Blue Jays in Central Florida

Hannah White, Huiru Liu, Rhys McCrosson,
Sharone Vinushka Suthaharan, Ula Wolanowska (Group 6)

1 Introduction

The distinctive bright blue colour and noisy presence isn't the only thing that makes the Blue Jay so remarkable. Not only are they beautiful to look at, they are also highly intelligent and very adaptable to new environments. Data was collected for 65 Blue Jays captured in Central Florida to investigate the habitat of these birds. The birds underwent several measurements before being released back into the wild. Variables included each birds sex, body mass, skull length and bill length. Mass was measured in grams (g), whilst bill length and skull length were measured in millimetres (mm). The sex of each bird was measured on a categorical scale (M / F), male and female respectively.

The aim of this report is to firstly, investigate the effectiveness of predicting the body mass of a Blue Jay using a linear model with bill length as an explanatory variable. Secondly, to assess whether specifying the bird's sex and/or skull length as further explanatory variables can improve the linear model. And, thirdly determine whether there is evidence to suggest that, on average, the bird's skull length differs for male and female birds.

2 Methodology

To collect the data, 65 Blue Jays were captured and measured before being released back into the wild.

The variables recorded are listed as follows:

- **Mass** (*continuous variable*) - body mass of the Blue Jay (in g) - Primary response variable
- **BillLength** (*continuous variable*) - length of the bill (in mm) - Primary explanatory variable
- **KnownSex** (*categorical variable*) - gender of the Blue Jay: Female (F), Male (M) - Secondary explanatory variable
- **Skull** (*continuous variable*) - distance from the base of the bill to the back of the skull (in mm) - Secondary explanatory variable

If any missing data on the continuous variables is recorded, the bird will be removed from further analysis. However, if data on KnownSex of the bird is missing then this can be imputed via substitution and will be included in the analysis.

3 Exploratory Analysis

Figure 1 displays a scatterplot showing the correlation between the bill length and body mass of each bird. The blue and pink points represent male and female birds respectively. There is a moderate positive linear correlation between the two continuous variables BillLength and Mass, which suggests that the longer the bill on the bird the greater their body mass will be.

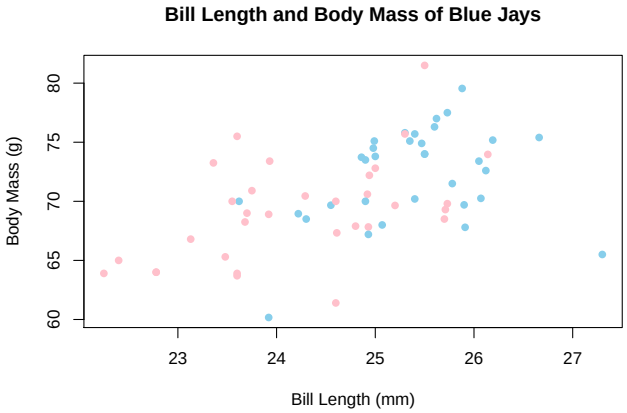

Figure 1: Bill Length and Body Mass of Blue Jays

Figure 2 demonstrates a pairs plot of scatterplots for variables BillLength, Mass and Skull to allow for brief analysis of the relationships. BillLength ~ Mass and Mass ~ Skull show a relatively strong positive linear correlation however, BillLength vs Skull shows no correlation.

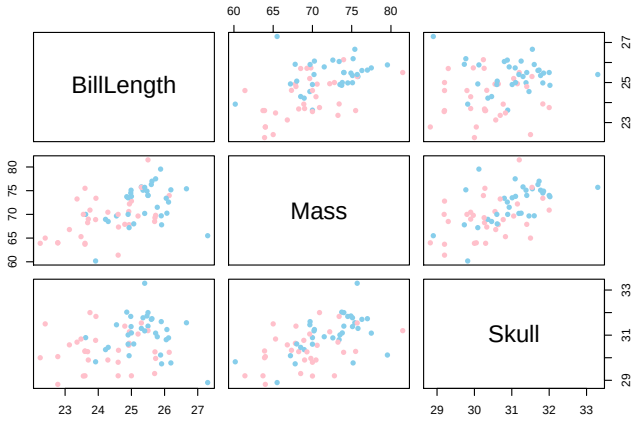

Figure 2: Exploring the relationships between variables

Figure 3 displays three boxplots for Skull, Mass and BillLength against KnownSex. Looking at the boxplots for Skull against KnownSex, it evident that females have a larger range of observed skull lengths in comparison to males. The median for males is higher than females implying that on average males have a larger Skull width than females by around 1mm. The boxplots for Mass against KnownSex show that on average male Blue Jays are 1g heavier than females. Whilst the female boxplot looks symmetrical, the boxplot for males appears left skewed. Lastly, the boxplots displayed for BillLength against KnownSex, show that the male

boxplot has a smaller interquartile range compared to females which suggests that there is less variation for BillLength in male birds. Again, BillLength appears to have higher measurements for males than females. Overall, from analysis of the boxplots it appears that male birds are larger than females in all aspects. Additionally, the female birds appear to have a higher variation in the data for all three measurements.

