Understanding a penguin's weight

Mac Kul

6/2/2022

Abstract

This analysis is motivated by my curiosity of the penguin anatomy and how culmen lengths, sex, and species type affect their body weight. To understand more about the penguin body weight, I ran a Welch's t-test to compare the means of penguin body weight between sexes, a two-way ANOVA test with the categorical variables, sex and species, and a linear regression model with the two categorical variables, along with a numeric culmen length variable to measure and predict the penguin body weight. Results show that all of the variables used, in fact affect how big penguins get.

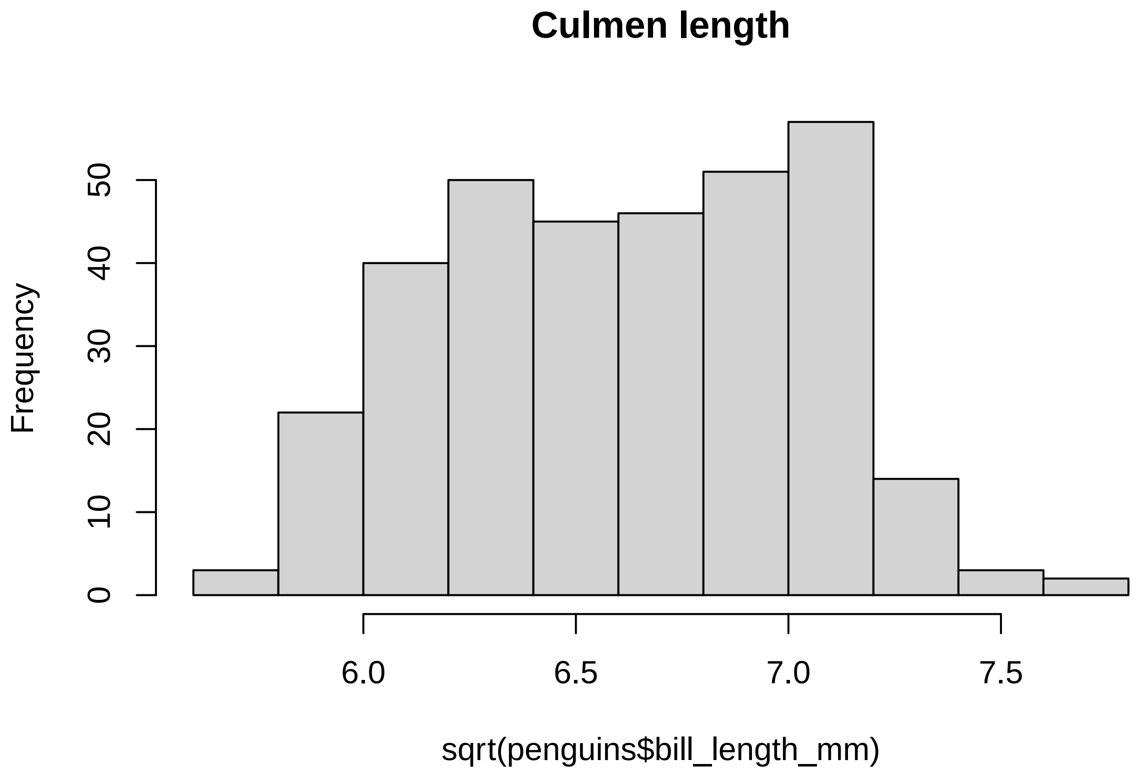
Introduction

The data were collected and made available as an R studio package ("palmerpenguins") by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. Within the package contains 2 datasets; The raw, unfiltered dataset and the cleaned up version consisting of non redundant columns. Both datasets contain 344 penguin observations, and there are 3 different species of penguins; Gentoo, Chinstrap, and Adelie. Because I am only looking at the biology of the penguin, I decided to choose sex and species type as the two categorical variable, and body weight and culmen length as the two numeric variables.

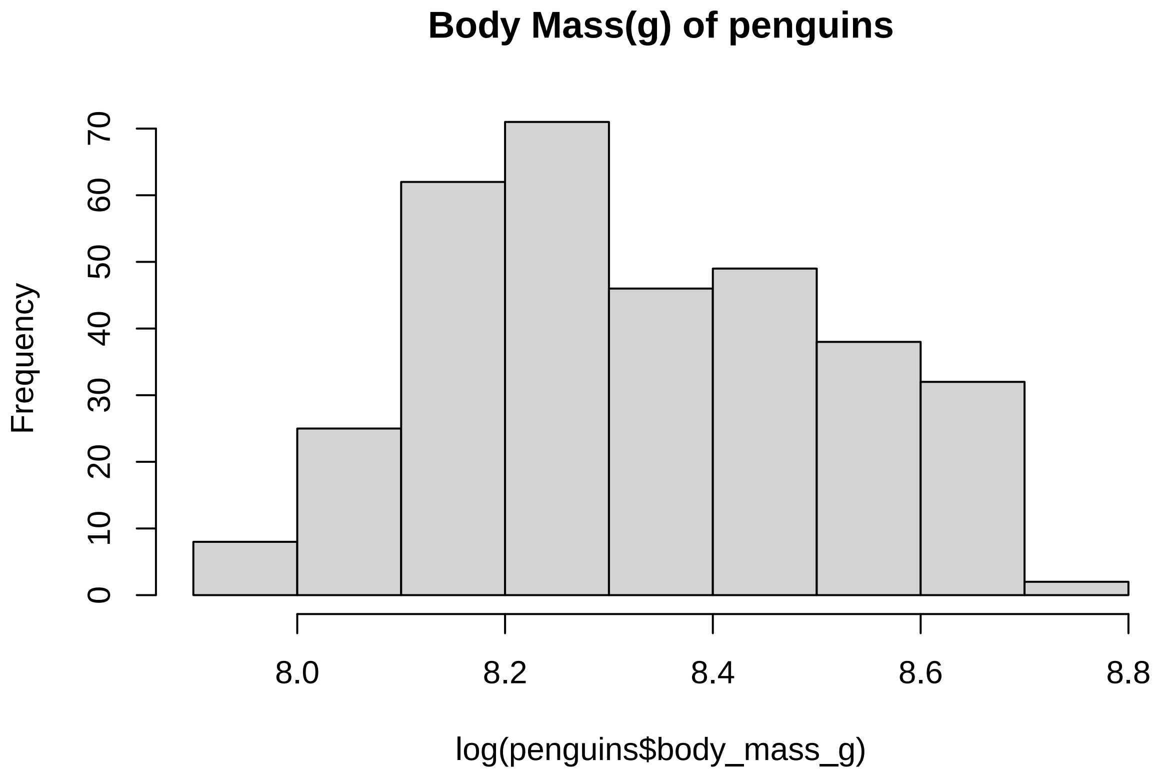
My hypothesis is that culmen lengths play a huge role in hunting. Penguins with longer culmen will have better reach and will ultimately have better chance at succeeding hunts. In addition, sexes and species types tend to generally dictate sizes of animals. Putting all of this information together, I believe that all of the variables used in this study will be significant, and can be used to predict more penguin body weights. Ultimately, the goal for this analysis is to test the significance of the predictors sexes, species, culmen length, and also create a prediction with synthetic data with the best chosen model.

