4
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 5), **multiple linear regression (MLR)** (Chapter 6), and **generalised linear models (GLMs)** (Chapter 7). 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 3) 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.

4.1 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 8.6).

The simple linear model is given by:

$$Y_i = \beta \cdot X_i + \alpha + \epsilon \tag{1}$$

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 8, 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.

4.2 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:** The independent variable (X) should be measured on a continuous scale, such as integers, real numbers, intervals, or ratios. Use an *ordinal regression* if the independent variable is ordinal. If you have more than one independent variable, use a *multiple linear regression*.
4. **Dependent Variable:** Similarly, the dependent variable (Y) should also be measured on a continuous scale, such as integers, real numbers, intervals, or ratios.[3] If your data are not continuous and the dependent variable is ordinal, use a *ordinal (logistic) regression*.

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.

Additional assumptions and requirements are discussed next in Section 4.3.

4.3 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 4.8 for more information about how to proceed when assumptions 1-3 are violated.

4.4 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.

4.5 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, ...)
```

TABLE 4.1. 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

where `formula` is a symbolic description using the notation by [2] 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`.

4.6 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 4.1.

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.

4.6.1 *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
```

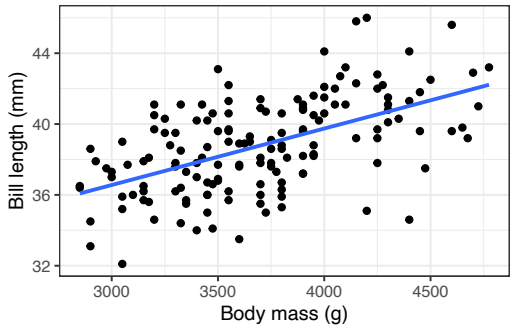

FIGURE 4.1. Scatter plot of the Palmer Station Adelie penguin data with a best fit line.

```
>                                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
```

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.

4.6.2 *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_bw()
```

Although there is some scatter in the data (Figure 4.1), 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.

4.6.3 *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:

$$H_0 : \beta = 0 \tag{2}$$

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.

4.6.4 *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
```

4.6.5 *Test the Assumptions*

Assumptions of normality, homoscedasticity, and linearity must be tested (Section 8.3).

We already noted that a linear model will probably be appropriate for the data (see Figure 4.1), 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 4.6.6).

```
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:

- 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 4.2 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 4.2 B).

Homoscedasticity

I now examine the homoscedasticity assumption. The residuals should be approximately equal across all values of the independent variable. There are several

FIGURE 4.2. Diagnostics plots the linear regression, `mod1`, for assumption testing.

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 4.2 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 4.2 D).

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

4.6.6 *Check for outliers*

How do we identify outliers in linear regression analysis? There are several approaches (see Figure 4.3):

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 4.4), what do we do with outliers? There are a

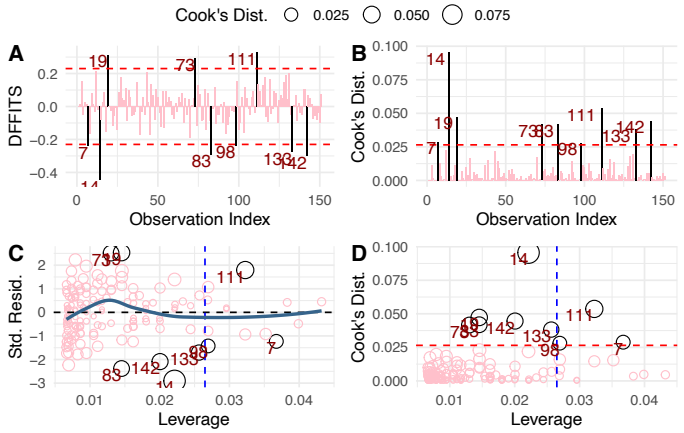

FIGURE 4.3. Diagnostic plots for visual inspection of outliers in the pernguin data. A) Difference in Fits (DFFITS) for `mod1`. B) Cook's distance. C) Residuals vs. leverage. D) Cook's distance vs. Lev./(1-Lev.). Outliers are identified beyond the Cook's distance threshold ($4/n$) and are plotted in 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))`.

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.

4.6.7 *Interpret the Results*

Now that we have tested the assumptions, we can interpret the results of the model fitted in Section 4.6.4. The slope of the regression line is 0.003188 mm/g, with a standard error of ± 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.

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

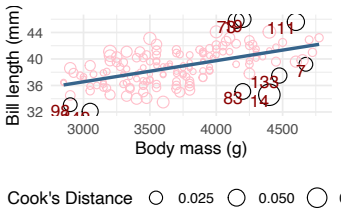

FIGURE 4.4. Plot of the linear regression resulting from `mod1` with the outliers identified using Cook's distance highlighted.

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.

4.6.8 *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.

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

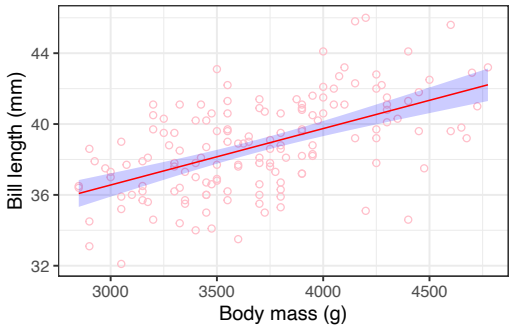

FIGURE 4.5. Plot of bill length as a function of body mass for Adelie penguins sampled at the Palmer Station. The straight line indicates the best fit regression line and the blue shading is the 95% confidence interval.

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

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 4.5).

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.

4.7 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 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 4.6.

