Penguins

Azriella Turkdogan

2024-12-11

## Penguins

This R Markdown document is made using the “palmerpenguins” package as a data set to show data visualization, and hypothesis testing, and data analysis. Below is the output for loading the packages used (tidyverse, lme4, palmerpenguins, car, multcomp).

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors  
## Loading required package: Matrix  
##   
##   
## Attaching package: 'Matrix'  
##   
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
##   
## Loading required package: carData  
##   
##   
## Attaching package: 'car'  
##   
##   
## The following object is masked from 'package:dplyr':  
##   
## recode  
##   
##   
## The following object is masked from 'package:purrr':  
##   
## some

## Warning: package 'multcomp' was built under R version 4.4.2

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.4.2

## Loading required package: survival  
## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.4.2

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 4.4.2

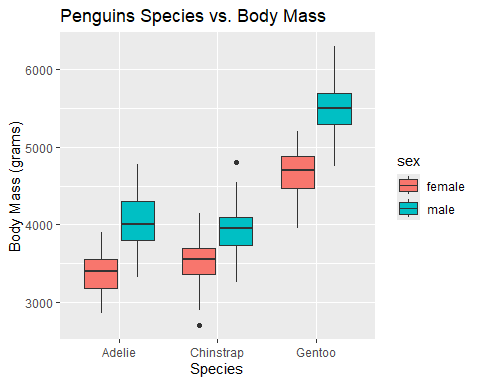
##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select  
##   
##   
## Attaching package: 'TH.data'  
##   
## The following object is masked from 'package:MASS':  
##   
## geyser

## # A tibble: 6 × 8  
## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## <fct> <fct> <dbl> <dbl> <int> <int>  
## 1 Adelie Torgersen 39.1 18.7 181 3750  
## 2 Adelie Torgersen 39.5 17.4 186 3800  
## 3 Adelie Torgersen 40.3 18 195 3250  
## 4 Adelie Torgersen NA NA NA NA  
## 5 Adelie Torgersen 36.7 19.3 193 3450  
## 6 Adelie Torgersen 39.3 20.6 190 3650  
## # ℹ 2 more variables: sex <fct>, year <int>

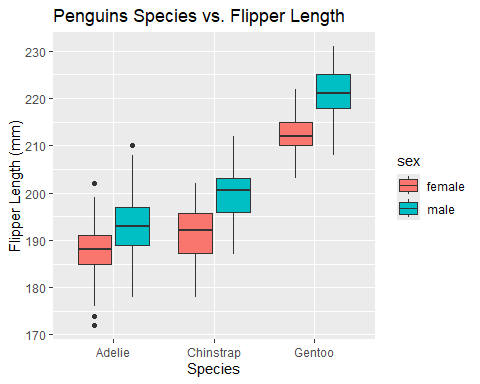
## Data Exploration

### Boxplots

I used boxplots as a way to visualize what is happening with the variables. I included species, sex, body mass, and flipper length for variables.



There seems to be differences shown on body mass vs. the penguin species in the boxplot above. Gentoo especially seems to have a higher mass compared to the other two species. An anova test will be used to test the hypothesis of if the average mass of each species is significantly different. The null hypothesis will be that the species do not have significantly different average body masses.



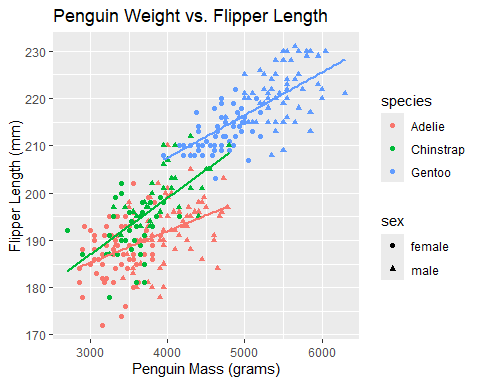
The flipper length between penguin species appears to be differentin the boxplot above. Gentoo seems to have the highest average flipper length among the species, and Adelie appears to have the lowest average flipper length. An anova test will be used to test the hypothesis for if the average flipper lengths of the penguin species are significantly different. The null hypothesis will be that the species do not have significantly different mean flipper lengths.

### Scatter Plot

I used a scatter plot to look at how body mass and flipper length change based on the species. I also included sex for visualization purposes.

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).



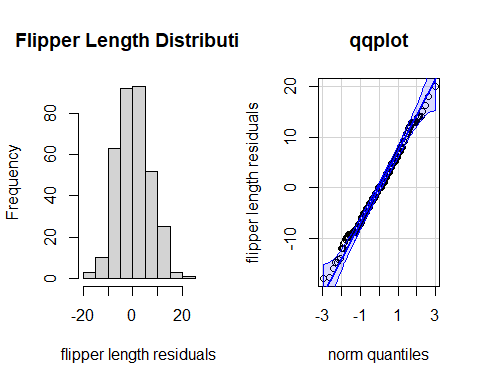
The scatterplot above shows linear trends between the penguin masses and the flipper length. With this scatterplot, a hypothesis was formed that there is a significant linear relationship between body mass and flipper length. The null hypothesis is there is no linear relationship between body mass and flipper length. Linear Models will be used to test these hypotheses.