Exploratory Data Analysis

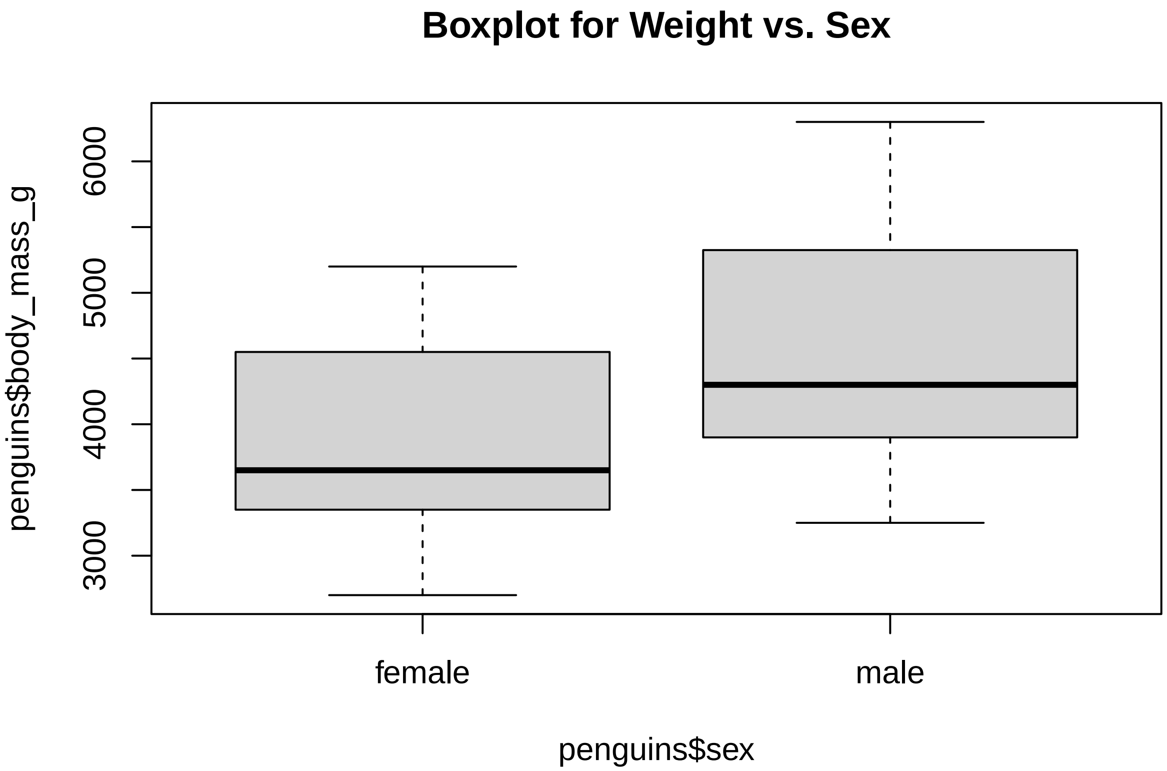
To get a better grasp of the data, I visualized numeric variables in histogram plots, and categorical variables in box plots in relation to the body weight. The main thing I looked for in histogram plots is a bell shaped curve. This tells me that the data is normally distributed. However, the culmen length (bill_length_mm) turned out to have a bimodal distribution, which required a transformation. For the reason why the data is bimodal, I believe that bill lengths vary between sexes. Male penguins will have longer culmen length compared to female penguins due to genetics and their role in hunting. I transformed the data using a square root transformation.

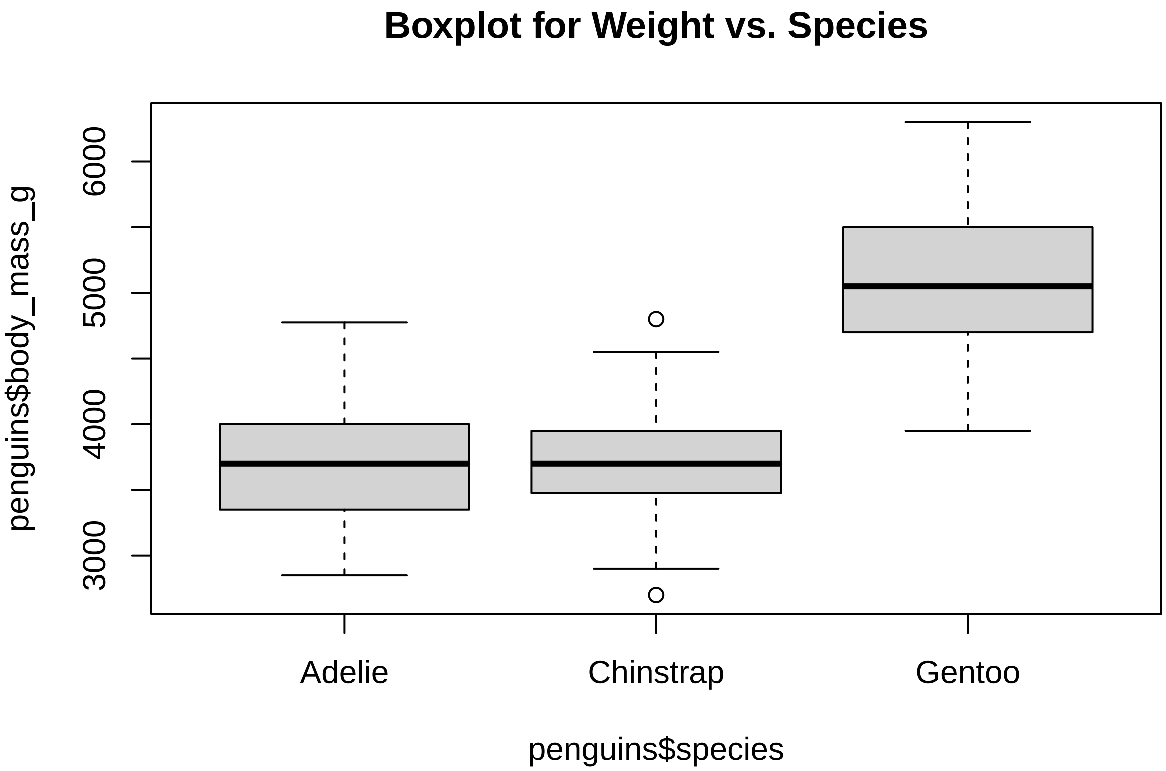


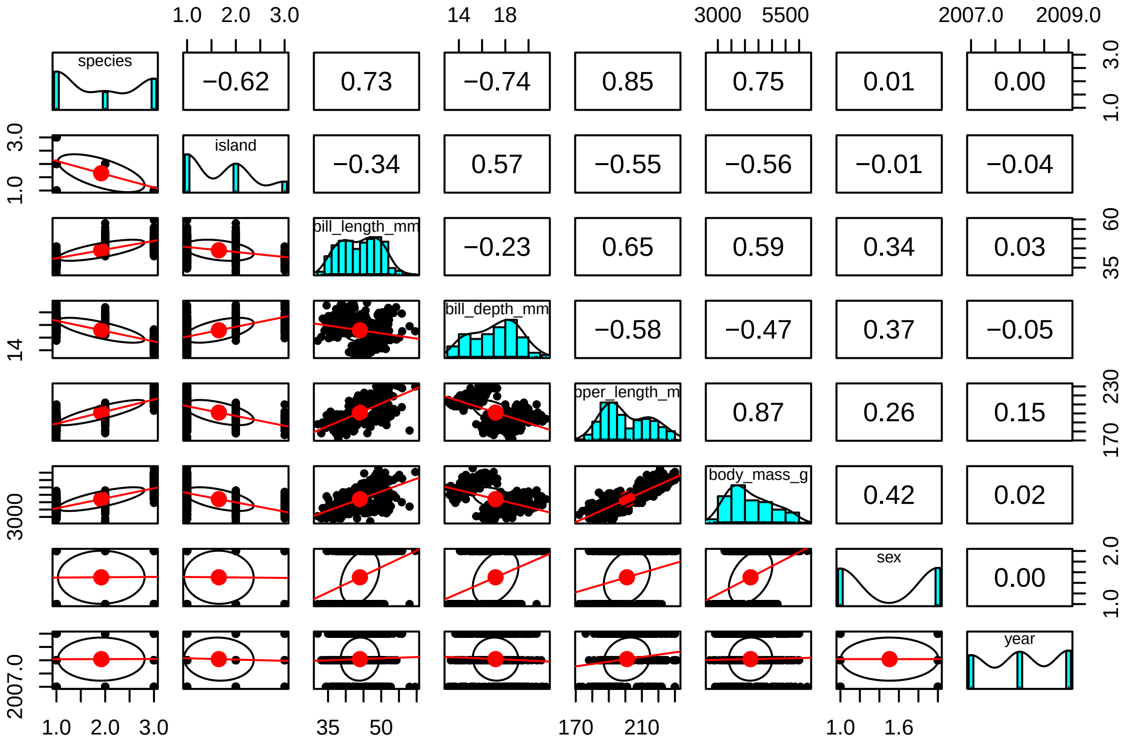
The histogram of the response variable, body weight of penguins (body_mass_g), the data seems to be slightly skewed to the right. Log transformation was applied to fix the curve.



Box plots were used to provide a visual summary of data to quickly identify mean values and potential outliers between groups. For this data, both sexes and species variables did not have many outliers, and the means of body weights are closely related with male and Chinstrap penguins being on the larger sides. Based on the visualization, female penguins have mean body mass of 8.2g, while male penguins have mean body mass of 8.4g. Interestingly enough, Adelie and Chinstrap penguins share the same mean body mass value of 8.5g while Gentoo penguins have average body mass of 8.5g.


