Analysis Steps and Results

1. We are analyzing Palmer Station penguin data from palmerpenguins package. After removing 11 rows that had missing values, we have a dataset with 7 variables and 333 observations.
2. First, we are interested in the relationship between the weight of a penguin and its sex. We can visualize the relationship in the following way.

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

From graphical analysis, it appears that male penguins weigh more than females. Let’s test if this difference is statistically significant with the **t-test**.

Text, letter

Description automatically generated

Based on p-value = 4.794e-16, we reject the null hypothesis. There appears to be a difference in the average weight of a penguin depending on its sex.

1. Next, we explored the relationship between the weight and the bill length. Visually, there appears to be a linear relationship between these two variables.

Chart, scatter chart

Description automatically generated

Let’s test if we can predict the weight of a penguin based on its bill length. The results of linear regression are as follows.

Text, letter

Description automatically generated

Based on p-value: < 2.2e-16, there appears to be a significant influence of bill length on weight. For every increase in bill length by 1mm (0.001m), the weight is expected to increase by 86g (0.086kg). However, the model only explains about 35% of the variation in body weight.

1. Next, let’s add more variables to the model and see if it helps explain more variation in the data.

Text

Description automatically generated

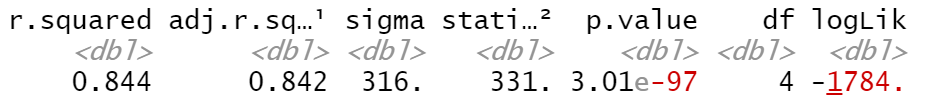
The model with the bill length, type of species, and sex explains about 85% of the variation in the body mass.

1. Finally, we’ll see how our multiple regression model performs on a new set of data.

We split the data 75/25 between the training and the test datasets and ran multiple regression on the training dataset. The results are as follows.

Table

Description automatically generated



Applied to a testing dataset, the model derived from a training dataset explains about 88% of the variation in the data.

Text

Description automatically generated