 0$ for $i = 1, 2, 3, 4, 5$

Where β_i are the coefficients of the predictors in the multiple linear regression model.

H_A: There is a linear relationship between the environmental variables (`augMean`, `febRange`, `febSD`, `augSD`, and `annMean`) and the species composition as measured by $\beta_{sør}$:

- $H_A : \beta_i \neq 0$ for $i = 1, 2, 3, 4, 5$

Overall hypothesis

In addition to testing the individual predictors, X_i, we can also test a hypothesis about the overall significance of the model (F-test), which examines whether the model as a whole explains a significant amount of variance in the response variable Y. A significant F-test would suggest that *at least one* predictor (excluding the intercept) in the model is likely to be significantly related to the response, but it requires further investigation of individual predictors and potential multicollinearity to fully understand the relationships. For the overall model hypothesis:

Null Hypothesis (H_0):

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

Alternative Hypothesis (H_A):

- $H_A : \exists \beta_i \neq 0$ for at least one i

5.6.3 Fit the Model

We fit two models:

- a full model that includes an intercept term and the five environmental variables, and
- a null model that includes only an intercept term.

The reason the null model is included is to compare the full model with a model that has no predictors. This comparison will help us determine which of the predictors are useful in explaining the response variable—we will see this in action in the forward model selection process later on (Section 5.6.5).

```
# Select only the variables that will be used in model building
sw_sub1 <- sw_ectz[, c("Y", "augMean", "febRange",
                       "febSD", "augSD", "annMean")]

# Fit the full and null models
full_mod1 <- lm(Y ~ augMean + febRange + febSD +
                  augSD + annMean, data = sw_sub1)
null_mod1 <- lm(Y ~ 1, data = sw_sub1)

# Add fitted values from the full model to the dataframe
```

```
sw_ectz$.fitted <- fitted(full_mod1)
```

5.6.4 Dealing With Multicollinearity

Some of the predictor variables may be correlated with each other and this can lead to multicollinearity. When predictor variables are highly correlated, the model may not be able to distinguish the individual effects of each predictor. Consequently, the model becomes less precise and harder to interpret due to the coefficients' inflated standard errors (Graham (2003)). One can create a plot of pairwise correlations to visually inspect the correlation structure of the predictors. I'll not do this here, but you can try it on your own.

A formal way to detect multicollinearity is to calculate the variance inflation factor (VIF) for each predictor variable. The VIF measures how much the variance of the estimated regression coefficients is increased due to multicollinearity. A VIF value greater than 5 or 10 indicates a problematic amount of multicollinearity.

```
initial_formula <- as.formula("Y ~ .")

threshold <- 10 # Define a threshold for VIF values

# Extract the names of the predictor variables
predictors <- names(vif(full_mod1))

# Iteratively remove collinear variables
while (TRUE) {
  # Calculate VIF values
  vif_values <- vif(full_mod1)
  print(vif_values) # Print VIF values for debugging
  max_vif <- max(vif_values)

  # Check if the maximum VIF is above the threshold
  if (max_vif > threshold) {
    # Find the variable with the highest VIF
    high_vif_var <- names(which.max(vif_values))
    cat("Removing variable:",
        high_vif_var,
        "with VIF:",
        max_vif,
        "\n")

    # Update the formula to exclude the high VIF variable
    updated_formula <- as.formula(paste("Y ~ . -", high_vif_var))

    # Refit the model without the high VIF variable
    full_mod1 <- lm(updated_formula, data = sw_sub1)

    # Update the environment data frame to reflect the removal
```

```
      sw_sub1 <- sw_sub1[, !(names(sw_sub1) %in% high_vif_var)]
    } else {
      break
    }
  }
```

```
  augMean   febRange     febSD     augSD   annMean
27.947767 10.806635  8.765732  2.497739 31.061900
Removing variable: annMean with VIF: 31.0619
  augMean   febRange     febSD     augSD
 2.290171 10.648752  8.637679  1.616390
Removing variable: febRange with VIF: 10.64875
 augMean    febSD    augSD
1.423601 1.674397 1.585055
```

5.6.5 Perform Forward Selection

It might be that not all of the variables included in the full model are necessary to explain the response variable. We can use a stepwise regression to select the best combination (subset) of predictors that best explains the response variable. To do this, we will use the `stepAIC` function that lives in the `MASS` package.

`stepAIC()` works by starting with the null model and then adding predictors one by one, selecting the one that improves the model the most as seen in the reduction of the AIC values along the way. This process continues until no more predictors can be added to improve the model (i.e. to further reduce the AIC). Progress is tracked as the function runs.

```
  # Perform forward selection
  mod1 <- stepAIC(null_mod1,
                  scope = list(lower = null_mod1, upper = full_mod1),
                  direction = "forward")
```

```
Start:  AIC=-1044.97
Y ~ 1

          Df Sum of Sq    RSS     AIC
+ augMean  1    6.0084 1.7108 -1478.4
+ febSD    1    3.0759 4.6433 -1189.9
+ augSD    1    2.6394 5.0797 -1163.9
<none>                 7.7192 -1045.0

Step:  AIC=-1478.41
Y ~ augMean

        Df Sum of Sq    RSS     AIC
+ febSD  1   0.36036 1.3504 -1544.8
+ augSD  1   0.31243 1.3984 -1534.7
<none>               1.7108 -1478.4
```

```
Step:  AIC=-1544.77
Y ~ augMean + febSD

        Df Sum of Sq    RSS     AIC
+ augSD  1   0.10568 1.2448 -1566.3
<none>               1.3504 -1544.8

Step:  AIC=-1566.32
Y ~ augMean + febSD + augSD
```

The model selection process shows that as we add more variables to the model, the AIC value decreases. We can infer from this that the multiple regression model provides a better fit that simple linear models that use the variables in isolation.

We also see that `stepAIC()` has not removed any variables from the full model. Probably one reason for failing to remove any variables is that the VIF process has already accomplished this by virtue of dealing with multicollinearity. This means that all the variables retained in `mod1` are important in explaining the response variable.

5.6.6 Added-Variable Plots (Partial Regression Plots)

Before looking at the output in more detail, I'll introduce partial regression plots as a means to examine the relationship between the response variable and each predictor variable. Although they can be calculated by hand, the **car** package provides a convenient function, `avPlots()`, to create these plots.

Added variable plots are also sometimes called 'partial regression plots' or 'individual coefficient plots.' They are used to display the relationship between a response variable and an individual predictor variable while accounting for the effect of other predictor variables in a multiple regression model (the marginal effect).

```
# Create partial regression plots
avPlots(mod1, col = "dodgerblue4", col.lines = "magenta")
```

What insights can we draw from the added-variable plots? Although there are better ways to assess the model fit, we can already make some observations about the linearity of the model or the presence of outliers. The slope of the line in an added variable plot corresponds to the regression coefficient for that predictor in the full multiple regression model. Seen in this way, it visually indicates the magnitude and direction of each predictor's effect. In Figure 5.2, the added-variable plot for `augMean` shows a tighter clustering of points around the regression line and a strong linear relationship (steep slope) with the response variable; the plots for `febSD` and `augSD`, on the other hand, show a weaker response and more scatter about the regression line. Importantly, this suggests that `augMean` has a stronger and more unique contribution to the multiple-variable model than the other two variables.

There are also insights to be made about possible multicollinearity using added-variable plots. These plots are not a definitive test for multicollinearity, but they can provide some clues. Notably, if a predictor shows a strong relationship with the response variable in a simple correlation but appears to have little relationship in the added-variable plot, it might indicate collinearity with

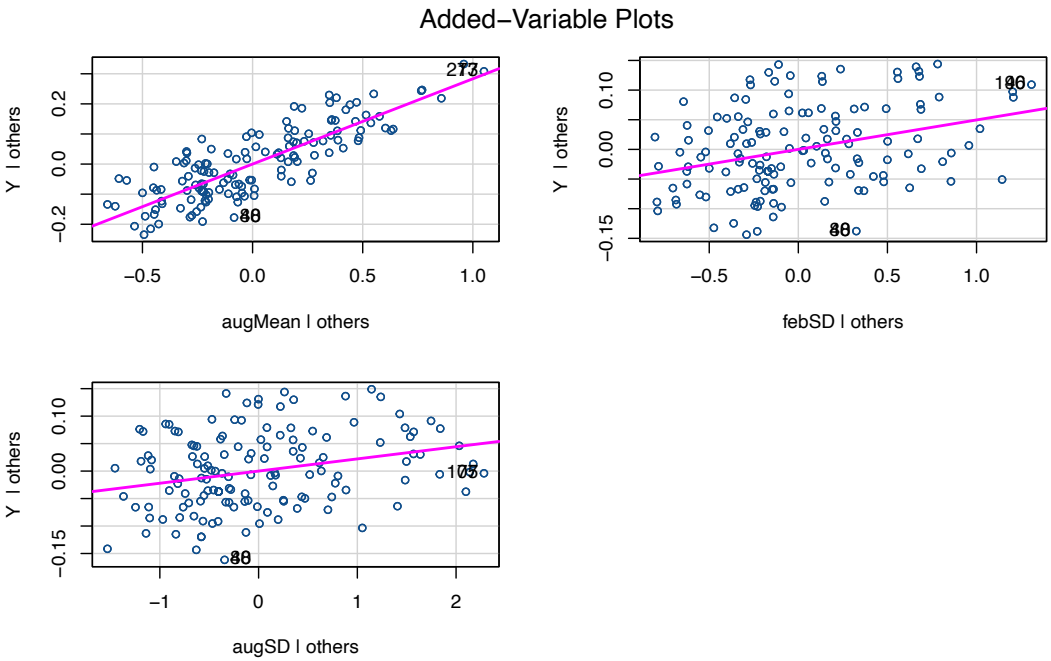

Figure 5.2: Partial regression plots for `mod1` with the selected variables `augMean`, `febSD`, and `augSD`.

other predictors. This discrepancy suggests that the predictor's effect on the response is being masked by the presence of other correlated predictors.