Boxplot for Weight vs. Sex
penguins$body_mass_g
3000
4000
5000
6000
female
male
penguins$sex


Boxplot for Weight vs. Species
penguins$body_mass_g
3000
4000
5000
6000
Adelie
Chinstrap
Gentoo
penguins$species



This visualization displays a quick summary for the overall dataset. Since I am only looking at 4 variables, I do not need to transform the rest of the data. bill_length_mm and body_mass_g may not have the best bell shape curve, but since the sample size is > 50, Central Limit Theorem can be applied and fix the normality issue.

Statistical Methods

ANOVA(Analysis of Variance)

Two-way ANOVA test is a statistical test used to determine the effect of two nominal predictor variables on a continuous outcome variable. Running an ANOVA test with penguin body weight as the dependent variable against sex and species as the independent variables will allow me to understand those effects at a statistical level.

The three primary assumptions in ANOVA include: -Random Sample: Data must be random(no bias) -Homoscedasticity: The variance should be equal between different groups -Normally distributed residuals: The actual value minus mean value must be normally distributed

To verify these assumptions, the following methods were conducted: -Random Sample: Random sampling was assumed for this study. The penguins in Antarctica were randomly chosen. -Homoscedasticity: Run a “residuals vs. fitted” plot. If the points are spread out evenly with no distinct patterns, then the variances are equal. Levene’s Test can also verify homoscedasticity where the null hypothesis states that the variances are equal vs. alternative hypothesis that the variances are not equal. -Normally distributed residuals: Run a “Normal Q-Q plot” after finding the residuals. If the points follow the line, then residuals are normally distributed. Shipro Wilk’s Test can also verify the normality of the data where the null hypothesis states that the data is normally distributed vs. alternative hypothesis that the data isn’t normally distributed.

Linear Regression

Linear regression is a linear approach for modeling the relationship between a scalar response nd one or more explanatory variables. The response variable(y) is the penguin body weight(body_mass_g), and the explanatory variables are sex, species, and culmen length(bill_length_mm).

The summary report of a linear model will also explain the significance of explanatory variables. Hypothesis testing is used to confirm if the beta coefficients(predictors) are significant in a linear regression model. Every time we run a linear model, we test if a line is significant or not by checking the line.

The three assumptions are the same as ANOVA: -Random sample -Homoscedasticity -Normally distributed residuals

Results

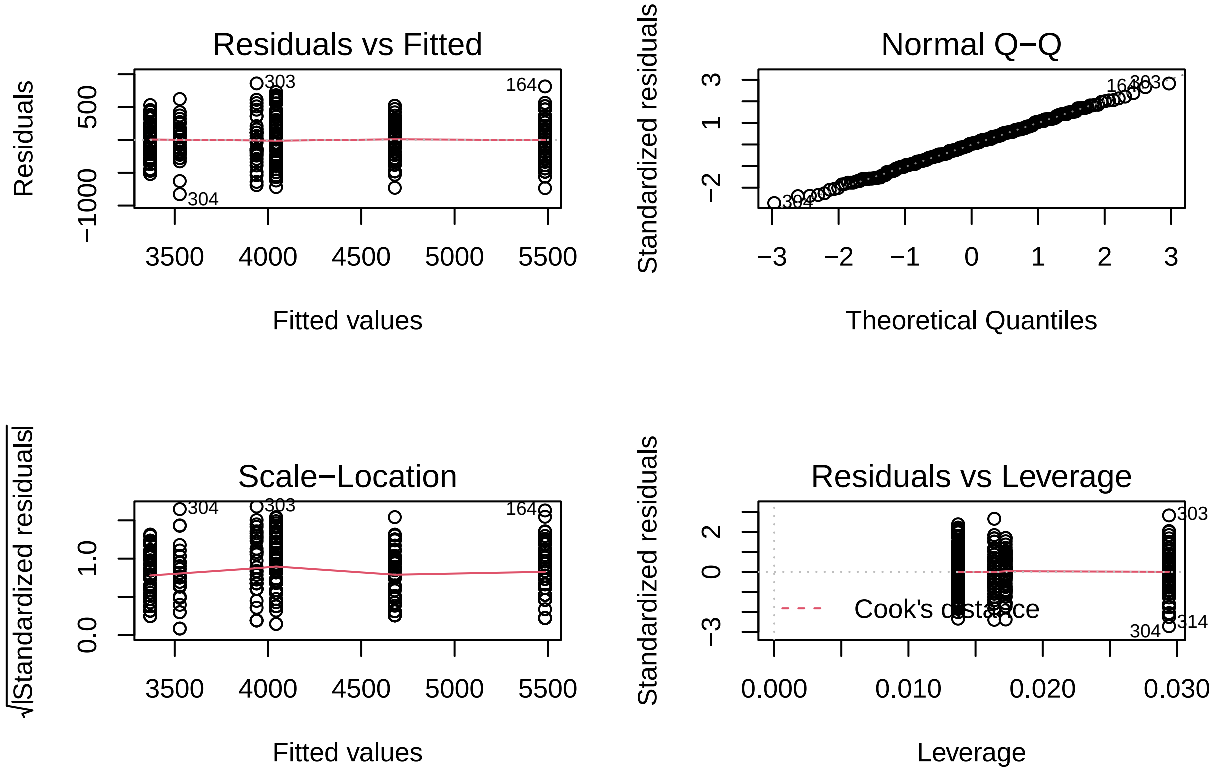
ANOVA

The two-way ANOVA summary reported p-values for the categorical variables sex, species, and the interaction between sex and species. p-values < 0.05 means that the variable of interest does have a significant effect on penguin body weight.

The ANOVA report on how penguin's sex affect body weight yields a p-value of 2*eˆ16(<0.05), which means that sex does have a significant effect on penguin body weight.

The ANOVA report on how penguin's species type affect body weight yields a p-value of 2*eˆ16(<0.05), which means that species type does have a significant effect on penguin body weight

The ANOVA report on how the interaction of penguin's sex and species type affect body weight yields a p-value of 0.00515 (<0.05). This p-value is small enough to conclude that the interaction of sex and species have a significant effect on penguin body weight.



The following figures show how the data satisfy homoscedasticity and the normality of residuals assumptions. In "Residuals vs. Fitted" plot, the points tend to center around y = 0 axis, signifying that the variances are equal. Similarly, the points in "Normal Q-Q" plot follow the y = x line, meaning that the data is normally distributed.

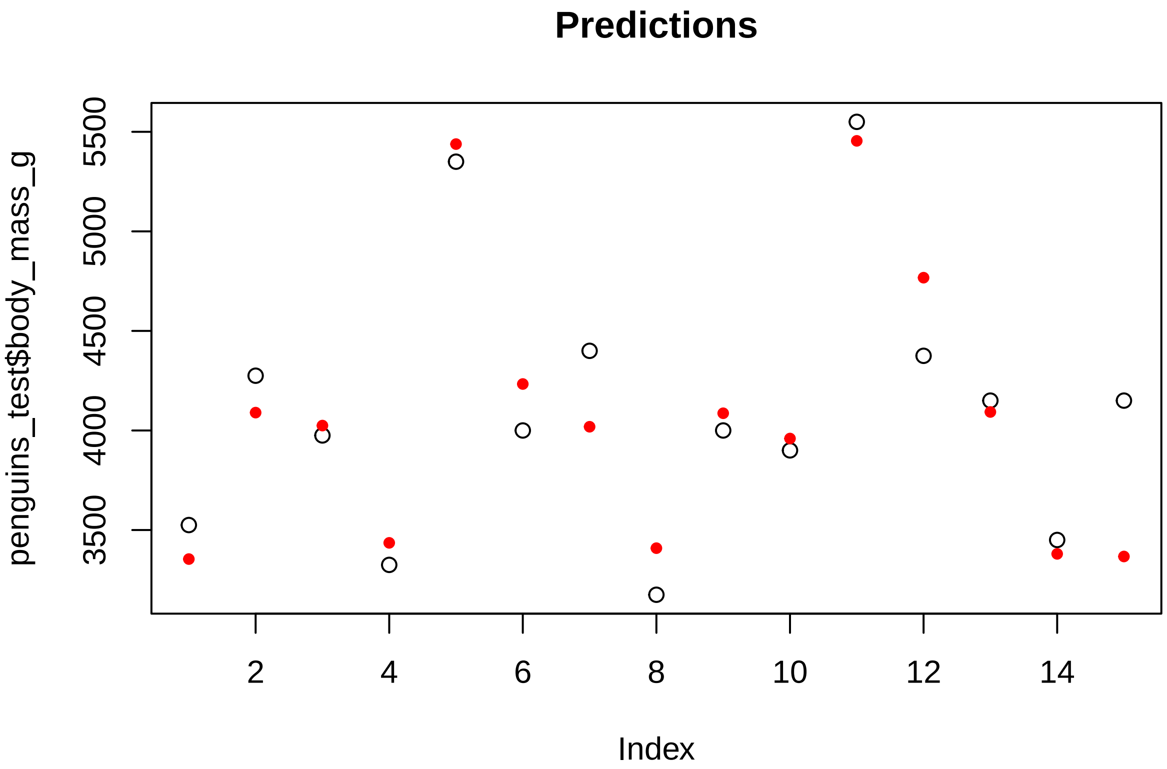
Linear Regression

The linear regression summary reported p-values for the explanatory variables and the intercept. p-values < 0.05 suggest that the coefficient of that predictor is statistically significant.