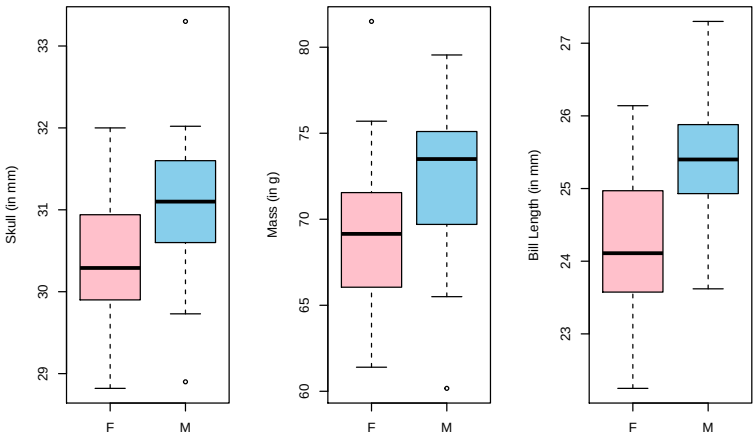

Figure 3: Summary Statistics for Blue Jays

4 Formal Analysis

4.1 Primary objective

The first model to be fitted (Model 1) will use simple linear regression, which will use the explanatory variable BillLength to predict the Mass of the Blue Jay.

$$Model\ 1: \quad y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Where y_i and x_i denote Mass and BillLength of the ith Blue Jay, respectively. The intercept of the linear regression line is denoted by α, while the slope is given by β. The random error component is denoted by ϵ_i, which we assume is normally distributed with mean zero and constant variance σ^2

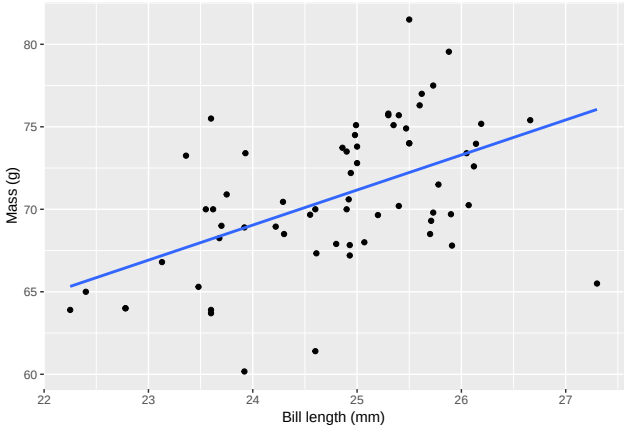

Figure 4: Relationship between body mass and bill length with regression line from Model 1

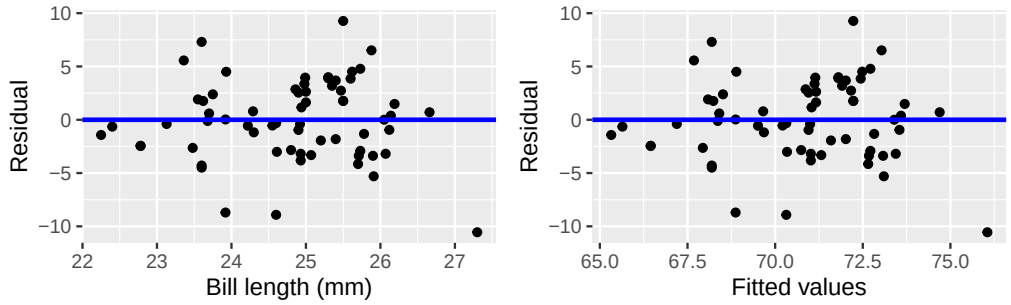

Figure 5: Scatterplots of the residuals against Bill Length (left) and the fitted values (right)

The assumptions corresponding to the linear regression model need to be assessed. Figure 5 shows scatterplots of the residuals against the explanatory variable bill length, as well as the fitted values. There is no obvious pattern in the residuals and the points appear evenly scattered above and below the zero line, hence have mean zero. Additionally, the spread of the residuals is constant across all values of the explanatory variable and the fitted values and display no obvious change in variability. Hence, the random error component of the regression model satisfies the assumptions of having mean zero and constant variance.