## Hypothesis Testing

### Anova Testing

An anova test and a Tukey test was used to test the hypothesis for if the means between species for flipper length and body mass are significantly different

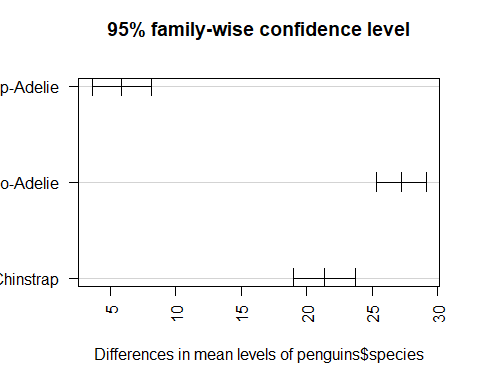
## Df Sum Sq Mean Sq F value Pr(>F)   
## penguins$species 2 52473 26237 594.8 <2e-16 \*\*\*  
## Residuals 339 14953 44   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness



The flipper length Anova test showed a low p-value (<2e-16) and a large F-value of 594.8. This indicates that the averages between penguin species are significantly different.

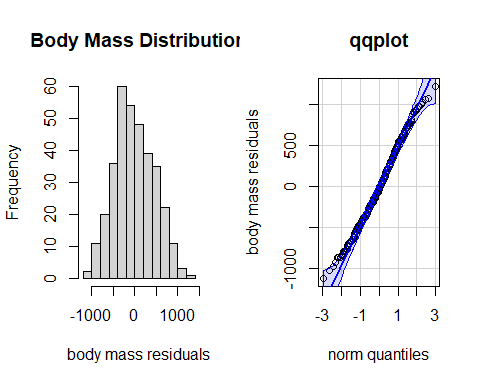
The flipper length distribution shows a normal distribution on the histogram. The qqplot is not a perfect line, but it also shows close to normal distribution.

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = penguins$flipper\_length\_mm ~ penguins$species, data = penguins)  
##   
## $`penguins$species`  
## diff lwr upr p adj  
## Chinstrap-Adelie 5.869887 3.586583 8.153191 0  
## Gentoo-Adelie 27.233349 25.334376 29.132323 0  
## Gentoo-Chinstrap 21.363462 19.000841 23.726084 0



From a Tukey test with a 95% confidence level, it can be seen that the adjusted p-values are all zero. This supports the hypothesis that all of the penguin species have significantly different flipper length averages from each other. The null hypothesis that the flipper lengths are on average the same between species is rejected.

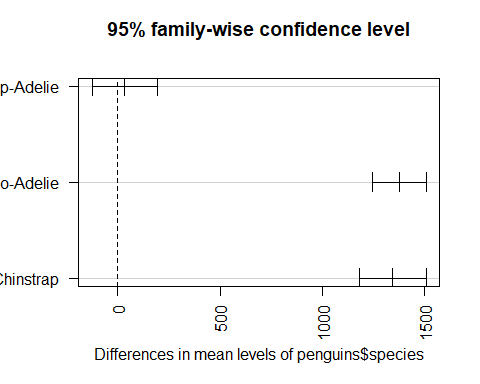
## Df Sum Sq Mean Sq F value Pr(>F)   
## penguins$species 2 146864214 73432107 343.6 <2e-16 \*\*\*  
## Residuals 339 72443483 213698   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness



The body mass Anova test showed a low p-value (<2e-16) and a large F-value of 343.6. This indicates that the averages between penguin species are significantly different.

The body mass distribution shows a normal distribution on the histogram, and the qqplot also shows normal distribution.

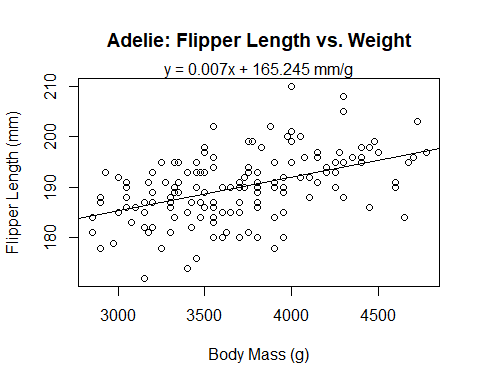
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = penguins$body\_mass\_g ~ penguins$species, data = penguins)  
##   
## $`penguins$species`  
## diff lwr upr p adj  
## Chinstrap-Adelie 32.42598 -126.5002 191.3522 0.8806666  
## Gentoo-Adelie 1375.35401 1243.1786 1507.5294 0.0000000  
## Gentoo-Chinstrap 1342.92802 1178.4810 1507.3750 0.0000000



A Tukey test of body mass with 95% confidence level shows that the species Chinstrap and Adelie do not have significantly different means (p=0.88). However, Gentoo and Chinstrap, as well as Gentoo and Adelie have significantly different mean body masses (p=0). The relationship between chinstrap and Adelie fails to reject the null hypothesis that the penguin body masses are the same on average.