	df	AIC	BIC	rsq	adj_rsq
fit_penguin	4	5034.55	5049.78	0.67	0.67
fit_penguin2	5	4785.59	4804.63	0.85	0.85
fit_penguin3	6	4768.03	4790.87	0.86	0.85

The best model chosen by AIC/BIC is model3, which includes the intercept, species, sex, and the culmen length(bill_length_mm). Model3 has the lowest AIC and BIC values, as well has the highest R squared value.

The p-values for each predictors in model 3 are as follows: -Intercept(Including Adelie species type and female sex): <2*e^-16 -Chinstrap: 0.00196 -Gentoo: <2*e^-16 -Male: <2*e^-16 -Culmen length: 3.46*e^-5



Using the linear regression model as the baseline, the following plot was created which shows the prediction performance of the model. The red points are the predicted values and the white points are actual values. While the predictions are not exactly on the dot, the red points are not far off from the actual values, and it leaves some room for error which accounts for randomness within the data.

Discussion

Real World Application

From this analysis, we can conclude that the length of penguins' bill, sex, and species types are all factors of body mass. This finding may help biologists identify penguins with abnormal characteristics (i.e. penguins with smaller bill length) and aid them with food to help the penguin population, especially during global warming.

Limitation

Potential limitation of this study is how I restricted myself to only 2 categorical variables and 2 numerical variables. Given more time, I would like to test out other variables such as island, flipper length, and age. I also believe that the cleaned up version of palmerpenguins dataset that I used for this study was handpicked to yield significant results compared to the raw version. In addition, the average penguin body weight in present time might have changed due to how global warming had affected the Antarctic ecosystems. Perhaps an updated dataset may yield different results.