5.6.7 Model Diagnostics

We are back in the territory of parametric statistics, so we need to check the assumptions of the multiple linear regression model (similar to those of simple linear regression). We can do this by making the various diagnostic plots. all of them consider various aspects of the residuals, which are simply the differences between the observed and predicted values.

Diagnostic plots of final model

You have been introduced to diagnostic plots in the context of simple linear regression (Section 3.1). They are also useful in multiple linear regression. Although `plot.lm()` can easily do this, here I use `autoplot()` from the **ggfortify** package. When applied to the final model, `mod1`, the plot will in its default setting show four diagnostic plots: residuals vs. fitted values, normal Q-Q plot, scale-location plot, and residuals vs. leverage plot. Note, this is for the full model inclusive of the combined contributions of all the predictors, so we will not see separate plots for each predictor as we have seen in the added-variable plots or component plus residual plots.

```
# Generate diagnostic plots
autoplot(mod1, shape = 21, colour = "dodgerblue4",
```

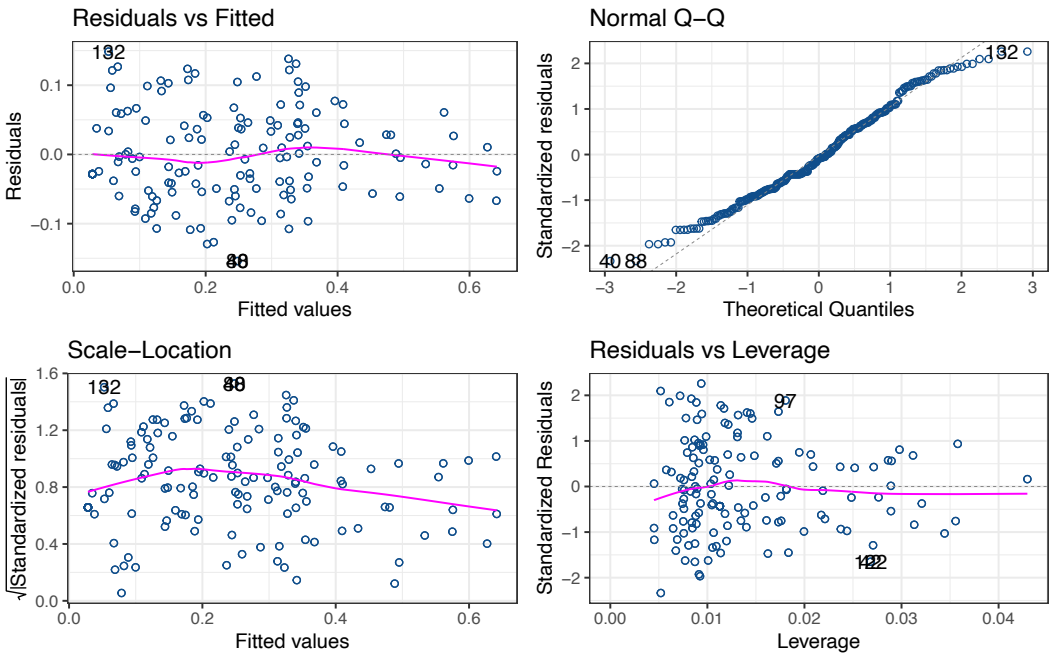

Figure 5.3: Diagnostic plots to assess the fit of the final multiple linear regression model, `mod1`.

```
        smooth.colour = "magenta") +
theme_bw()
```

Residuals vs. Fitted Values: In this plot we can assess linearity and homoscedasticity of the residuals. If the seaweed gods were with us, we'd expect the points to be randomly scattered about a horizontal line situation at zero. This would indicate that the relationship between the predictors selected by the forward selection process (`augMean`, `febSD`, and `augSD`) and the response variable (`Y`) is linear, and the variance of the residuals is constant across the range of fitted values. In this plot, there's a very slight curvature which might suggest a potential issue with the linearity assumption—it is minute and I'd suggest not worrying about it. The variance of the residuals seems to decrease slightly at higher fitted values, indicating a mild case of heteroscedasticity.

Q-Q Plot (Quantile-Quantile Plot): This plot is used to check the normality of the residuals. The points should fall approximately along a straight diagonal line if the residuals are normally distributed. Here we see that the points generally follow the line although some deviations may be seen at the tails. These deviations are not that extreme and again I don't think this is not a big concern.

Scale-Location Plot: This plot should reveal potential issues with homoscedasticity. The square root of the standardised residuals is used here to make it easier to spot patterns, so we would like the points to be randomly scattered around the horizontal red line. Here, the line slopes slightly downward and this indicates that the variance of the residuals might decrease as the fitted values increase. We can also see evidence of this in a plot of the observed values vs. the predictors in Figure 5.3.

Residuals vs. Leverage: This diagnostic highlights influential points (outliers). Points with high leverage (far from the mean of the predictors) can be expected to exert a strong influence on the regression line, tilting it in some direction. Cook's distance (indicated by the yellow line) helps identify such outliers. In our seaweed data a few points could have a high leverage, but since they don't seem to cross the Cook's distance thresholds, I doubt they are overly worrisome.

Considering that no glaring red flags were raised by the diagnostic plots, I doubt that they are severe enough to invalidate the model. However, if you cannot stand these small issues, you could i) consider transforming the predictor or response variables to address your concerns about heteroscedasticity, ii) investigate the outliers (high leverage points) to confirm if they are valid data points or errors, or iii) try robust regression methods that are less sensitive to outliers and heteroscedasticity.

Component plus residual plots

Component plus residual plots offer another way to assess the fit of the model in multiple regression models. Unlike simple linear regression where we only had one predictor variable, here we have several. So, we need to assure ourselves that there is a linear relationship between each predictor variable and the response variable (we could already see this in the added-variable plots in Section 5.6.6). We can make component plus residual plots using the `crPlots()` function in the **car** package. It displays the relationship between the response variable and each predictor variable. If the relationship is linear, the points should be randomly scattered about a best fit line and the spline (in pink in Figure 5.4) should plot nearly on top of the linear regression line.

```
# Generate component plus residual plots
crPlots(mod1, col = "dodgerblue4", col.lines = "magenta")
```

5.6.8 Understanding the Model Fit

The above model selection process has led us to the `mod1` model, which can be stated formally as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{5.4}$$

Where:

- Y: The response variable, the mean Sørensen dissimilarity.
- X_1, X_2, and X_3: The predictors corresponding to `augMean`, `febSD`, and `augSD`, respectively.
- ϵ: The error term.

We have convinced ourselves that the model is a good fit for the data, and we can proceed to examine the model's output. The fitted model can be explored in two ways: by applying the `summary()` function or by using the `anova()` function. The `summary()` function provides a detailed output of the model, while the `anova()` function provides a table of deviance values that can be used to compare models.

The model summary

```
# Summary of the selected model
summary(mod1)
```

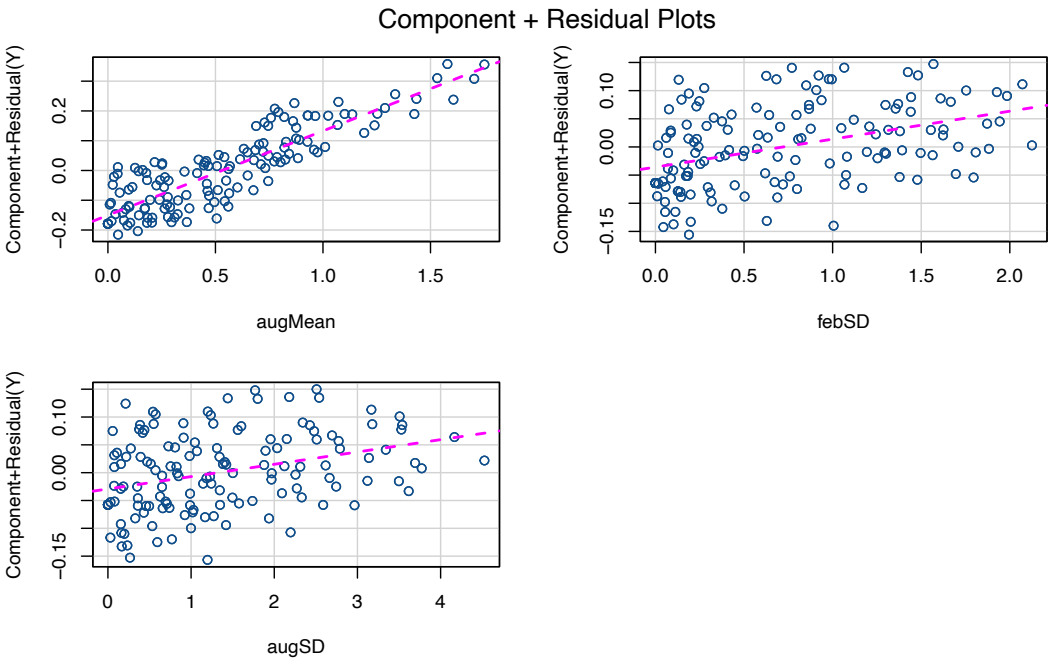

Figure 5.4: Component plus residual diagnostic plots to assess the fit of the final multiple linear regression model, `mod1`.

```
Call:
lm(formula = Y ~ augMean + febSD + augSD, data = sw_sub1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153994 -0.049229 -0.006086  0.045947  0.148579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.028365   0.007020   4.040 6.87e-05 ***
augMean     0.283335   0.011131  25.455  < 2e-16 ***
febSD       0.049639   0.008370   5.930 8.73e-09 ***
augSD       0.022150   0.004503   4.919 1.47e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06609 on 285 degrees of freedom
Multiple R-squared:  0.8387,    Adjusted R-squared:  0.837
F-statistic: 494.1 on 3 and 285 DF,  p-value: < 2.2e-16
```