To examine whether they are normally distributed, Figure 3 displays a Q-Q normal plot of the residuals. The red line represents perfect quantile matching. Here, the points on the plot fall roughly on the straight line. This indicates that the residuals have the same distribution as our theoretical quantiles: they are normally distributed.

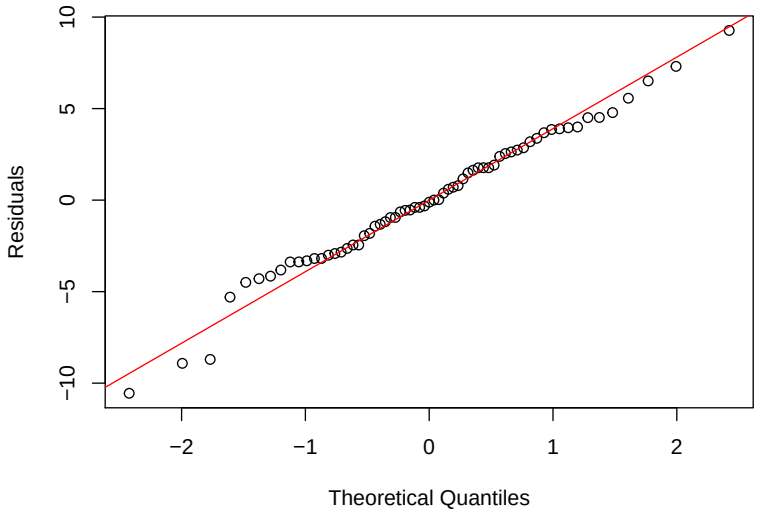

Figure 6: Q-Q plot of residuals for model 1

To check for the significance of BillLength as an explanatory variable, a 5% significance level is used. As shown in Table 1, the p-value for BillLength (produced using F value/F-test) is much smaller than 0.05. This concludes significance of BillLength as an explanatory variable in predicting the body mass.

Table 1: ANOVA table for Model 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BillLength	1	328.1771	328.17708	22.76458	1.13e-05
Residuals	63	908.2162	14.41613	NA	NA

4.2 Secondary objective

To further investigate if the linear model could be improved by adding the skull length and/or sex as additional explanatory variables, we fit Model 2:

$$Model\ 2: \quad y_i = \alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_{12} x_{1i}x_{2i}+\beta_{13} x_{1i}x_{3i}+\beta_{23} x_{2i}x_{3i}+\beta_{123} x_{1i}x_{2i}x_{3i}+\epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2)$$

where $i = 1, 2, ..., 65$ is the ith Blue Jay. y_i represents the mass of the ith Blue Jay and α is the intercept of the regression line for the baseline sex (females in our case). Additionally $x_{1i} = bill\ length_i$, $x_{2i} = skull\ length_i$ and $x_{3i} = \mathbb{I}_{\text{male}}(x)$ such that:

$$\mathbb{I}_{\text{male}}(x) = \begin{cases} 1 & \text{if Gender of } the\ i\text{th observation is Male,} \\ 0 & \text{Otherwise.} \end{cases}$$

As there are several potential explanatory variables, a stepwise regression procedure is used to identify which explanatory variables are most significant predictors and which variables should be removed from the model. Specifically, forward selection is used in this report. Model 1 is used as the null model and Model 2 as a full model. The type of test used is the F-test.

The final model produced (using forward selection)is Model 3:

$$Model\ 3: \quad \widehat{Mass}_i = -42.119 + 1.763 \cdot Bill\ Length_i + 2.251 \cdot Skull\ Width_i$$

On average, while keeping all other variables constant, for one mm increase in the bill length of a Blue Jay, its body mass increases by 1.763 g. Similarly, for one mm increase in the skull width, the body mass increases by 2.251 g while keeping all other variables constant.