References

```
##
## To cite palmerpenguins in publications use:
##
##   Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer
##   Archipelago (Antarctica) penguin data. R package version 0.1.0.
##   https://allisonhorst.github.io/palmerpenguins/
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {palmerpenguins: Palmer Archipelago (Antarctica) penguin data},
##     author = {Allison Marie Horst and Alison Presmanes Hill and Kristen B Gorman},
##     year = {2020},
##     note = {R package version 0.1.0},
##     url = {https://allisonhorst.github.io/palmerpenguins/},
##   }
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D’Agostin
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }
```

```
##
## To cite package ’EnvStats’ in publications use:
##
## Millard SP (2013). _EnvStats: An R Package for Environmental
## Statistics_. Springer, New York. ISBN 978-1-4614-8455-4, <URL:
## https://www.springer.com>.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{EnvStats-book,
##     title = {EnvStats:  An R Package for Environmental Statistics},
##     author = {Steven P. Millard},
##     year = {2013},
##     publisher = {Springer},
##     address = {New York},
##     isbn = {978-1-4614-8455-4},
##     url = {https://www.springer.com},
##   }
```

```
##
## To cite the car package in publications use:
##
##   John Fox and Sanford Weisberg (2019). An {R} Companion to Applied
##   Regression, Third Edition. Thousand Oaks CA: Sage. URL:
##   https://socialsciences.mcmaster.ca/jfox/Books/Companion/
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     title = {An {R} Companion to Applied Regression},
##     edition = {Third},
##     author = {John Fox and Sanford Weisberg},
##     year = {2019},
##     publisher = {Sage},
##     address = {Thousand Oaks {CA}},
##     url = {https://socialsciences.mcmaster.ca/jfox/Books/Companion/},
##   }

##
## To cite the psych package in publications use:
##
##   Revelle, W. (2022) psych: Procedures for Personality and
##   Psychological Research, Northwestern University, Evanston, Illinois,
##   USA, https://CRAN.R-project.org/package=psych Version = 2.2.5.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {psych: Procedures for Psychological, Psychometric, and Personality Research},
##     author = {William Revelle},
##     organization = { Northwestern University},
##     address = { Evanston, Illinois},
##     year = {2022},
##     note = {R package version 2.2.5},
##     url = {https://CRAN.R-project.org/package=psych},
##   }

##
## To cite the 'knitr' package in publications use:
##
##   Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report
##   Generation in R. R package version 1.36.
##
##   Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition.
##   Chapman and Hall/CRC. ISBN 978-1498716963
##
##   Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible
##   Research in R. In Victoria Stodden, Friedrich Leisch and Roger D.
##   Peng, editors, Implementing Reproducible Computational Research.
##   Chapman and Hall/CRC. ISBN 978-1466561595
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

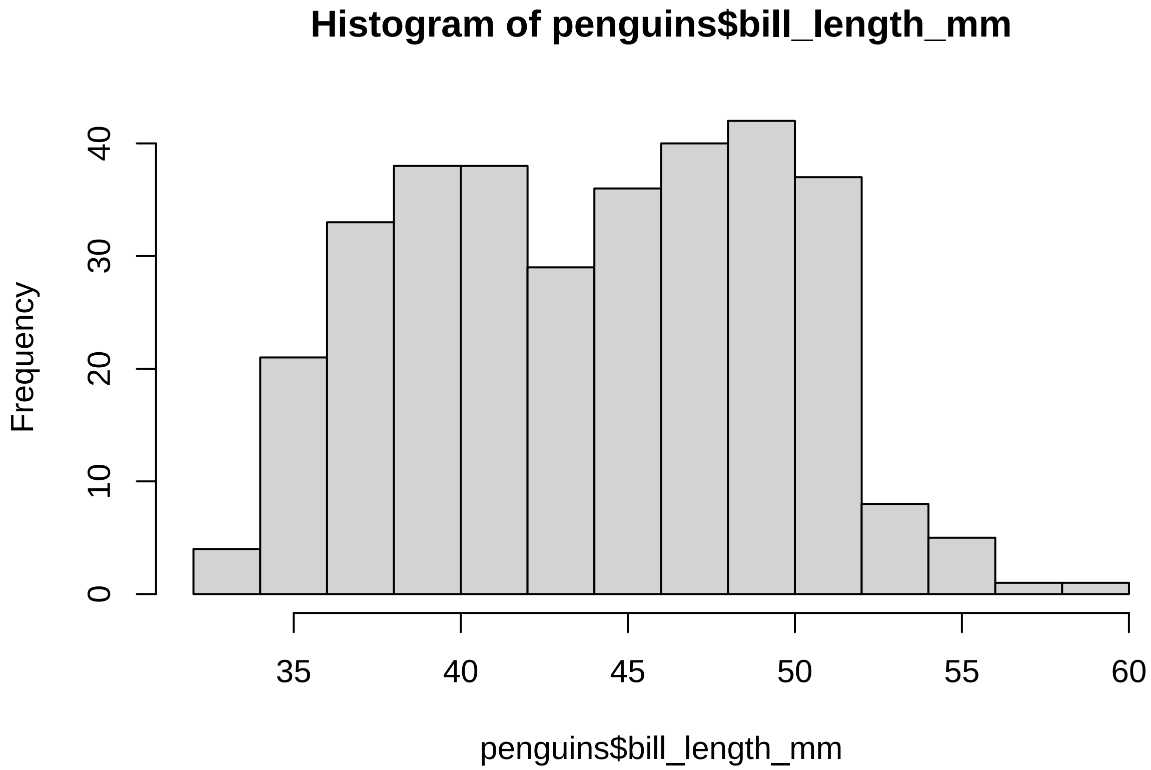
Appendix

Data cleaning

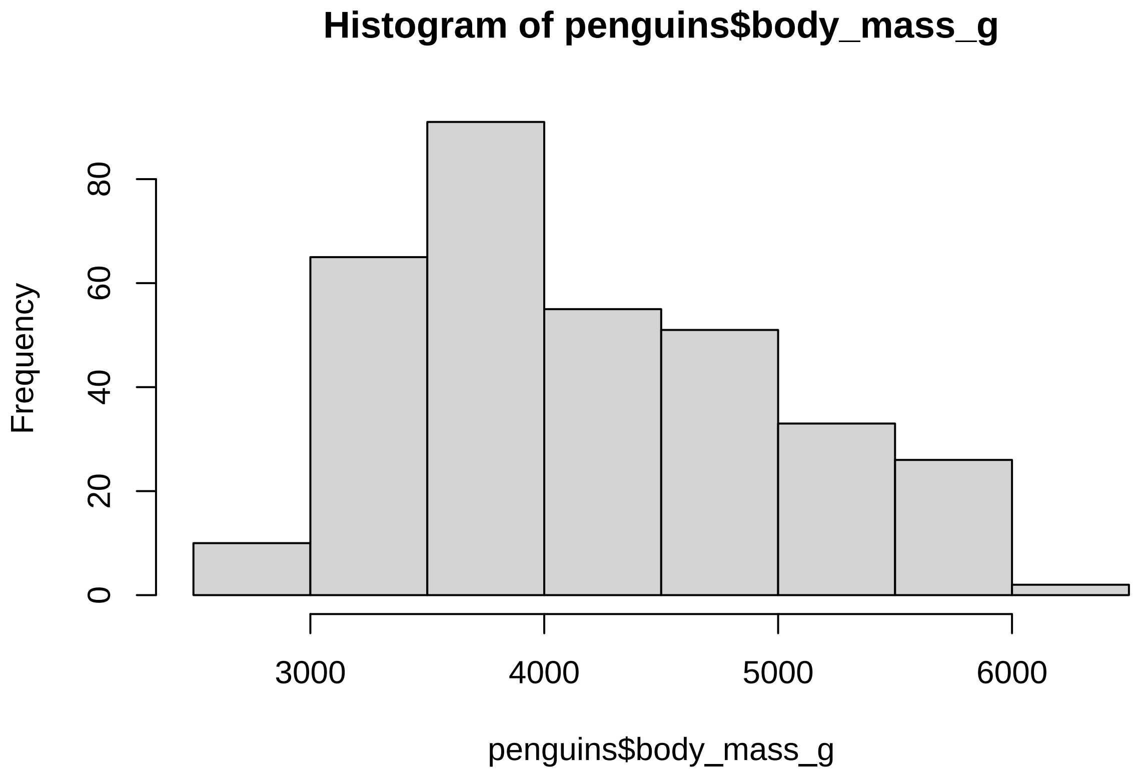
```
#filter missing values
penguins <- penguins %>%
              na.omit()
```

EDA

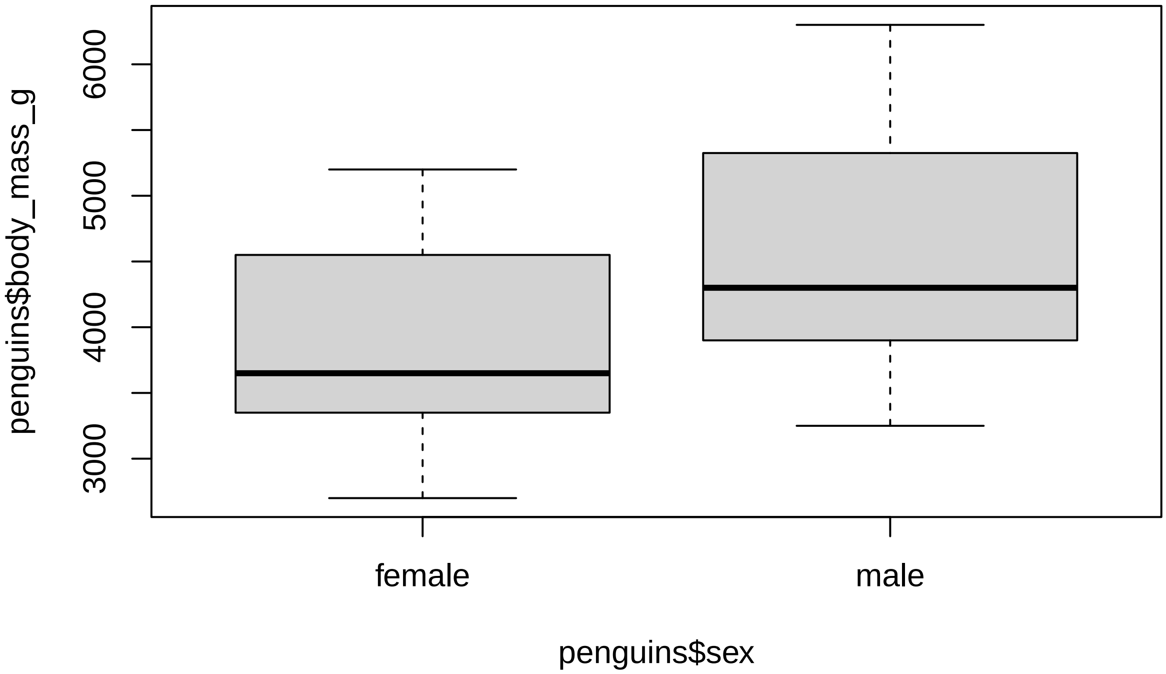
```
hist(penguins$bill_length_mm)
```



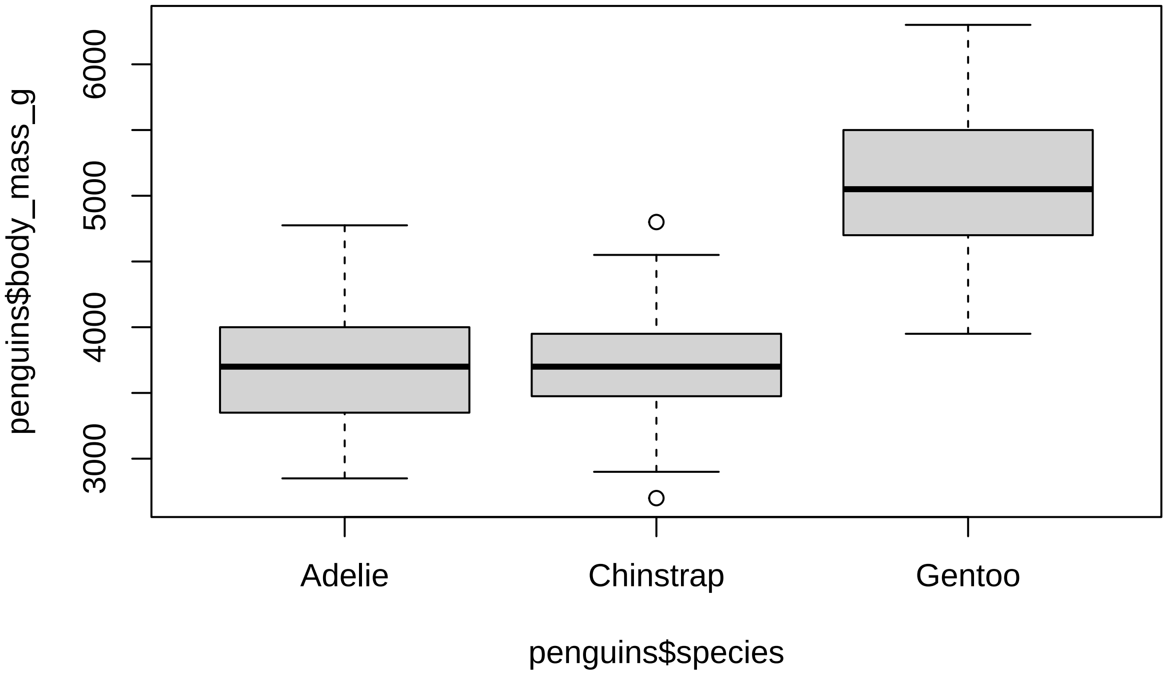
```
hist(penguins$body_mass_g)
```



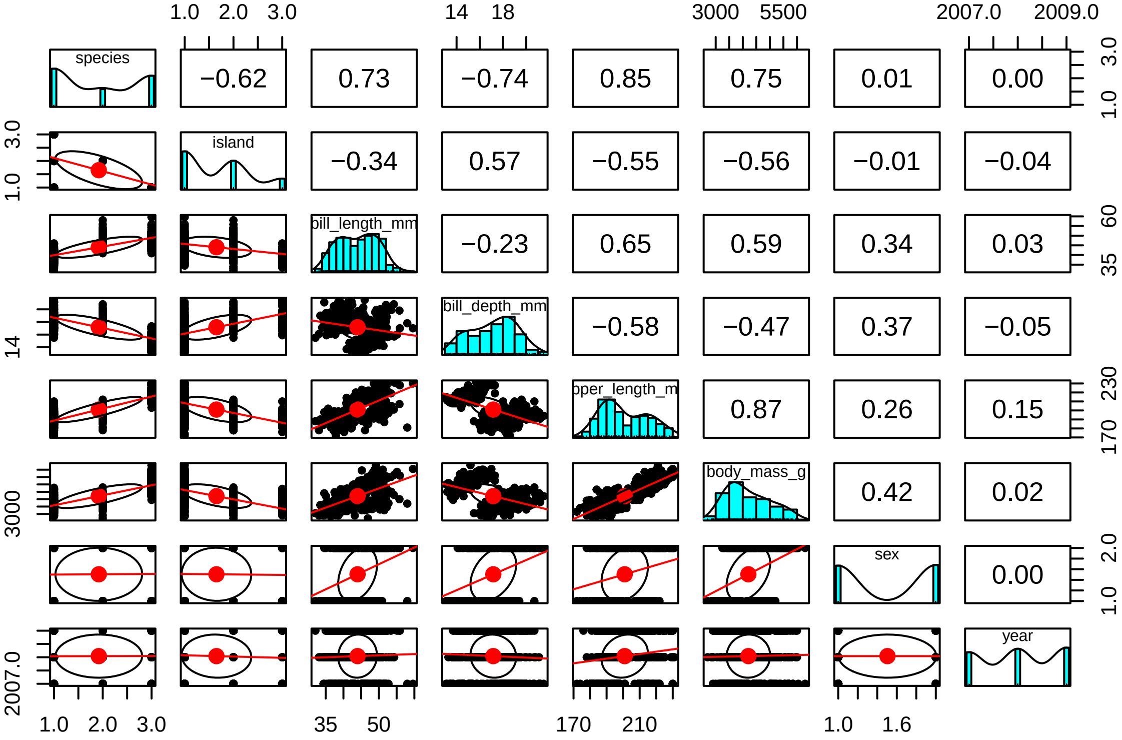
```
boxplot(penguins$body_mass_g~penguins$sex)
```