The first part of the `summary()` function's output is the `Coefficients` section. This is where the main effects hypotheses are tested (this model does not have interactions—if there were, they'd appear here, too). The important components of the coefficients part of the model summary are:

- `(Intercept)`: This row provides information about where the regression line intersects the *y*-axis.
- Main Effects:
  - `augMean`, `febSD`, and `augSD`: These rows give the model coefficients associated with the slopes of the regression lines fit to those predictor variables. They indicate the rate of change in the response variable for a one-unit change in the predictor variable.
  - `Estimate`, `Std. Error`, `t value`, and `Pr(>|t|)`: These columns contain the statistics used to interpret the hypotheses about the main effects. In the `Estimate` column are the coefficients for the *y*-intercept and the main effects' slopes, and `Std. Error` indicates the variability of the estimate. The `t value` is obtained by dividing the coefficient by its standard error. The *p*-value tests the null hypothesis that the coefficient is equal to zero and significance codes are provided as a quick visual reference (their use is sometimes frowned upon by statistics purists). Using this information, we can quickly see that, for example, `augMean` has a coefficient of 0.2833 ± 0.0111 and the slope of the line is highly significant, i.e. there is a significant effect of `Y` due to the temperature gradient set up by `augMean`.

i The intercept and slope coefficients

The interpretation of the coefficients is a bit more complicated in multiple linear regression compared to what we are accustomed to in simple linear regression. Let us look at some greater detail at the intercept and the slope coefficients:
Intercept (α):) The intercept is the expected value of the response variable, Y, when all predictor variables are zero. It is not always meaningful, but it can be useful in some cases.
Slope Coefficients ($\beta_1, \beta_2, \ldots, \beta_k$): Each slope coefficient, β_j, represents the expected change in the response variable, Y, for a one-unit increase in the predictor variable, X_j, holding all other predictor variables constant. This partial effect interpretation implies that β_j accounts for the direct contribution of X_j to Y while removing the confounding effects of other predictors in the model. Figure 5.2 provides a visual representation of this concept and isolates the effect of each predictor variable on the response variable.
Therefore, in the context of our model (Equation 5.4) for this analysis, the partial interpretation is as follows:

- β_1: Represents the change in Y for a one-unit increase in X_1, holding X_2 and X_3 constant.
- β_2: Represents the change in Y for a one-unit increase in X_2, holding X_1 and X_3 constant.
- β_3: Represents the change in Y for a one-unit increase in X_3, holding X_1 and X_2 constant.

There are also several overall model fit statistics—it is here where you'll find the information you need to assess the hypothesis about the overall significance of the model. `Residual standard error` indicates the average distance between observed and fitted values. `Multiple R-squared` and `Adjusted R-squared` values tell us something about the model's goodness of fit. The latter adjusts for the number of predictors in the model, and is the one you must use and report in multiple linear regressions. As you also know, higher numbers approaching 1 are better, with 1 suggesting that the model perfectly captures all of the variability in the data. The `F-statistic` and its associated *p*-value test the overall significance of the model and examines whether all regression coefficients are simultaneously equal to zero. You can also use the brief overview of the residuals, but I don't find this particularly helpful—best examine the residuals in a histogram.

The ANOVA tables

```
anova(mod1)
```

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq  F value    Pr(>F)
augMean     1 6.0084  6.0084 1375.660 < 2.2e-16 ***
febSD       1 0.3604  0.3604   82.507 < 2.2e-16 ***
augSD       1 0.1057  0.1057   24.196 1.473e-06 ***
Residuals 285 1.2448  0.0044
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This function provides a sequential analysis of variance (Type I ANOVA) table for the regression model (see more about Type I ANOVA, below). As such, this function can also be used to compare nested models. Used on a single model, it gives a more interpretable breakdown of the variability in the response variable `Y` and assesses the contribution of each predictor variable in explaining this variability.

The ANOVA table firstly shows the degrees of freedom (`Df`) for each predictor variable added sequentially to the model, as well as the residuals. For each predictor, the degrees of freedom is typically 1. For the residuals, however, it represents the total number of observations minus the number of estimated parameters. The Sum of Squares (`Sum Sq`) indicates the variability in `Y` attributable to each predictor, and the mean sum of squares (`Mean Sq`) is the sum of squares divided by the degrees of freedom.

The `F value` is calculated as the ratio of the predictor's mean square to the residual mean square tests. It is used in testing the null hypothesis that the predictor has no effect on `Y`. Whether or not we accept the alternative hypothesis (reject the null) is given by the *p*-value (`Pr(>F)`) that goes with each *F*-statistic. You know how that works.

Because this is a sequential ANOVA, the amount of variance in `Y` explained by each predictor (or group of predictors) is calculated by adding the predictors to the model in sequence (as specified in the model formula). For example, the Sum of Squares for `augMean` (6.0084) represents the amount of variance explained by adding `augMean` to a model that doesn't include any predictors yet. The Sum of Squares for `febSD` 0.3604) represents the amount of variance explained by adding `febSD` to a model that already includes `augMean`—this improvement indicates that `febSD` explains some of the variance in `Y` that `augMean` doesn't.

> **i** Order in which predictors are assessed in multiple linear regression
>
> The interpretation of sequential ANOVA (Type I) is inherently dependent on the order in which predictors are entered. In `mod1` the order is first `augMean`, then `febSD`, and last comes `augSD`. This order might not be the most meaningful for interpreting the sequential sums of squares and their significance in the ANOVA table. How, then, does one decide on the order of predictors in the model?
>
> - If you have a strong theoretical or causal basis for thinking that certain predictors influence others, you can enter them in that order.
> - If you have a hierarchy of predictors based on their importance or general vs. specific

nature, you can enter them hierarchically.
- You can manually fit models with different predictor orders and compare the ANOVA tables to see how the results change. This can be time-consuming but might offer insights into the sensitivity of your conclusions to the order of entry.
- You can use automated model selection procedures, such as stepwise regression, to determine the best order of predictors. This is a more objective approach but can be criticised for being data-driven and not theory-driven.
- Use Type II or Type III ANOVAs, which are are not order-dependent and can be used to assess the significance of predictors after accounting for all other predictors in the model. However, they have their own limitations and assumptions that need to be considered.

My advice would be to have sound theoretical reasons for the order of predictors in the model.

Both ways of looking at the model fit of `mod1`—`summary()` and `anova()`—show that forward selection retained the variables `augMean`, `febSD`, and `augSD`. These three predictors should be used together to explain the response, `Y`.

Let's make a plot of the full model with all the initial predictors and the selected model with the predictors chosen by the forward selection process.

```
# Add fitted values from the selected model to the dataframe
sw_ectz$.fitted_selected <- fitted(mod1)

# Create the plot of observed vs fitted values for the selected model
ggplot(sw_ectz, aes(x = .fitted_selected, y = Y)) +
  geom_point(shape = 1, colour = "black", alpha = 1.0) +
  geom_point(aes(x = .fitted), colour = "red",
             shape = 1, alpha = 0.4) +
  geom_abline(intercept = 0, slope = 1,
              color = "blue", linetype = "dashed") +
  labs(x = "Fitted Values",
       y = "Observed Values") +
  theme_bw()
```

5.6.9 Reporting

A Results section should be written in a format sutable for inclusion in your report or publication. Present the results in a clear and concise manner, with tables and figures used to help substantiate your findings. The results should be interpreted in the context of the research question and the study design. The limitations of the analysis should also be discussed, along with any potential sources of bias or confounding. Here is an example.

Results

The model demonstrates a strong overall fit, as indicated by the high R^2 value of 0.839 and an adjusted R^2 of 0.837, suggesting that approximately 83.7% of the variance in the mean Sørensen dissimilarity is explained by the predictors `augMean`, `febSD`, and `augSD`. All predictors in the model are statistically significant, with `augMean` showing the strongest effect ($\beta_1 = 0.283$, $p <$

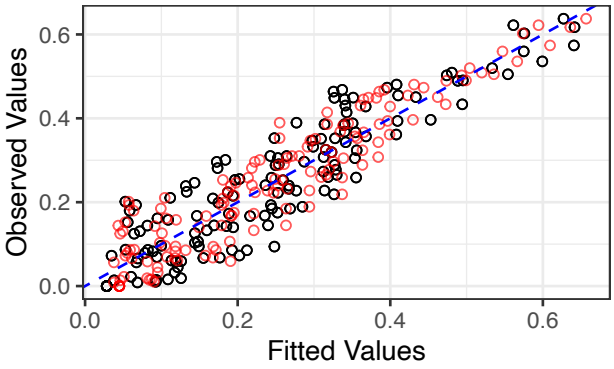

Figure 5.5: Plot of observed vs. predicted value obtained from the final multiple linear regression model (`mod`) with the selected variables `augMean`, `febSD`, and `augSD` as predictors (black points), and the initial model with also `annMean` and `febRange` (red points).

0.0001) (Figure 5.2). The predictors `febSD` and `augSD` also have significant positive relationships with the response variable ($\beta_2 = 0.050$, $p = 0.0001$; $\beta_3 = 0.022$, $p = 0.0001$). A sequential ANOVA further confirms the significance of each predictor variable in the model, with all F-values indicating that the inclusion of each predictor significantly improves the model fit ($p < 0.0001$ in all cases). Our model therefore provides clear support for the mean temperatures in August, the standard deviation of temperatures in February, and the standard deviation of temperatures in August as strong predictors of the mean Sørensen dissimilarity, with each contributing uniquely to the explanation of variability in the response variable.

5.7 Example 2: Interaction of Distance and Bioregion

Our seaweed dataset includes two additional variables that we have not yet considered. These are the continuous variable `dist` which represents the geographic distance between the seaweed samples taken along the coast of South Africa, and the categorical variable `bio` which is the bioregional classification of the seaweed samples.