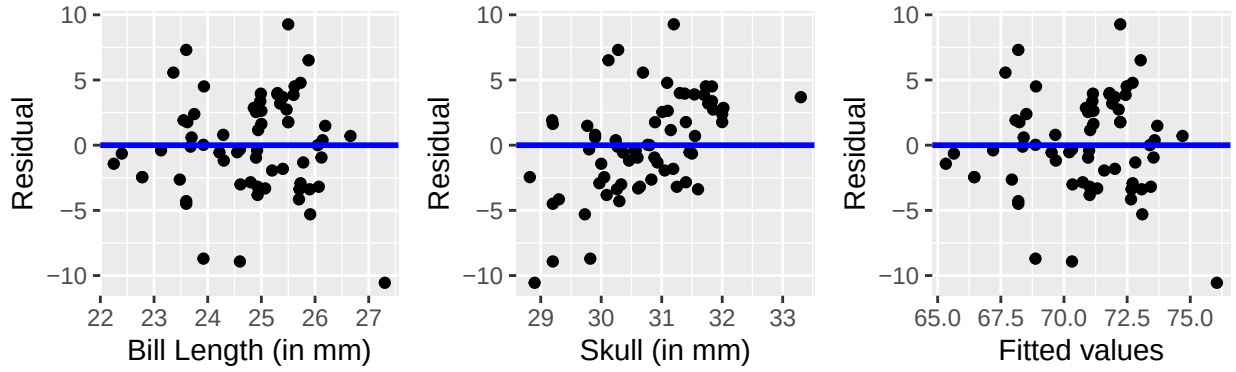

Figure 7: Scatterplots of the residuals against bill length (left) and skull (centre)

Figure 7 demonstrates that the final model satisfies the following assumptions: residuals have mean zero and constant variance at all values of the explanatory variable. Figure 8 confirms that the residuals are normally distributed as the points fall roughly on the straight line.

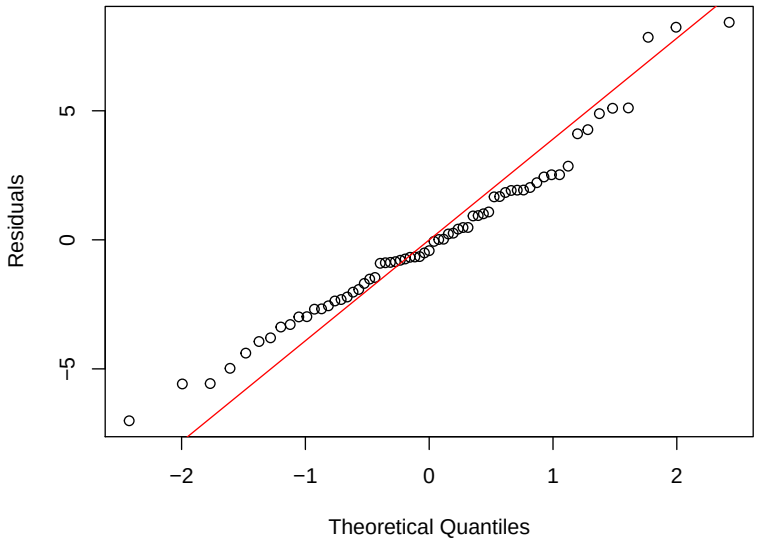

Figure 8: Q-Q plot of residuals for the final model

4.3 Tertiary objective

To investigate whether there is evidence to suggest that skull lengths differ, on average, between males and females, Model 4 is fitted:

$$Model\ 4: \quad \widehat{Skull\ Width}_i = \alpha + \beta \cdot \mathbb{I}_{\text{male}}(x)$$

where the intercept α represents the mean skull length for females and β is the difference in the mean skull length between females and males.

Table 2: ANOVA table for Model 4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
KnownSex	1	7.343907	7.3439073	10.41312	0.0019867
Residuals	63	44.431071	0.7052551	NA	NA

The one-way ANOVA shown in Table 2 reveals that the p-value for known sex is smaller than 0.05 (when using the F-test). Therefore, it can be concluded that sex is a significant explanatory variable and the skull length differs, on average, between males and females.

Table 3: Parameter estimates from Model 4

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	30.373	0.148	204.591	0.000	30.076	30.669
KnownSexM	0.672	0.208	3.227	0.002	0.256	1.089

Based on Table 3, the average skull length for females equals 24.23 and for males 25.36.

5 Conclusion

In summary, it is clear that there is in fact significant evidence to suggest that BillLength is an effective predictor of Mass within a linear model. It can also be demonstrated that the model fit can be improved by including Skull as an additional explanatory variable. Finally, there is significant evidence to suggest a difference in the average Skull between male and female Blue Jays. Possible limitations could include the fact that there is no way to satisfy the following assumptions of the explanatory variables being recorded without error & independence of the errors. Further limitations may also come from the fact that the Blue Jays can only be kept for so long due to consensus ethics; the sample size is relatively small; female Blue Jays are oversampled, and the age of the Blue Jays is unknown.

6 Appendix