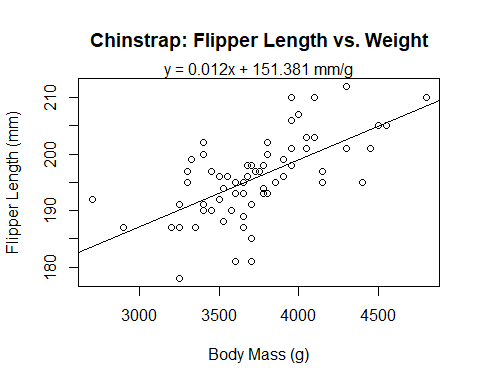
### Linear Models

Linear models were used to test for linearity between penguins’ body mass and flipper length.



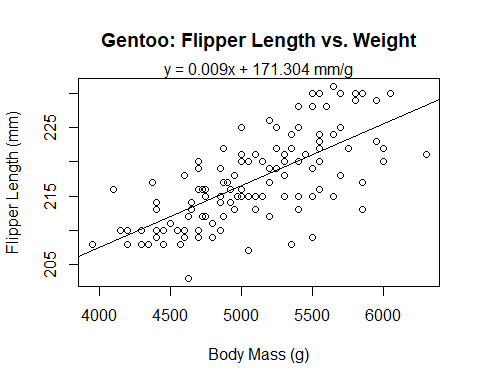
##   
## Call:  
## lm(formula = adelie$flipper\_length\_mm ~ adelie$body\_mass\_g)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.2769 -3.6192 0.0569 3.4696 18.0477   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.652e+02 3.849e+00 42.929 < 2e-16 \*\*\*  
## adelie$body\_mass\_g 6.677e-03 1.032e-03 6.468 1.34e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.798 on 149 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.2192, Adjusted R-squared: 0.214   
## F-statistic: 41.83 on 1 and 149 DF, p-value: 1.343e-09

The linear model for the Adelie penguins shows a significant linear relationship with a slope of 0.00667 mm/gram (p = 1.34e-09). This model has 149 degrees of freedom and a low R-squared value at 0.214, indicating that body mass has a weak correlation with flipper length.



##   
## Call:  
## lm(formula = chinstrap$flipper\_length\_mm ~ chinstrap$body\_mass\_g)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.4296 -3.3315 0.4097 2.8889 11.5941   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.514e+02 6.575e+00 23.024 < 2e-16 \*\*\*  
## chinstrap$body\_mass\_g 1.191e-02 1.752e-03 6.795 3.75e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.512 on 66 degrees of freedom  
## Multiple R-squared: 0.4116, Adjusted R-squared: 0.4027   
## F-statistic: 46.17 on 1 and 66 DF, p-value: 3.748e-09

The linear model for the Chinstrap penguins shows a significant linear relationship with a slope of 0.0119 mm/gram (p = 3.75e-09). This model has 66 degrees of freedom and a moderate R-squared value at 0.4027, indicating that body mass has a moderate correlation with flipper length.



##   
## Call:  
## lm(formula = gentoo$flipper\_length\_mm ~ gentoo$body\_mass\_g)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.0194 -2.7401 0.1781 2.9859 8.9806   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.713e+02 4.244e+00 40.36 <2e-16 \*\*\*  
## gentoo$body\_mass\_g 9.039e-03 8.321e-04 10.86 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.633 on 121 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.4937, Adjusted R-squared: 0.4896   
## F-statistic: 118 on 1 and 121 DF, p-value: < 2.2e-16

The linear model for the Gentoo penguins shows a significant linear relationship with a slope of 0.00904 mm/gram (p = 2.2e-16). This model has 121 degrees of freedom and a high R-squared value at 0.4937, indicating that body mass has a high correlation with flipper length.

The hypothesis that there is a linear relationship between penguin flipper length and body mass is supported by the linear models above for all three species of penguins. The null hypothesis that there is no linear relationship between body mass and flipper length is rejected.

#### Citations

R Core Team (2024). \_R: A Language and Environment for Statistical Computing\_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” \_Journal of Open Source Software\_, \*4\*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

Fox J, Weisberg S (2019). \_An R Companion to Applied Regression\_, Third edition. Sage, Thousand Oaks CA. https://www.john-fox.ca/Companion/.

Hothorn T, Bretz F, Westfall P (2008). “Simultaneous Inference in General Parametric Models.” \_Biometrical Journal\_, \*50\*(3), 346-363.