# Data exploration using the palmerpenguins dataset

Julien Martin

18 January, 2021

## Data exploration

### Motivation

In this section, we **explore** the data from package `palmerpenguins`. A recent publication from the researcher, Dr Kristen Gorman, who shared the data is Connors et al. (2020).

### Data

The data are displayed below (first 10 rows) :

```
penguins %>%
  slice(1:10) %>%
  knitr::kable()
```

| species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex | year |
|---|---|---|---|---|---|---|---|
| Adelie | Torgersen | 39.1 | 18.7 | 181 | 3750 | male | 2007 |
| Adelie | Torgersen | 39.5 | 17.4 | 186 | 3800 | female | 2007 |
| Adelie | Torgersen | 40.3 | 18.0 | 195 | 3250 | female | 2007 |
| Adelie | Torgersen | NA | NA | NA | NA | NA | 2007 |
| Adelie | Torgersen | 36.7 | 19.3 | 193 | 3450 | female | 2007 |
| Adelie | Torgersen | 39.3 | 20.6 | 190 | 3650 | male | 2007 |
| Adelie | Torgersen | 38.9 | 17.8 | 181 | 3625 | female | 2007 |
| Adelie | Torgersen | 39.2 | 19.6 | 195 | 4675 | male | 2007 |
| Adelie | Torgersen | 34.1 | 18.1 | 193 | 3475 | NA | 2007 |
| Adelie | Torgersen | 42.0 | 20.2 | 190 | 4250 | NA | 2007 |

### Numerical exploration

There are 344 penguins in the dataset, and 3 different species. The data were collected in 3 islands of the Palmer archipelago in Antarctica.

The mean of all traits that were measured on the penguins are:

```
## # A tibble: 3 x 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g  year
##   <fct>              <dbl>         <dbl>             <dbl>       <dbl> <dbl>
## 1 Adelie              38.8          18.3              190.       3701. 2008.
## 2 Chinstrap           48.8          18.4              196.       3733. 2008.
## 3 Gentoo              47.5          15.0              217.       5076. 2008.
```

## Graphical exploration

A histogram of body mass per species:

```
penguins %>%
  ggplot() +
  aes(x = body_mass_g) +
  geom_histogram(aes(fill = species),
                 alpha = 0.5,
                 position = "identity") +
  scale_fill_manual(values = c("darkorange","purple","cyan4")) +
  theme_minimal() +
  labs(x = "Body mass (g)",
       y = "Frequency",
       title = "Penguin body mass")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

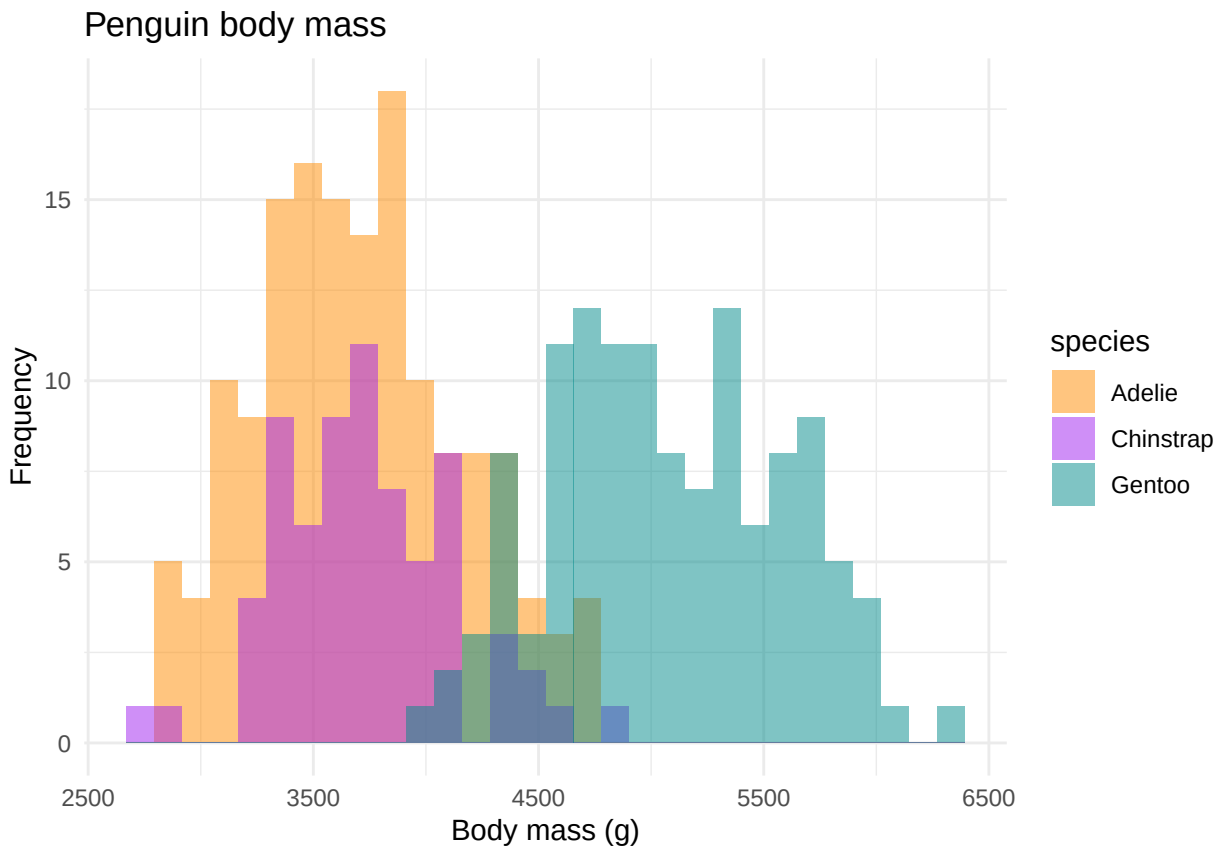



Figure 1: Distribution of body mass by species of penguins

## Linear regression

And here is a nice model with graphical output