```
boxplot(penguins$body_mass_g~penguins$species)
```



```
par(mfrow = c(2,2))
pairs.panels(penguins, lm = TRUE, cor = T)
```

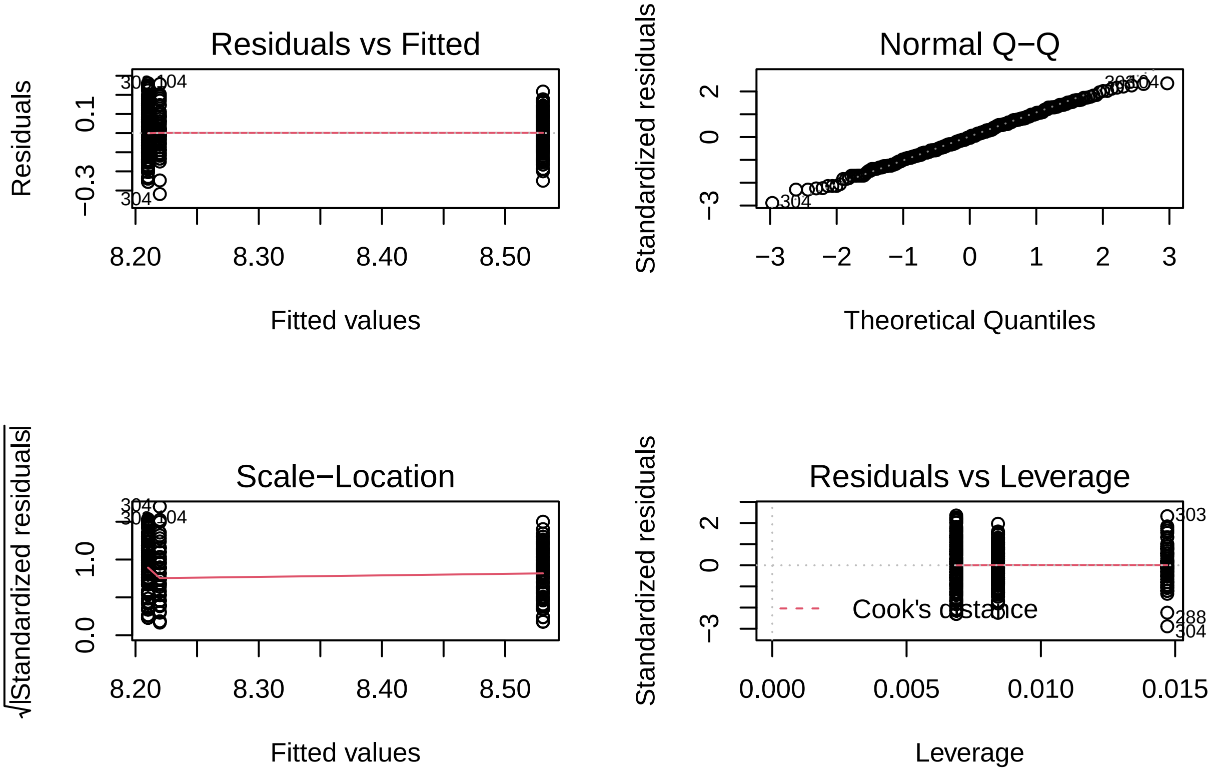


Data transformation

```
penguins$bill_length_mm <- sqrt(penguins$bill_length_mm)
penguins$body_mass_g <- log(penguins$body_mass_g)
```

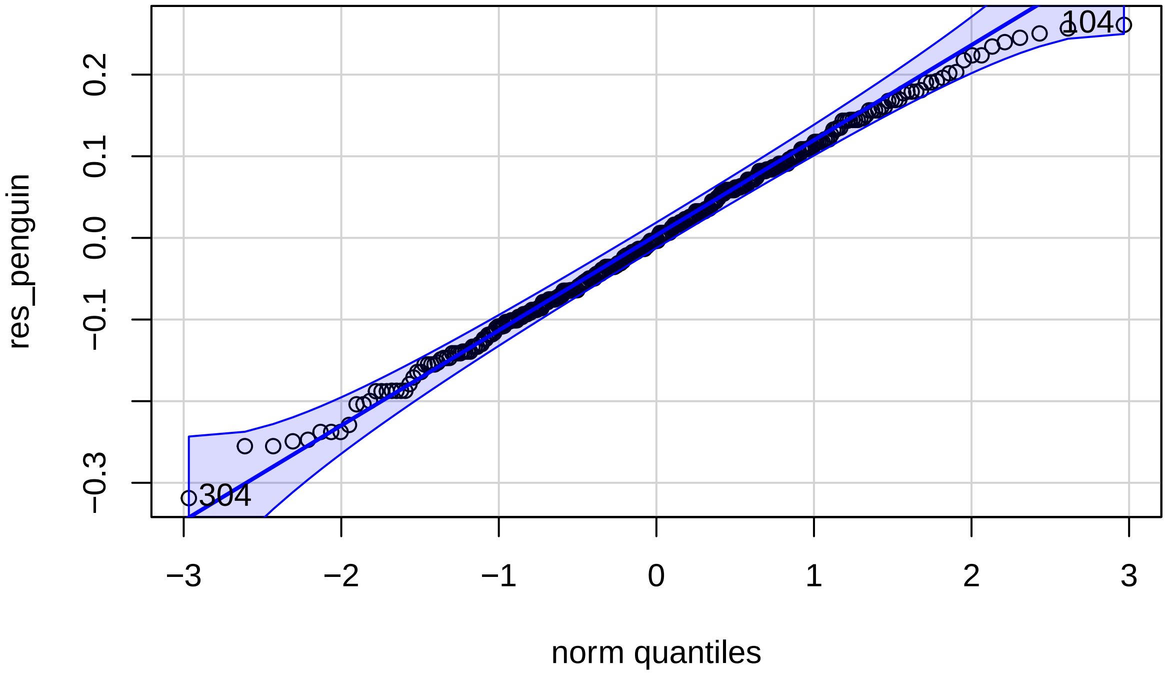
Check assumptions

```
fit_penguin <- lm(body_mass_g~species, data = penguins)
par(mfrow=c(2,2))
plot(fit_penguin)
```



Residuals and normality

```
res_penguin <- fit_penguin$residuals
qqPlot(res_penguin)
```



```
## [1] 304 104
```

```
shapiro.test(res_penguin)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_penguin
## W = 0.99591, p-value = 0.5419
```

p-value is 0.05118 > 0.05, fail to reject null hypothesis thus data is normal. But since the p-value is still very close to 0.05, we must check the number of samples of each species and see if we can apply CLT.