These two new variables lend themselves to a few interesting questions. For example:

1. Is the geographic distance between samples related to the Sørensens dissimilarity of the seaweed flora?
2. Does the average Sørensens dissimilarity vary among the bioregions to which the samples belong?
3. Is the effect of geographic distance on the Sørensens dissimilarity different for each bioregion?

The most complex model is (3), the one that answers the question about whether the effect of `dist` on the response variable Y is different for each bioregion. Questions (1) and (2) are subsets of this more inclusive question. To fully answer these quesitons, let's first consider the full model, which includes an *interaction term* between the continuous predictor `dist` and the categorical predictor `bio`. When we finally test our model, we will also have to consider the simpler models that do not include the interaction term.

'Interaction' means that the effect of one predictor on the response variable is contingent on the value of another predictor. For example, we might have reason to suspect that the relationship of the Sørensens dissimilarity with the geographic distance between samples is different between the west coast compared to, say, the east coast. This is indeed a plausible expectation, but we will test this formally below.

The full multiple linear regression model with the interaction terms can be formally expressed as Equation 5.5:

$$\begin{aligned} Y = \alpha &+ \beta_1 \text{dist} + \beta_2 \text{bio}_{\text{B-ATZ}} + \beta_3 \text{bio}_{\text{BMP}} \\ &+ \beta_4 \text{bio}_{\text{ECTZ}} + \beta_5 (\text{dist} \times \text{bio}_{\text{B-ATZ}}) \\ &+ \beta_6 (\text{dist} \times \text{bio}_{\text{BMP}}) + \beta_7 (\text{dist} \times \text{bio}_{\text{ECTZ}}) + \epsilon \end{aligned} \tag{5.5}$$

Where:

- Y: The response variable, the mean Sørensen dissimilarity.
- α: The intercept term.
- dist: The continuous predictor variable representing distance.
- bio: The categorical predictor variable representing bioregional classification with four levels: `AMP` (reference category), `B–ATZ`, `BMP`, and `ECTZ`.
- $\text{bio}_{\text{B-ATZ}}, \text{bio}_{\text{BMP}}, \text{bio}_{\text{ECTZ}}$: Dummy variables for the bioregional classification, where:
  - $\text{bio}_{\text{B-ATZ}} = 1$ if `bio` = `B–ATZ`, and 0 otherwise,
  - $\text{bio}_{\text{BMP}} = 1$ if `bio` = `BMP`, and 0 otherwise, and
  - $\text{bio}_{\text{ECTZ}} = 1$ if `bio` = `ECTZ`, and 0 otherwise.
- $\text{dist} \times \text{bio}_{\text{B-ATZ}}, \text{dist} \times \text{bio}_{\text{BMP}}, \text{dist} \times \text{bio}_{\text{ECTZ}}$: Interaction terms between distance and the bioregional classification dummy variables.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$: The coefficients to be estimated for the main effects and interactions.
- ϵ: The error term.

If this seems tricky, it is because of the dummy variable coding used to represent interactions in multiple linear regression. The `bio` variable is a categorical variable with four levels, so we need to create three dummy variables to represent the bioregional classification. The `dist` variable is then interacted with each of these dummy variables to create the interaction terms. The `lm()` function in R takes care of this for us in a far less complicated model statement. I'll explain the details around the interpretation of dummy variable coding when we look at the output of the model with the `summary()` function.

5.7.1 State the Hypotheses for a Multiple Linear Regression with Interaction Terms

Equation 5.5 expands into the following series of hypotheses that concern the main effects, the interactions between the main effects, and the overall hypothesis:

Main effects hypotheses

In the main effects hypotheses we are concerned with the effect of each predictor variable on the response variable. For the main effect of distance we have the null:

- $H_0 : \beta_1 = 0$

vs. the alternative:

- $H_A : \beta_1 \neq 0$

For the main effect of bioregional classification, the nulls are:

- $H_0 : \beta_2 = 0 \quad (\text{bio}_{\text{B-ATZ}})$
- $H_0 : \beta_3 = 0 \quad (\text{bio}_{\text{BMP}})$
- $H_0 : \beta_4 = 0 \quad (\text{bio}_{\text{ECTZ}})$

vs. the alternatives:

- $H_A : \beta_2 \neq 0 \quad (\text{bio}_{\text{B-ATZ}})$
- $H_A : \beta_3 \neq 0 \quad (\text{bio}_{\text{BMP}})$
- $H_A : \beta_4 \neq 0 \quad (\text{bio}_{\text{ECTZ}})$

Hypotheses about interactions

This is where the hypothesis tests whether the effect of distance on the response variable is different for each bioregional classification. The null hypotheses are:

- $H_0 : \beta_5 = 0 \quad (\text{dist} \times \text{bio}_{\text{B-ATZ}})$
- $H_0 : \beta_6 = 0 \quad (\text{dist} \times \text{bio}_{\text{BMP}})$
- $H_0 : \beta_7 = 0 \quad (\text{dist} \times \text{bio}_{\text{ECTZ}})$

vs. the alternatives:

- $H_A : \beta_5 \neq 0 \quad (\text{dist} \times \text{bio}_{\text{B-ATZ}})$
- $H_A : \beta_6 \neq 0 \quad (\text{dist} \times \text{bio}_{\text{BMP}})$
- $H_A : \beta_7 \neq 0 \quad (\text{dist} \times \text{bio}_{\text{ECTZ}})$

Overall hypothesis

The overall hypothesis states that all coefficients associated with the predictors (distance, bioregional categories, and their interactions) are equal to zero, therefore indicating no relationship between these predictors and the response variable, the Sørensen index. The null hypothesis is:

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

vs. the alternative:

- $H_A : \exists\, \beta_i \neq 0$ for at least one i

5.7.2 Visualise the Main Effects

To facilitate the interpretation of the main effects hypotheses and make an argument for why an interaction term might be necessary, I've visualised the main effects (Figure 5.6). I see this as part of my exploratory data analysis ensemble of tests. We see that fitting a straight line to the Υ vs. distance relationship seems unsatisfactory as there is too much scatter around that single line to adequately capture all the structure in the variability of the points. Colouring the points by bioregion reveals the hidden structure. The model could benefit from including an additional level of complexity: see how points in the same bioregion show less scatter compared to points in different bioregions.

Now look at the boxplots of the Sørensen dissimilarity index for each bioregional classification. It shows that the median values of the Sørensen dissimilarity index are different for each bioregion.

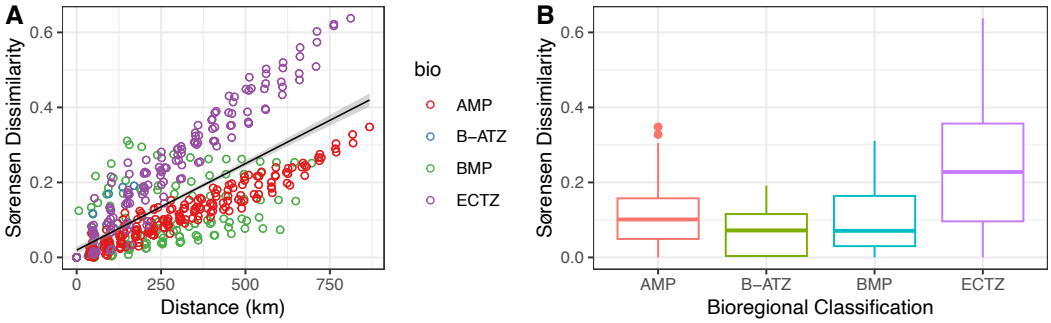

Figure 5.6: Plot of main effects of A) distance along the coast and B) bioregional classification on the Sørensen dissimilarity index.

Taken together, Figure 5.6 (A, B) provide a good indication that adding the bioregional classification might be an important predictor of the Sørensen dissimilarity index as a function of distance between pairs of sites along the coast.

Next, we will move ahead and fit the model inclusive of the distance along the coast and bioregion as per Equation (5.5).

5.7.3 Fit and Assess Nested Models

I have a suspicion that the full model (`mod2`; see below) with the interaction terms will be a better fit than reduced models with only the effect due to distance (seen independently). How can we have greater certainty that we should indeed favour a slightly more complex model (with two predictors) over a simpler one with only (distance only)?

One way to do this is to use a nested model comparison. We will fit a reduced model (one slope for all bioregions) and compare this model to the full model (slopes are allowed to vary among bioregions).

```
# Fit the linear regression model with only distance
mod2a <- lm(Y ~ dist, data = sw)

# Fit the multiple linear regression model with interaction terms
mod2 <- lm(Y ~ dist * bio, data = sw)
```

This is a nested model where `mod2a` is nested within `mod2`. 'Nested' means that the reduced model is a subset of the full model. Nested models can be used to test hypotheses about the significance of the predictors in the full model—does adding more predictors to the model improve the fit? Comparing a nested model with a full model can be done with a sequential ANOVA, which is what the `anova()` function also does (in addition to its use in Section 5.6.8).