```
m1 <- lm(bill_length_mm ~  flipper_length_mm + body_mass_g + sex, data = penguins)
summary(m1)
```

```
##
```

```
## Call:
## lm(formula = bill_length_mm ~ flipper_length_mm + body_mass_g +
##     sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6131 -2.8005 -0.6307  2.1699 20.1682
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.0754117  4.6732129  -1.728   0.0849 .
## flipper_length_mm  0.2672650  0.0335203   7.973 2.57e-14 ***
## body_mass_g       -0.0006670  0.0006232  -1.070   0.2853
## sexmale            2.3047154  0.5055670   4.559 7.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.029 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.4621, Adjusted R-squared:  0.4572
## F-statistic:  94.2 on 3 and 329 DF,  p-value: < 2.2e-16
```

```
par(mfrow= c(2,2))
plot(m1)
```

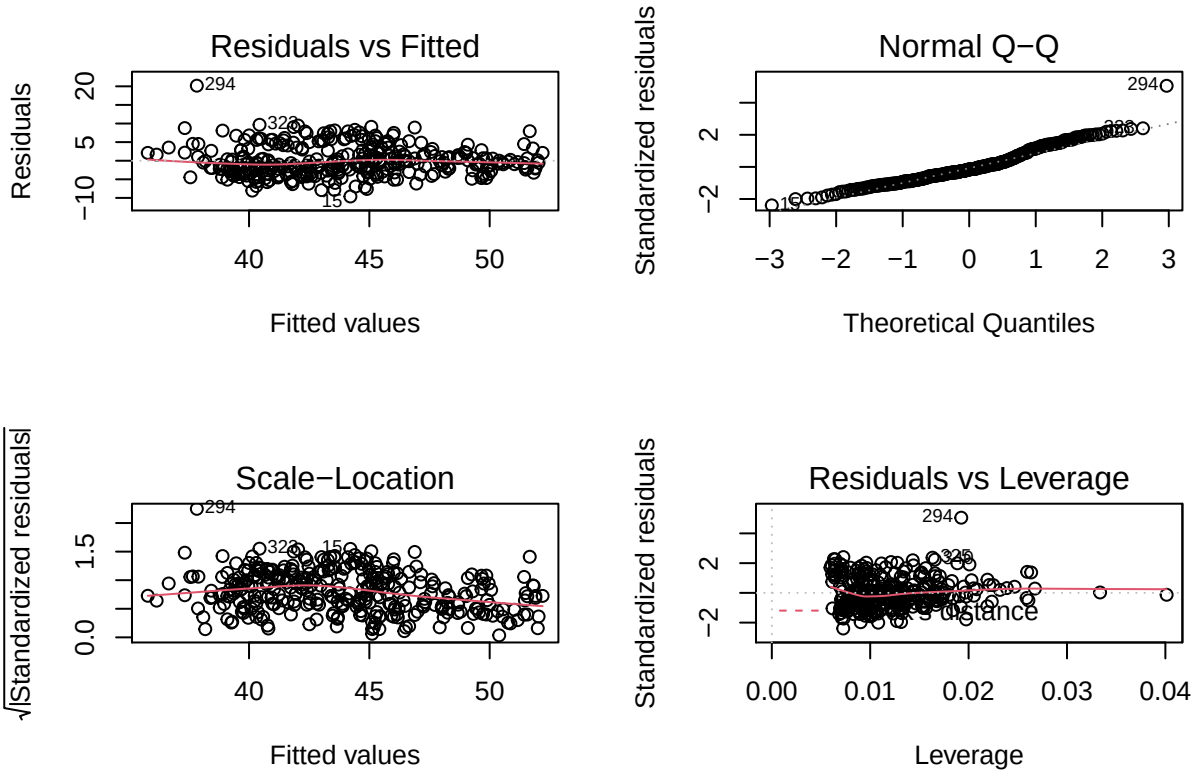
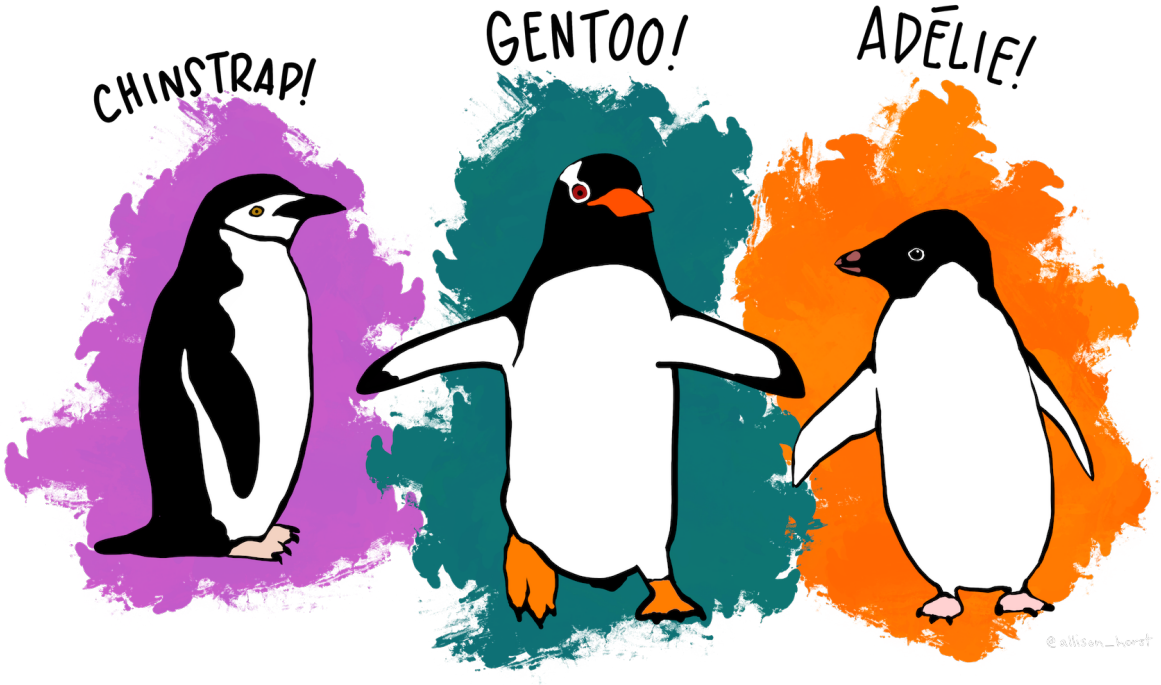



Figure 2: Checking assumptions of the model

## The end

The 3 species of penguins:




## References

Connors, Brendan, Michael J. Malick, Gregory T. Ruggerone, Pete Rand, Milo Adkison, James R. Irvine, Robert Campbell, and Kristen Gorman. 2020. "Climate and Competition Influence Sockeye Salmon Population Dynamics Across the Northeast Pacific Ocean." *Canadian Journal of Fisheries and Aquatic Sciences* 77 (6): 943–49. https://doi.org/10.1139/cjfas-2019-0422.