```
# 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 uncer-

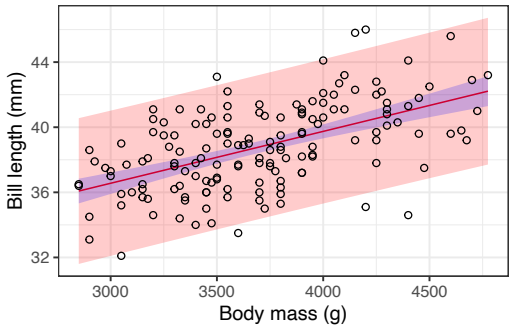

FIGURE 4.6. Plot of pernguin data with the confidence interval (blue) and prediction interval (pink) around the fitted values.

tainty associated 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.

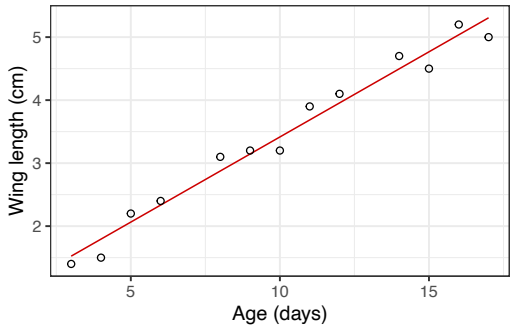

FIGURE 4.7. Scatter plot of the sparrow dataset with a best fit line.

4.8 WHAT DO I DO WHEN SOME ASSUMPTIONS FAIL?

4.8.1 *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 4.7 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

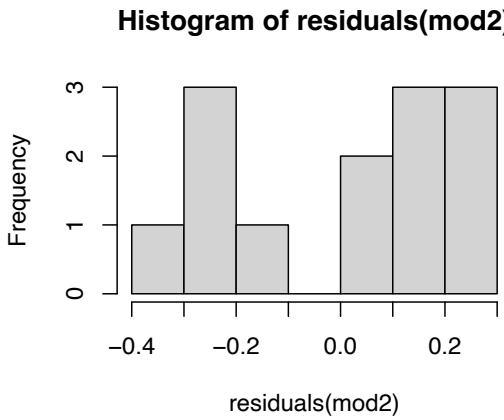

FIGURE 4.8. A histogram of the residual of the linear regression, `mod2`.

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 4.8 and Figure 4.9). 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.

```
bptest(mod2)
>
```

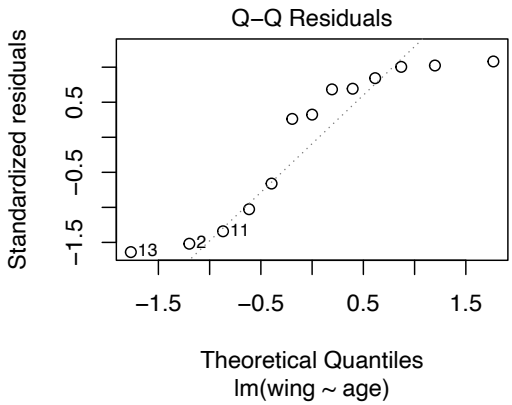

FIGURE 4.9. A Residual Q-Q plot of the linear regression, `mod2`.

```
>   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 4.10).

```
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.

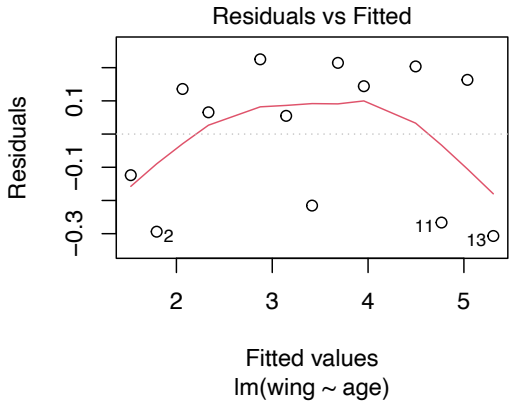

FIGURE 4.10. A plot of residuals against fitted values for the linear regression, `mod2`.

```
library(mblm)

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

4.8.2 *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. 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 4.8 for more details on how to proceed.

Outliers in the data can significantly impact the regression line, leading to misleading results (Section 4.6.6). 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 6) 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 7), 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 8) or generalised additive models (GAM; Chapter 12) might be necessary to describe the relationship between the variables accurately.