So, comparing `mod2a` to `mod2` with an *F*-test tests the significance of adding the `bio` and using it together with `dist`. The interaction is built into `mod2` but we are not yet testing the significance of the interaction terms. We will do that later.

```
  anova(mod2a, mod2, test = "F")
```

```
Analysis of Variance Table

Model 1: Y ~ dist
Model 2: Y ~ dist * bio
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    968 7.7388
2    962 2.2507  6    5.4881 390.95 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The sequential ANOVA shows that there is significant merit to consider an interaction term in the model. This model would then allow us to have a separate slope for the Sørensen index as function of distance for each bioregion. The residual sum of squares (RSS) decreases from 7.7388 in Model 1 to 2.2507 in Model 2, which indicates that Model 2 explains a significantly larger proportion of the variance in the response variable. The *F*-test for comparing the two models yields an *F*-value of 390.95 with a highly significant *p*-value (< 0.0001). The improvement in model fit due to the inclusion of the interaction term is therefore statistically significant.

The above analyses skirted around the questions stated in the beginning of Section 5.7. I've provided statistical evidence that full model is a better fit than the reduced model (the sequential *F*-test tested this), so we should use both `dist` and `bio` in the model. I have not looked explicitly at the main effects of the predictors. However, we can easily address questions (1) and (2):

- Question 1: looking at the summary of `mod2a` tells us that the main effect of `dist` is a significant ($p < 0.0001$) predictor of the Sørensen dissimilarity index.
- Question 2: the main effect of `bio` is also significant ($p < 0.0001$), which is what we'd see if we fit the model `mod2b <- lm(Y ~ bio, data = sw)`.

Question 3 warrants deeper investigation. Next, we will look at the interaction terms in the full model `mod2` to see if the effect of `dist` on `Y` is different for each level of `bio`.

5.7.4 Interpret the Full Model

The model summary

```
  # Summary of the model
  summary(mod2)
```

```
Call:
lm(formula = Y ~ dist * bio, data = sw)

Residuals:
      Min        1Q    Median        3Q       Max
-0.112117 -0.030176 -0.004195  0.023698  0.233520

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.341e-03  4.177e-03   1.279   0.2013
```

```
dist           3.530e-04  1.140e-05  30.958  < 2e-16 ***
bioB-ATZ      -6.140e-03  1.659e-02  -0.370   0.7114
bioBMP         3.820e-02  6.659e-03   5.737 1.29e-08 ***
bioECTZ        1.629e-02  6.447e-03   2.527   0.0117 *
dist:bioB-ATZ  7.976e-04  1.875e-04   4.255 2.30e-05 ***
dist:bioBMP   -1.285e-04  2.065e-05  -6.222 7.31e-10 ***
dist:bioECTZ   4.213e-04  1.801e-05  23.392  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04837 on 962 degrees of freedom
Multiple R-squared:  0.8607,    Adjusted R-squared:  0.8597
F-statistic: 849.2 on 7 and 962 DF,  p-value: < 2.2e-16
```

In the output returned by `summary(mod2)`, we need to pay special attention to the use of dummy variable encoding for the categorical predictor. The `Coefficients` section is similar to that of `mod1` (see Section 5.6.8), but now it includes the categorical predictor `bio*` and the interaction terms `dist:bio*` (`*` indicating the levels of the categorical variable). The `bio` variable has four levels, `BMP`, `B-ATZ`, `AMP`, and `ECTZ`, and `AMP` is selected as reference level. This decision to selected `AMP` as reference is entirely arbitrary, and alphabetical sorting offers a convenient approach to selecting the reference. The coefficients for the other levels of `bio` are interpreted as the sum of the response variable and the reference level.

The following are the key coefficients in the model summary:

- `(Intercept)`: This is the estimated average value of `Y` when `dist` is zero and `bio` is the reference category (`AMP`). Its *p*-value (> 0.05) suggests it's not significantly different from zero.
- Main Effects:
  - `dist`: This represents the estimated change in `Y` for a one-unit increase in `dist` when the bioregion is the reference category, `AMP`. The highly significant *p*-value (< 0.0001) indicates a strong effect of distance in the `AMP`.
  - `bioB-ATZ`, `bioBMP`, `bioECTZ`: These are dummy variables representing different bioregions. Their coefficients indicate the difference in the average value of `Y` between each of these bioregions and the reference bioregion when `dist` is zero. Only `bioBMP` and `bioECTZ` are significantly different from the reference bioregion, `AMP`.
- Interaction Effects:
  - `dist:bioB-ATZ`, `dist:bioBMP`, `dist:bioECTZ`: These interaction terms capture how the effect of `dist` on `Y` varies across different bioregions. For instance, `dist:bioB-ATZ` indicates the additional change in the effect of `dist` in the `B-ATZ` bioregion compared to the reference bioregion, `AMP`. All interaction terms are highly significant, suggesting the effect of distance is different across bioregions.

Given this explanation, we can now interpret the coefficients of, for example, the `bioB-ATZ` main effect and `dist:bioB-ATZ` interaction. Since `AMP` is the reference bioregion, its effect is absorbed into the intercept term. Therefore, the coefficient for `bioB-ATZ` directly reflects the difference we are interested in. The coefficient for `bioB-ATZ` is -0.0061 $\pm$ 0.0166 lower than that of the reference, but the associated *p*-value (> 0.05) indicates that the average value of `Y` in the `B-ATZ` bioregion is not significantly different from the reference bioregion, `AMP`.

If we'd want to report the actual coefficient for `B-ATZ`, we'd calculate the sum of the coefficients

for `(Intercept)` and `bioB–ATZ`. This would give us the estimated average value of `Y` in the `B–ATZ` bioregion when `dist` is zero. The associated SE is calculated as the square root of the sum of the squared SEs of the two coefficients. Therefore, the coefficient for `B–ATZ` is $-8 \times 10^{-4} \pm 0.0171$.

The coefficient of 8×10^{-4} for `dist:bioB–ATZ` indicates that the effect of distance on `Y` is 8×10^{-4} units greater in the `B–ATZ` bioregion compared to the `AMP` bioregion. The SE of 2×10^{-4} suggests a high level of precision in this estimate, and the *p*-value (< 0.0001) indicates that this difference is statistically significant.

As before, to calculate the actual coefficient for `dist` in the `B–ATZ` bioregion, we'd sum the coefficients for `dist` and `dist:bioB–ATZ`. The associated SE of this sum is calculated as the square root of the sum of the squared SEs of the two coefficients. Therefore, the coefficient for `dist` in the `B–ATZ` bioregion is $0.0012 \pm 2 \times 10^{-4}$.

Concerning the overall hypothesis, the `Adjusted R–squared` value of 0.8597 indicates that the model explains 85.97% of the variance in the response variable `Y`. The `F–statistic` and associated `p–value` (< 0.0001) indicate that the model as a whole is highly significant, meaning at least one of the predictors (including interactions) has a significant effect on `Y`.

The ANOVA table

```
# The ANOVA table
anova(mod2)
```

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
dist        1 8.4199  8.4199 3598.79 < 2.2e-16 ***
bio         3 3.6232  1.2077  516.21 < 2.2e-16 ***
dist:bio    3 1.8648  0.6216  265.69 < 2.2e-16 ***
Residuals 962 2.2507  0.0023
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table's interpretation is intuitive and simple: the `Pr(>F)` column shows the *p*-value for each predictor in the model. The `dist` predictor has a highly significant effect on `Y` (< 0.0001), as do all the bioregions and their interactions with `dist`. This confirms the results we obtained from the coefficients. We don't need to overthink this result.

5.8 Example 3: The Final Model

I'll now expand `mod1` to include `bio` as a predictor alongside `augMean`, `febSD`, and `augSD` (`mod1` was applied only to data pertaining to `ECTZ`, one of the four levels in `bio`).

$$
\begin{aligned}
Y = \alpha &+ \beta_1 \text{augMean} + \beta_2 \text{febSD} + \beta_3 \text{augSD} \\
&+ \beta_4 \text{bio}_{\text{B-ATZ}} + \beta_5 \text{bio}_{\text{BMP}} + \beta_6 \text{bio}_{\text{ECTZ}} \\
&+ \beta_7 (\text{augMean} \times \text{bio}_{\text{B-ATZ}}) + \beta_8 (\text{augMean} \times \text{bio}_{\text{BMP}}) \\
&+ \beta_9 (\text{augMean} \times \text{bio}_{\text{ECTZ}}) + \beta_{10} (\text{febSD} \times \text{bio}_{\text{B-ATZ}}) \\
&+ \beta_{11} (\text{febSD} \times \text{bio}_{\text{BMP}}) + \beta_{12} (\text{febSD} \times \text{bio}_{\text{ECTZ}}) \\
&+ \beta_{13} (\text{augSD} \times \text{bio}_{\text{B-ATZ}}) + \beta_{14} (\text{augSD} \times \text{bio}_{\text{BMP}}) \\
&+ \beta_{15} (\text{augSD} \times \text{bio}_{\text{ECTZ}}) + \epsilon
\end{aligned} \tag{5.6}
$$

Where:

- Y: The response variable (mean Sørensen dissimilarity).
- α: The intercept term, representing the expected value of `Y` when all predictors are zero and `bio` is at the reference level `AMP`).
- β_1: The coefficient for the main effect of `augMean`.
- β_2: The coefficient for the main effect of `febSD`.
- β_3: The coefficient for the main effect of `augSD`.
- $\beta_4, \beta_5, \beta_6$: The coefficients for the main effects of the categorical predictor `bio` (for levels `B-ATZ`, `BMP`, and `ECTZ` respectively, with `AMP` as the reference category).
- $\beta_7, \beta_8, \beta_9$: The coefficients for the interaction effects between `augMean` and `bio` (for levels `B-ATZ`, `BMP`, and `ECTZ` respectively).
- $\beta_{10}, \beta_{11}, \beta_{12}$: The coefficients for the interaction effects between `febSD` and `bio` (for levels `B-ATZ`, `BMP`, and `ECTZ` respectively).
- $\beta_{13}, \beta_{14}, \beta_{15}$: The coefficients for the interaction effects between `augSD` and `bio` (for levels `B-ATZ`, `BMP`, and `ECTZ` respectively).
- ϵ: The error term, representing the unexplained variability in the response variable.

In this multiple regression model, we aim to understand the complex and interacting relationships between the response variables and the set of predictors. It allows us to investigate not only the individual effects of the continuous predictors on `Y`, but also how these effects might vary across the different bioregions.

The model therefore incorporates interaction terms between each continuous predictor (`augMean`, `febSD`, and `augSD`) and the categorical variable `bio`. This allows us to assess whether the relationships between `augMean`, `febSD`, or `augSD` and `Y` change depending on the specific bioregion. Essentially, we are testing whether the slopes of these relationships are different in different bioregions.

Additionally, the model examines the main effects of the bioregions themselves on `Y`. This means we're testing whether the average value of `Y` differs significantly across bioregions, after accounting for the influence of the continuous predictors.

This is how these different insights pertain to the model components:

- Main Effects: The coefficients for the main effects of `augMean`, `febSD`, and `augSD` represent the effect of each predictor when `bio` is at its reference level.
- Coefficients for `bio`: The coefficients for `bio` (e.g., $\beta_4 \text{bio}_{\text{B-ATZ}}$) represent the difference in the intercept for the corresponding level of `bio` compared to the reference level.

- Interaction Terms: The interaction terms allow the slopes of `augMean`, `febSD`, and `augSD` to vary across the different levels of `bio`. For example, β_7(augMean $\times$ $\text{bio}_{\text{B-ATZ}}$) represents how the effect of `augMean` on `Y` changes when `bio` is `B-ATZ` compared to `AMP`.

5.8.1 State the Hypotheses

Overall hypothesis

I'll only state the overall hypothesis for this model as the expansion of the individual hypotheses for each predictor and interactions (all the β-coefficients in Equation 5.6) is quite voluminous.

The null is that there is no relationship between the response variable `Y` and the predictors (including their interactions):

- H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$

The alternative is that at least one predictor or interaction term has a significant relationship with the response variable `Y`:

- H_A : At least one $\beta_i \neq 0$ for $i \in \{1, 2, ..., 15\}$

5.8.2 Fit the Model

In Section 5.6 I included the ECTZ seaweed flora in my analysis, but here I expand it to the full dataset. To assure myself that there is not a high degree of multicollinearity between the predictors, I have calculated the variance inflation factors (VIFs) for the full model (not shown). This allowed me to retain the same three predictors used in `mod1`, i.e. `augMean`, `febSD`, and `augSD`. This is the point of departure for `mod3`.

Now I fit the model with those three continuous predictors and their interactions with the categorical variable `bio`.