```
length(penguins[penguins$species == "Adelie",]$species)
```

```
## [1] 146
```

```
length(penguins[penguins$species == "Chinstrap",]$species)
```

```
## [1] 68
```

data size is large enough for CLT, residuals are normal

```
length(penguins[penguins$species == "Gentoo",]$species)
```

```
## [1] 119
```

```
length(penguins[penguins$sex == "male",]$sex)
```

```
## [1] 168
```

```
length(penguins[penguins$sex == "female",]$sex)
```

```
## [1] 165
```

T-test:

data size is large enough for CLT, residuals are normal

Check Variance

```
leveneTest(penguins$body_mass_g~penguins$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.3152 0.5749
##       331
```

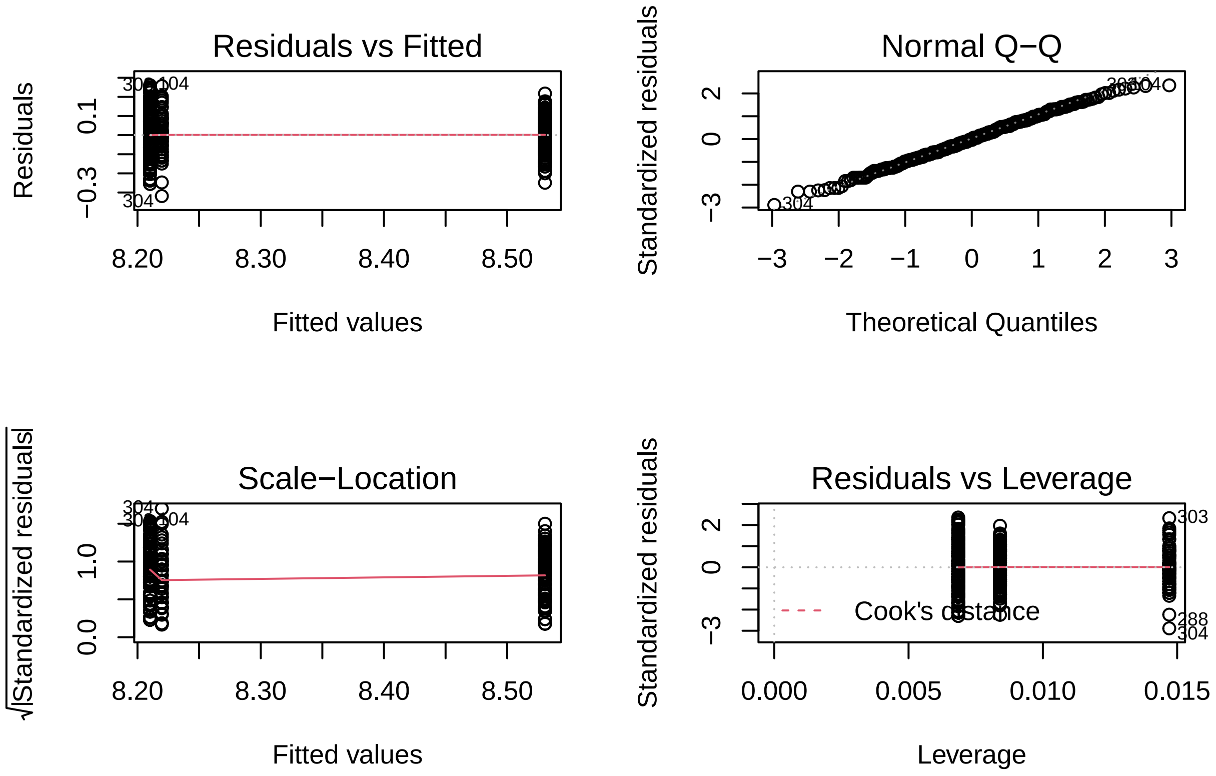
p-value = 0.7587 > 0.05, fail to reject null hypothesis thus variances are equal. Use Welch's t-test

```
t.test(penguins$body_mass_g~penguins$sex, equal.var = TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  penguins$body_mass_g by penguins$sex
## t = -8.7214, df = 330.98, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -0.1993807 -0.1259914
## sample estimates:
## mean in group female   mean in group male
##             8.244592             8.407278
```

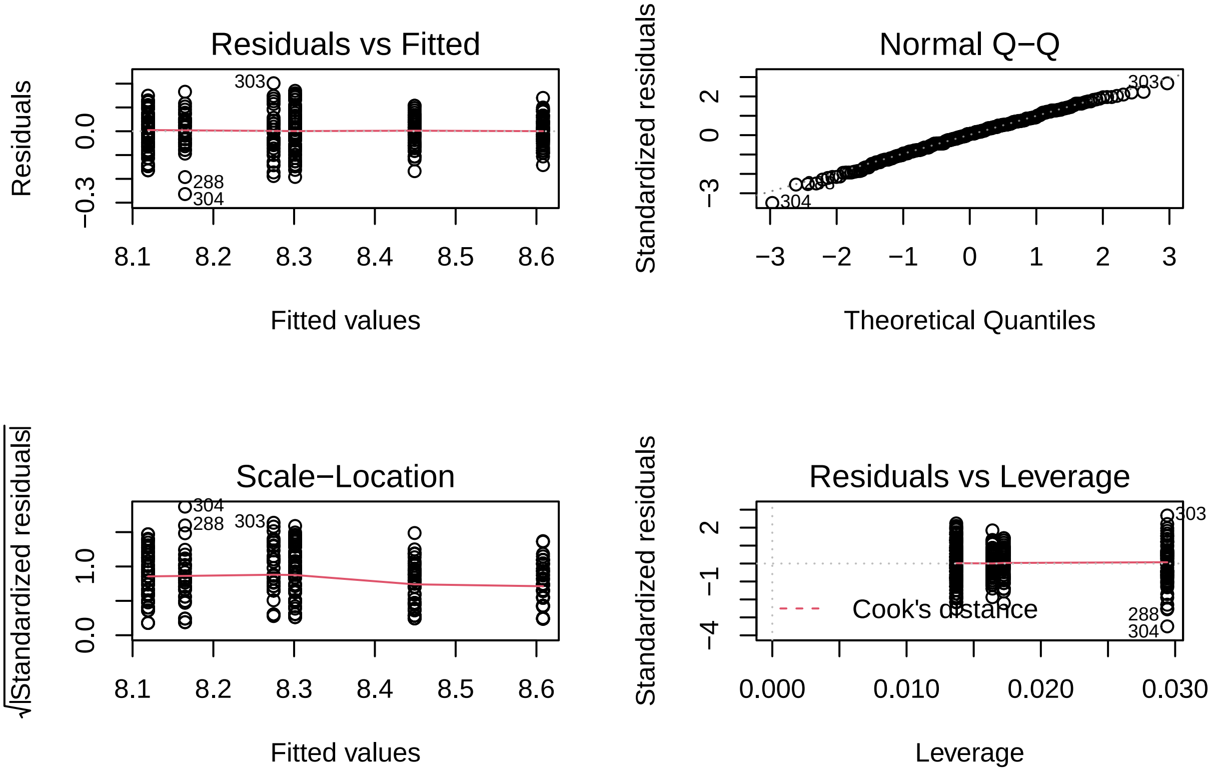
Reject the null hypothesis. There is a difference between the means of male and female groups

```
fit_penguin2 <- lm(body_mass_g~sex, data = penguins)
par(mfrow=c(2,2))
plot(fit_penguin)
```



Two way ANOVA:

```
anova_penguin2 <- aov(body_mass_g~species*sex, data = penguins)
par(mfrow=c(2,2))
plot(anova_penguin2)
```



```
summary(anova_penguin2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## species       2  7.710   3.855 657.900  < 2e-16 ***
## sex           1  2.105   2.105 359.238  < 2e-16 ***
## species:sex   2  0.061   0.030   5.195  0.00601 ** 
## Residuals   327  1.916   0.006                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit_penguin <- lm(body_mass_g~species, data = penguins)
fit_penguin2 <- lm(body_mass_g~species + sex, data = penguins)
fit_penguin3 <- lm(body_mass_g~species + sex + bill_length_mm, data = penguins)
```