```
# Make a dataframe with only the relevant columns
sw_sub2 <- sw |>
  dplyr::select(Y, augMean, febSD, augSD, bio)

# Fit the multiple linear regression model with interaction terms
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)
```

Model `full_mod3a` is similar to `full_mod3` but without the interaction terms. This will allow me to compare the two models and assess the importance of the interactions.

```
# Compare the models
anova(full_mod3, full_mod3a)
```

```
Analysis of Variance Table

Model 1: Y ~ (augMean + febSD + augSD) * bio
Model 2: Y ~ augMean + febSD + augSD
```

```
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    954 3.5603
2    966 5.6890 -12   -2.1288 47.535 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(full_mod3, full_mod3a)
```

```
           df       AIC
full_mod3  17 -2652.498
full_mod3a  5 -2221.852
```

The AIC value for `full_mod3` is lower than that of `full_mod3a`, indicating that including the interaction with `bio` is necessary. Likewise, the ANOVA test also shows that the full model (lower residual sum of squares) is significantly better than the reduced model.

I therefore use `full_mod3` going forward. This is a complex model so I have used the stepwise selection function, `stepAIC()`, to identify the most important predictors and interactions (code and output not shown). I hoped that this might have simplified the model somewhat, but the simplification I had hoped for did not materialise.

5.8.3 Interpret the Model

The model summary

The model summary provides a detailed look at the individual predictors and their interactions in the model.

```
# Summary of the model
summary(mod3) # full_mod3 renamed to mod3 during stepAIC()
```

```
Call:
lm(formula = Y ~ augMean + bio + augSD + febSD + augMean:bio +
    bio:augSD + bio:febSD, data = sw_sub2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15399 -0.03841 -0.01475  0.03464  0.24051

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.0299094  0.0062756   4.766 2.17e-06 ***
augMean           0.3441099  0.0158575  21.700  < 2e-16 ***
bioB-ATZ         -0.0459611  0.0242519  -1.895 0.058374 .
bioBMP            0.0160756  0.0100749   1.596 0.110906
bioECTZ          -0.0015444  0.0090275  -0.171 0.864197
augSD            -0.0059012  0.0034011  -1.735 0.083044 .
febSD            -0.0006481  0.0027954  -0.232 0.816706
augMean:bioB-ATZ -0.0461775  0.0874044  -0.528 0.597400
augMean:bioBMP   -0.2406297  0.0211404 -11.382  < 2e-16 ***
```

```
augMean:bioECTZ  -0.0607745  0.0189030  -3.215 0.001348 **
bioB-ATZ:augSD    0.0655983  0.0371033   1.768 0.077382 .
bioBMP:augSD      0.0410220  0.0114706   3.576 0.000366 ***
bioECTZ:augSD     0.0280513  0.0053752   5.219 2.21e-07 ***
bioB-ATZ:febSD    0.0409425  0.0818927   0.500 0.617223
bioBMP:febSD      0.0056433  0.0150126   0.376 0.707070
bioECTZ:febSD     0.0502867  0.0082266   6.113 1.43e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06109 on 954 degrees of freedom
Multiple R-squared:  0.7797,    Adjusted R-squared:  0.7762
F-statistic: 225.1 on 15 and 954 DF,  p-value: < 2.2e-16
```

The first thing to notice is that the model function has been rewritten in the forward selection process (but none of the variables were deemed insignificant and removed):

- Initial specification: `Y ~ (augMean + febSD + augSD) * bio`
- Specification after `stepAIC()`: `Y ~ augMean + bio + augSD + febSD + augMean:bio + bio:augSD + bio:febSD`

Functionally, these two are identical, but the order in which the terms are presented differs. Although this has affected the order in which the coefficients are presented in the summary output, the coefficients are the same. The coefficients are:

- `(Intercept)`: This is the estimated average value of `Y` when all predictor variables are zero and the observation is in the reference bioregion (`AMP`).
- Main Effects:
  - `augMean`: For every one-unit increase in `augMean`, `Y` increases by 0.3441, on average, assuming all other predictors are held constant. This effect is highly significant.
  - `augSD` and `febSD`: The main effects of these variables are not statistically significant, suggesting they might not have a direct impact on `Y` when averaged across all bioregions.
  - `bioB-ATZ`, `bioBMP`, `bioECTZ`: These coefficients represent the average difference in `Y` between each of these bioregions and the reference bioregion, when the continuous predictors are held at zero.
- Interaction Effects:
  - `augMean` interactions: The significant interactions of `augMean` with bioregion indicate that the effect of `augMean` on `Y` varies across bioregions. Notably, the interaction with `bioBMP` has a strong, significant negative effect, suggesting that the positive effect of `augMean` is much weaker in this bioregion compared to the reference.
  - `augSD` and `febSD` interactions: These interactions with bioregions are sometimes significant, providing good support for the alternative hypothesis that the effects of `augSD` and `febSD` on `Y` depend on the specific bioregion.

Since dummy coding returns differences with respect to reference levels, how would we calculate the actual coefficients for, say, `augMean`? Since there are significant interaction effects, we must consider the main effect of `augMean` in conjunction with bioregion.

For `bio = B-ATZ`:

- $\beta_{\text{augMean}} + \beta_{\text{augMean:bioB-ATZ}} = 0.3441099 + (-0.0461775) = 0.2979324$

For `bio = BMP`:

- $\beta_{\text{augMean}} + \beta_{\text{augMean:bioBMP}} = 0.3441099 + (-0.2406297) = 0.1034802$

For `bio = ECTZ`:

$\beta_{\text{augMean}} + \beta_{\text{augMean:bioECTZ}} = 0.3441099 + (-0.0607745) = 0.2833354$

The respective SEs for these coefficients can be calculated using the formula for the standard error of the sum of two variables. For example:

- $SE_{\text{augMean}} = \sqrt{SE^2_{\text{augMean}} + SE^2_{\text{augMean:bio}}}$

The ANOVA table

The ANOVA table assesses the overall significance of groups of predictors or the sequential addition of predictors to the model.

```
anova(mod3)
```

```
Analysis of Variance Table

Response: Y
             Df Sum Sq Mean Sq  F value    Pr(>F)
augMean       1 9.9900  9.9900 2676.902 < 2.2e-16 ***
bio           3 1.1901  0.3967  106.296 < 2.2e-16 ***
augSD         1 0.1393  0.1393   37.331 1.451e-09 ***
febSD         1 0.0053  0.0053    1.422    0.2334
augMean:bio   3 0.7910  0.2637   70.647 < 2.2e-16 ***
bio:augSD     3 0.3426  0.1142   30.602 < 2.2e-16 ***
bio:febSD     3 0.1401  0.0467   12.517 4.953e-08 ***
Residuals   954 3.5603  0.0037
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that the model is highly significant, with very low p-values throughout (< 0.0001). This indicates that the model as a whole is a good fit for the data.

5.8.4 Reporting

Here is what the reporting of the findings could look like in the Results section in your favourite journal.

Results

A multiple linear regression model examining the effects of the August climatological mean temperature (`augMean`), the August and February climatological SD of temperature (`augSD` and `febSD`, respectively), and the bioregion classification (`bio`) on the response variable, the Sørensen dissimilarity (`Y`), including their interaction terms, revealed several significant findings (Table 5.1). This model allows a separate regression slope for each predictor within the bioregions (Figure 5.7). The model explains a substantial portion of the variance in `Y` ($R^2 = 0.780$, adjusted $R^2 = 0.776$), and the overall model fit is highly significant ($F(15, 954) = 225.1$, $p < 0.0001$).

Table 5.1: Summary of the multiple linear regression model examining the effects of augMean, augSD, febSD, and bio on Y.

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0299	0.0063	4.766	< 0.0001 ***
augMean	0.3441	0.0159	21.700	< 0.0001 ***
bioB–ATZ	-0.0460	0.0243	-1.895	> 0.05
bioBMP	0.0161	0.0101	1.596	> 0.05
bioECTZ	-0.0015	0.0090	-0.171	> 0.05
augSD	-0.0059	0.0034	-1.735	> 0.05
febSD	-0.0006	0.0028	-0.232	> 0.05
augMean:bioB–ATZ	-0.0462	0.0874	-0.528	> 0.05
augMean:bioBMP	-0.2406	0.0211	-11.382	< 0.0005 ***
augMean:bioECTZ	-0.0608	0.0189	-3.215	< 0.005 **
bioB–ATZ:augSD	0.0656	0.0371	1.768	> 0.05
bioBMP:augSD	0.0410	0.0115	3.576	< 0.0005 ***
bioECTZ:augSD	0.0281	0.0054	5.219	< 0.0005 ***
bioB–ATZ:febSD	0.0409	0.0819	0.500	> 0.05
bioBMP:febSD	0.0056	0.0150	0.376	> 0.05
bioECTZ:febSD	0.0503	0.0082	6.113	< 0.0005 ***

The main effect of augMean was highly significant (Estimate = 0.3441, $p < 0.0001$), indicating a strong positive relationship with Y. The interaction term augMean:bioBMP (Estimate = -0.2406, $p < 0.0001$) and augMean:bioECTZ (Estimate = -0.0608, $p < 0.005$) were also significant, suggesting that the effect of augMean on Y varies significantly for BMP and ECTZ bioregions compared to the reference category (AMP). The bioBMP (Estimate = 0.0161, $p > 0.05$) and bioECTZ (Estimate = -0.0015, $p > 0.05$) terms were not significant, indicating no significant difference from AMP.

For augSD, the main effect was not significant (Estimate = -0.0059, $p > 0.05$). Significant interaction terms for bioBMP:augSD (Estimate = 0.0410, $p < 0.001$) and bioECTZ:augSD (Estimate = 0.0281, $p < 0.0001$) indicate that the effect of augSD on Y varies by bioregion.

The main effect of febSD was not significant (Estimate = -0.0006, $p > 0.05$), suggesting no direct relationship with Y. However, the interaction term bioECTZ:febSD (Estimate = 0.0503, $p = 0.0001$) was significant, indicating that the effect of febSD on Y differs for the ECTZ bioregion.

The ANOVA further highlights the overall significance of each predictor. augMean had a highly significant contribution to the model ($F = 2676.902$, $p < 0.0001$), as did bio ($F = 106.296$, $p < 0.0001$), and their interactions (augMean:bio, $F = 70.647$, $p < 0.0001$; bio:augSD, $F = 30.602$, $p < 0.0001$; bio:febSD, $F = 12.517$, $p = 4.953 \times 10^{-8}$). The main effect of augSD was also significant ($F = 37.331$, $p = 1.451 \times 10^{-9}$), while febSD did not significantly contribute to the model on its own ($F = 1.422$, $p = 0.2334$).

These findings suggest that the effects of augMean, augSD, and febSD on Y are influenced by the bioregional classification, with significant variations in the relationships depending on the specific bioregion.

5.9 Alternative Categorical Variable Coding Schemes (Contrasts)

Throughout the book, we have used dummy variable coding the specify the categorical variables in the multiple linear regression models. But, should dummy variable coding not be to your liking, there are other coding schemes that can be used to represent the categorical variables. These alternative coding schemes are known as contrasts. The choice of contrast coding can affect the interpretation of the regression coefficients.

I'll provide some synthetic data to illustrate a few different contrasts. The data consist of a continuous variable `x`, a categorical variable `cat_var` with four levels, and a response variable `y` that has some relationship with `x` and `cat_var`. I'll use dummy variable coding as the reference (haha!).

```
head(data)
```

```
           y           x cat_var
1  0.6667876 -0.56047565       B
2  1.3086873 -0.23017749       B
3  0.4496192  1.55870831       D
4  2.1326402  0.07050839       A
5 -2.8608771  0.12928774       D
6  0.1497346  1.71506499       D
```

Categorical variable coding (any scheme) only affects the interpretation of the categorical variable main effects and their interactions, so I'll not discuss the coefficient associated with the continuous variable `x` (the slope) in the model throughout the explanations offered below.

Dummy Variable Coding (Treatment Contrasts)

This is the most commonly used coding scheme, and `lm()`'s default. One level is the reference category (`A`) and the other levels are compared against it. Contrast matrices can be assigned and/or inspected using the `contrasts()` function. For the dummy coding, the reference level `A` will remain 0 and the other levels will be independently coded as 1 in three columns. You'll now understand why, when we have four levels within a categorical variable, we only need three dummy variables to represent them.

```
# Dummy coding (treatment coding) ... default
contrasts(data$cat_var)
```

```
  B C D
A 0 0 0
B 1 0 0
C 0 1 0
D 0 0 1
```

When we have four levels in a categorical variable, there are three dummy variable columns in the contrast matrix. The first row, consisting of all zeros (0, 0, 0), represents the reference level, which in this case is `A`. The other rows represent the different levels of the categorical variable, with a 1 in the respective column indicating that level. For example, level `A` is represented by (0, 0, 0), `B` by (1, 0, 0), `C` by (0, 1, 0), and `D` by (0, 0, 1). In the regression model, these contrasts are used to estimate the differences between each level and the reference level. Specifically, the first contrast column indicates that the coefficient for this column will represent the difference between the

mean of the response variable for level `B` and the mean for the reference level `A`, holding all other variables constant. Similarly, the second and third columns represent the differences between levels `C` and `A`, and `D` and `A`, respectively. This coding allows for a straightforward interpretation of how each level of the categorical variable affects the response variable relative to the reference level.

```
model_dummy <- lm(y ~ x + cat_var, data = data)
summary(model_dummy)
```

```
Call:
lm(formula = y ~ x + cat_var, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6615 -0.6297 -0.1494  0.4978  2.9305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8176     0.1635  17.232  < 2e-16 ***
x             1.8274     0.1040  17.572  < 2e-16 ***
cat_varB     -1.7201     0.2499  -6.883 6.24e-10 ***
cat_varC     -3.9056     0.2678 -14.586  < 2e-16 ***
cat_varD     -5.4880     0.2512 -21.850  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9246 on 95 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8822
F-statistic: 186.4 on 4 and 95 DF,  p-value: < 2.2e-16
```

The model summary shows that the coefficients for `cat_varB`, `cat_varC`, and `cat_varD` represent the differences in the mean of the response variable `y` between the reference category `A` and categories `B`, `C`, and `D`, respectively, while controlling for the effect of the continuous variable `x`.

Interpretation:

- `(Intercept)` (2.8176): The intercept represents the estimated mean value of the response (`y`) when `x` is zero and the categorical variable is at the reference level `A`. This is the baseline from which other categories are compared.
- `x` (1.8274): For each one-unit increase in `x`, `y` is expected to increase by 1.8274 units, holding the categorical variable constant. This effect is consistent across all levels of the categorical variable because the model does not have an interaction effect present.
- `cat_varB` (-1.7201): On average, the value of `y` for level `B` is 1.7201 units lower than that for the reference level `A`, when `x` is held constant. This corresponds to the (1, 0, 0) row in the contrast matrix.
- `cat_varC` (-3.9056): Similarly, on average, the value of `y` for level `C` is 3.9056 units lower than that for the reference level, when `x` is held constant. This corresponds to the (0, 1, 0) row in the contrast matrix.
- `cat_varD` (-5.4880): Lastly, on average, the value of `y` for level `D` is 5.4880 units lower

compared to the reference , when `x` is held constant. This is row (0, 0, 1) row in the contrast matrix.

All these coefficients are highly significant ($p < 0.0001$), indicating strong evidence for differences between each category and the reference category `A`.

The model explains a large proportion of the variance in `y` (Adjusted R-squared: 0.8822), suggesting a good fit. The F-statistic (186.4) with a very low p-value (< 0.0001) indicates that the model as a whole is statistically significant.

If you want to change the reference level, you can use the `relevel()` function. For example, to change the reference level of `cat_var` variable to `C_2`, you can use:

```
# Set "C" as the reference level for cat_var
data$cat_var <- relevel(data$cat_var, ref = "C")
contrasts(data$cat_var)
```

```
  A B D
C 0 0 0
A 1 0 0
B 0 1 0
D 0 0 1
```

This may be useful when you want to compare the other levels to a different reference level.

Effect Coding (Sum Contrasts)

This coding method compares the levels of a categorical variable to the overall mean of the dependent variable. The coefficients represent the difference between each level and the grand mean. Instead of using 0 and 1 as we did with dummy variable coding, effect coding uses -1, 0, and 1 to represent the different levels of the categorical variable.

```
# Reset the reference level to "A"
data <- data.frame(y, x, cat_var)

# Effect coding
contrasts(data$cat_var) <- contr.sum(4)
contrasts(data$cat_var)
```

```
  [,1] [,2] [,3]
A    1    0    0
B    0    1    0
C    0    0    1
D   -1   -1   -1
```

In effect coding (sum contrasts), each level of the categorical variable is compared to the overall mean rather than a specific reference category. This contrast matrix with four levels (A, B, C, D) and three columns can be interpreted as follows:

- Level `A` (1, 0, 0): The first row indicates that level `A` is included in the first contrast (`cat_var1`), which means the mean of level `A` is being compared to the overall mean. Since the other columns are zero, level `A` does not contribute to the other contrasts.