```
summary(fit_penguin)
```

```
## 
## Call:
## lm(formula = body_mass_g ~ species, data = penguins)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.31875 -0.07535 -0.00352  0.08184  0.26096 
## 
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.210187   0.009205 891.962   <2e-16 ***
## speciesChinstrap 0.009573   0.016329   0.586    0.558
## speciesGentoo    0.320478   0.013736  23.331   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1112 on 330 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6517
## F-statistic: 311.7 on 2 and 330 DF,  p-value: < 2.2e-16
```

```
summary(fit_penguin2)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239237 -0.053548  0.004346  0.057436  0.190620
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      8.130671   0.007695 1056.658   <2e-16 ***
## speciesChinstrap 0.009573   0.011381    0.841    0.401
## speciesGentoo    0.318474   0.009574   33.263   <2e-16 ***
## sexmale          0.159033   0.008497   18.716   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.07752 on 329 degrees of freedom
## Multiple R-squared:  0.8323, Adjusted R-squared:  0.8308
## F-statistic: 544.5 on 3 and 329 DF,  p-value: < 2.2e-16
```

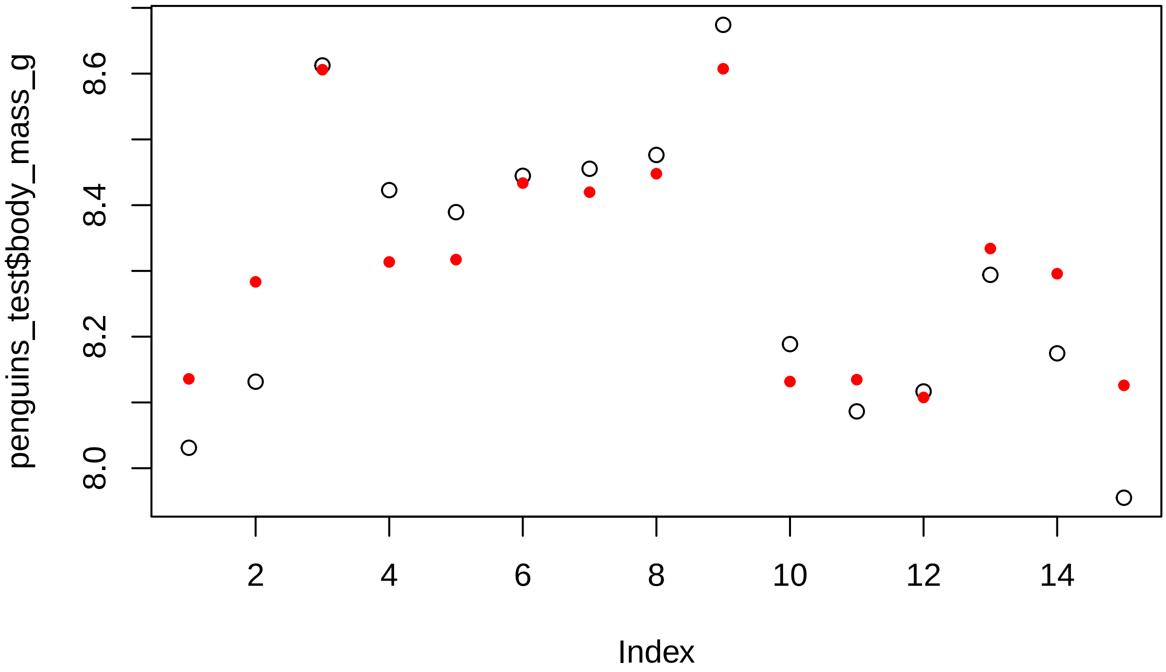
```
summary(fit_penguin3)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex + bill_length_mm, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23956 -0.04722  0.00217  0.04733  0.19692
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.52045    0.14636  51.385  < 2e-16 ***
## speciesChinstrap -0.06628    0.02129  -3.113  0.00202 **
## speciesGentoo     0.25206    0.01845  13.663  < 2e-16 ***
## sexmale           0.13119    0.01064  12.328  < 2e-16 ***
## bill_length_mm    0.10023    0.02401   4.175 3.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07565 on 328 degrees of freedom
## Multiple R-squared:  0.8408, Adjusted R-squared:  0.8389
## F-statistic: 433.1 on 4 and 328 DF,  p-value: < 2.2e-16
```

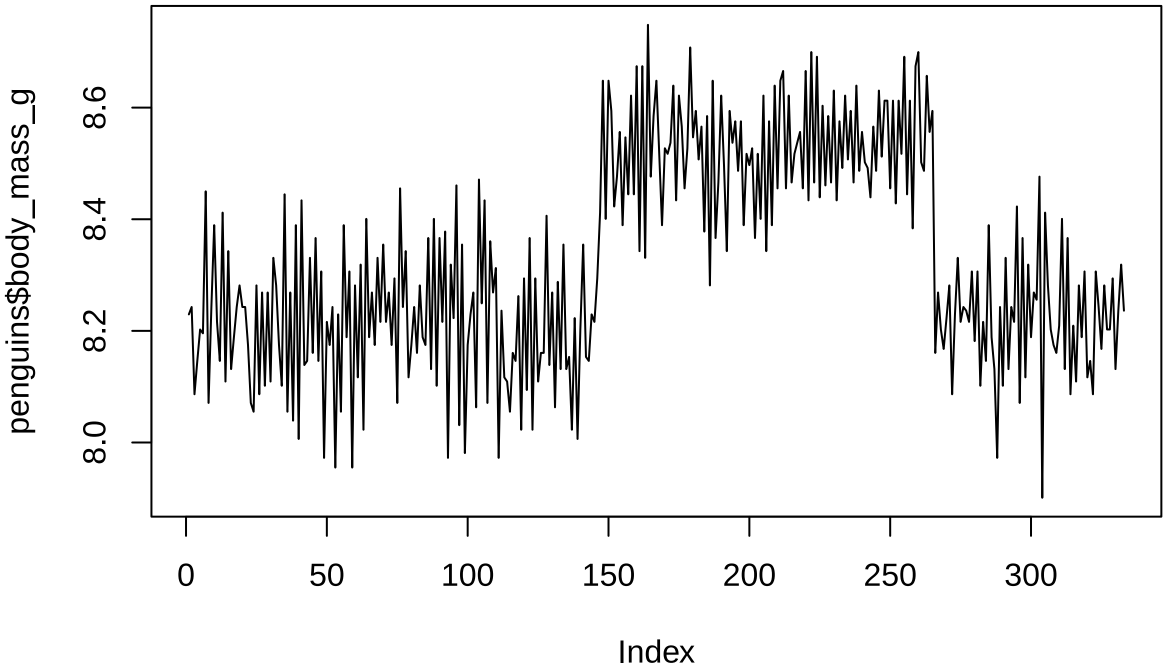
```
# choose the best model
# calculate AIC of each model
result <- AIC(fit_penguin, fit_penguin2, fit_penguin3)
models <- list(fit_penguin, fit_penguin2, fit_penguin3)
result$BIC <- sapply(models, BIC)
model_summary <- lapply(models, summary)
# for loop to extract the R^2 and adj R^2
for (i in 1:length(models)){
result$rsq[i] <- model_summary[[i]]$r.squared
result$adj_rsq[i] <- model_summary[[i]]$adj.r.squared
}
kable(result, digits = 2, align = "c")
```

	df	AIC	BIC	rsq	adj_rsq
fit_penguin	4	-512.70	-497.47	0.65	0.65
fit_penguin2	5	-752.13	-733.09	0.83	0.83
fit_penguin3	6	-767.37	-744.52	0.84	0.84

```
# use 15 random data points as test dataset
splitter <- sample(1:nrow(penguins), 15, replace = F)
penguins_train <- penguins[-splitter,] # leave those rows out of the training data
penguins_test <- penguins[splitter,] # use them to create a set of test data
# train the data
fit_penguin3_split <- lm(body_mass_g~species + sex + bill_length_mm, data = penguins)
# predict
prediction <- predict(fit_penguin3_split, penguins_test)
# plot the predicted points
plot(penguins_test$body_mass_g, pch = 1)
points(prediction, pch = 20, col = "red")
```



```
plot(penguins$body_mass_g, type = "l")
```


penguins$body_mass_g
8.0
8.2
8.4
8.6
0
50
100
150
200
250
300
Index