- Level B (0, 1, 0): The second row indicates that level B is included in the second contrast (cat_var2). The mean of level B is being compared to the overall mean, and it does not contribute to the other contrasts.
- Level C (0, 0, 1): The third row indicates that level C is included in the third contrast (cat_var3). The mean of level C is being compared to the overall mean, and it does not contribute to the other contrasts.
- Level D (-1, -1, -1): The fourth row is a balancing row, ensuring that the sum of the contrasts for each level equals zero. This indicates that level D is being compared to the overall mean indirectly by balancing the contributions of levels A, B, and C.

```
model_effect <- lm(y ~ x + cat_var, data = data)
summary(model_effect)
```

```
Call:
lm(formula = y ~ x + cat_var, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6615 -0.6297 -0.1494  0.4978  2.9305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03921    0.09452   0.415    0.679
x            1.82741    0.10400  17.572  < 2e-16 ***
cat_var1     2.77844    0.14968  18.563  < 2e-16 ***
cat_var2     1.05832    0.16329   6.481 4.04e-09 ***
cat_var3    -1.12720    0.17765  -6.345 7.53e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9246 on 95 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8822
F-statistic: 186.4 on 4 and 95 DF,  p-value: < 2.2e-16
```

Interpretation:

- (Intercept) 0.03921: The intercept represents the grand mean of the response variable (y). Since the intercept is not statistically significant ($p > 0.05$), it indicates that the overall mean is not significantly different from zero when considering the average effect of all levels of the categorical variable.
- x (1.82741): For each one-unit increase in (x), the response (y) increases by approximately 1.82741 units. This effect is highly significant ($p < 0.0001$).
- cat_var1 (2.77844): Level A has a mean (y) that is 2.77844 units higher than the grand mean. This effect is highly significant ($p < 0.0001$).
- cat_var2 (1.05832): Level B has a mean (y) that is 1.05832 units higher than the grand mean. This effect is also highly significant ($p < 0.0001$).
- cat_var3 (-1.12720): Level C has a mean (y) that is 1.12720 units lower than the grand mean. This effect is highly significant (p < 0.0001).

All these coefficients are highly significant ($p < 0.0001$), indicating strong evidence for differences

between each category and the overall mean of all levels.

The model explains a large proportion of the variance in `y` (Adjusted *R*-squared: 0.8822), suggesting a good fit. The *F*-statistic (186.4) with a very low *p*-value (< 0.0001) indicates that the model as a whole is statistically significant.

Helmert Coding

Helmert coding compares each level of a categorical variable to the mean of the subsequent levels. It is useful for testing ordered differences.

```
# Helmert coding
contrasts(data$cat_var) <- contr.helmert(4)
contrasts(data$cat_var)
```

```
  [,1] [,2] [,3]
A   -1   -1   -1
B    1   -1   -1
C    0    2   -1
D    0    0    3
```

The contrast matrix for a categorical variable with four levels (`A`, `B`, `C`, `D`) and three columns can be interpreted as follows:

- Level `A` (-1, -1, -1): Level `A` is compared to the mean of levels `B`, `C`, and `D`. The negative values indicate that level A is being subtracted in these comparisons.
- Level `B` (1, -1, -1): Level `B` is compared to the mean of levels `C` and `D`. The positive value in the first column indicates that level `B` is being added in this comparison.
- Level `C` (0, 2, -1): Level `C` is compared to the mean of level `D`. The positive value in the second column indicates that level `C` is being added in this comparison, while the negative value in the third column is part of the comparison for subsequent levels.
- Level `D` (0, 0, 3): Level `D` is compared on its own in the final contrast. The positive value in the third column indicates that level `D` is being added in this comparison.

```
model_helmert <- lm(y ~ x + cat_var, data = data)
summary(model_helmert)
```

```
Call:
lm(formula = y ~ x + cat_var, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6615 -0.6297 -0.1494  0.4978  2.9305 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.03921    0.09452   0.415    0.679    
x            1.82741    0.10400  17.572  < 2e-16 ***
cat_var1    -0.86006    0.12495  -6.883 6.24e-10 ***
cat_var2    -1.01519    0.08206 -12.371  < 2e-16 ***
cat_var3    -0.90319    0.05477 -16.491  < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9246 on 95 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8822
F-statistic: 186.4 on 4 and 95 DF,  p-value: < 2.2e-16
```

Interpretation:

- `(Intercept)` (0.03921): The grand mean of `y` when `x` is zero.
- `x` (1.82741): For each unit increase in `x` , y increases by 1.82741 units.
- `cat_var1` (-0.86006): The mean of level `A` is 0.86006 units lower than the combined mean of levels `B`, `C`, and `D`.
- `cat_var2` (-1.01519): The mean of level `B` is 1.01519 units lower than the combined mean of levels `C` and `D`.
- `cat_var3` (-0.90319): The mean of level `C` is 0.90319 units lower than the mean of level `D`.

The interpretation of the overall model remains more-or-less similar to before:

All these coefficients are highly significant ($p < 0.0001$), indicating strong evidence for differences between each level and the overall mean of all subsequent levels.

The model explains a large proportion of the variance in `y` (Adjusted R-squared: 0.8822), suggesting a good fit. The F-statistic (186.4) with a very low p-value (< 0.0001) indicates that the model as a whole is statistically significant.

5.10 Exercises

! Task G

Use the data loaded at the start of this chapter for this task.
In this task you will develop data analysis, undertake model building, and provide an interpretation of the findings. Your goal is to explore the species composition and assembly processes of the seaweed flora around the coast of South Africa. See Smit et al. (2017) for more information about the data and the analysis.

a. **Analysis:** Please develop multiple linear regression models for the seaweed species composition (β_{sim} and β_{sne}, i.e. columns called `Y1` and `Y2`, respectively) using the all the predictors in this dataset. At the end, the final model(s) that best describe(s) the species assembly processes operating along the South African coast should be presented. The final model may/may not contain all the predictors in the dataset, and it is your goal to justify the variable and model selection.
    - Accomplishing a) will require that you work through the whole model-building process as outlined in the chapter. This includes the following steps:
        - Data exploration and visualisation (EDA)
        - Model building (providing hypothesis statements, variable selection using VIF and forward selection, comparisons of nested models, justifications for model selection)
        - Model diagnostics
        - Explanation of `summary()` and `anova()` outputs

    - Producing the Results section
    - **[60%]**

b. **Interpretation:** Once you have arrived at the best model, discuss your findings in the light of the appropriate ecological hypotheses that explain the relationships between the predictors and the seaweed species composition. Include insights drawn from the analysis of $\beta_{sør}$ that I developed in this chapter, and also rely on the theory you have developed for the lecture material the class presented in Task A2.
    - Accomplishing b) is thus all about model interpretation and discussing the ecological relevance of the results.
    - **[40%]**

The format of this task is a Quarto file that will be converted to an HTML file. The HTML file will contain the graphs, all calculations, and the text sections. The task should be written up as a publication (i.e. use appropriate headings) using a journal style of your choice. Aside from this, there are no limitations.

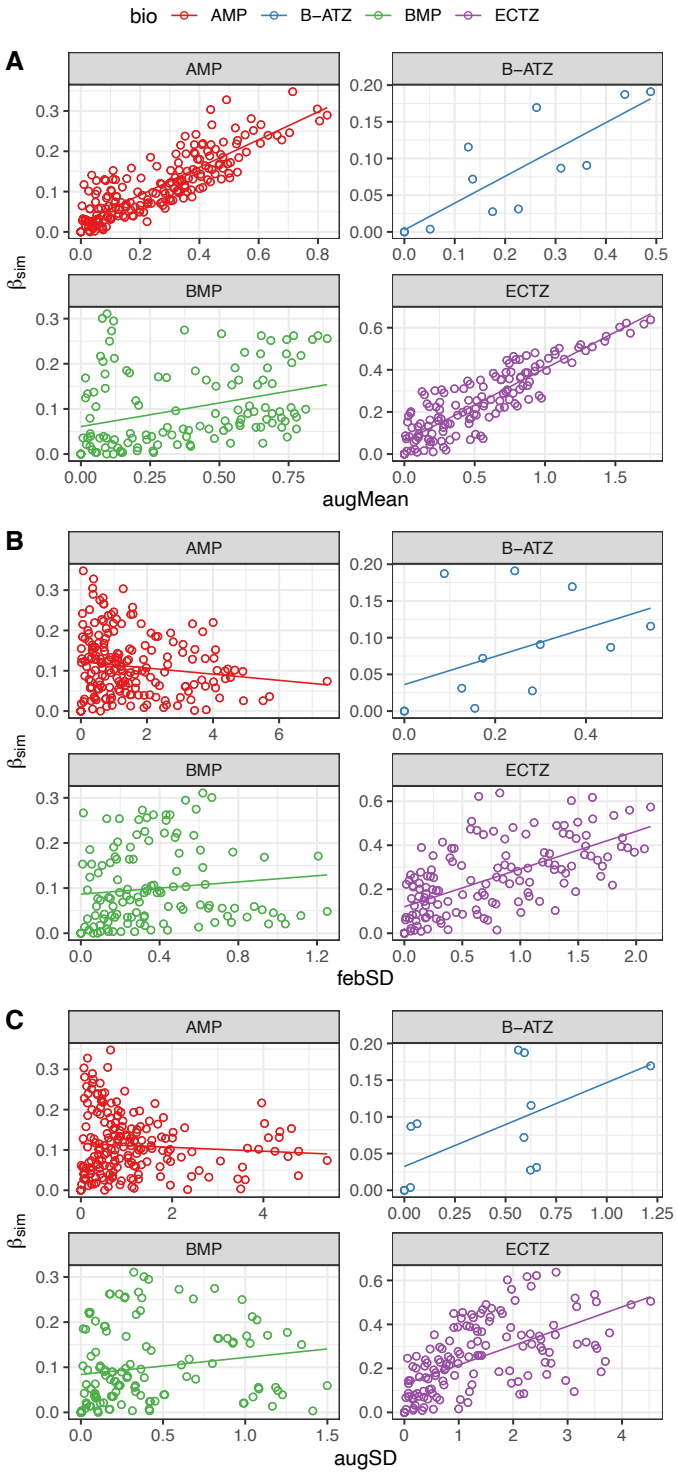

Figure 5.7: Individual linear regression fit to the variables `augMean`, `febSD`, and `augSD` for each bioregion as predictors of the seaweed species composition.