PSTAT 126

Lab 10

Roupen Khanjian

Spring 2021

```
library(faraway) # Functions and Datasets for Books by Julian Faraway
library(alr4) # Data to Accompany Applied Linear Regression 4th Edition
library(tidyverse) # Easily Install and Load the 'Tidyverse'
library(patchwork) # The Composer of Plots
library(GGally) # Extension to 'ggplot2'
library(janitor) # Simple Tools for Examining and Cleaning Dirty Data
library(palmerpenguins) # Palmer Archipelago (Antarctica) Penguin Data
library(broom) # Convert Statistical Objects into Tidy Tibbles
library(lmtest) # Testing Linear Regression Models
library(ballr) # Access to Current and Historical Basketball Data
```

Contents

Review of Linear Regression Assumptions

Linear Regression Model Assumptions

1) The relationship between each Y_n and each x_n, respectively, is linear. **L**inearity
2) Errors have **E**qual variance. $\text{Var}(Y_n) = \sigma^2$ for every n (homoscedasticity)
3) Errors are **N**ormally distributed
4) Errors are **I**ndependent

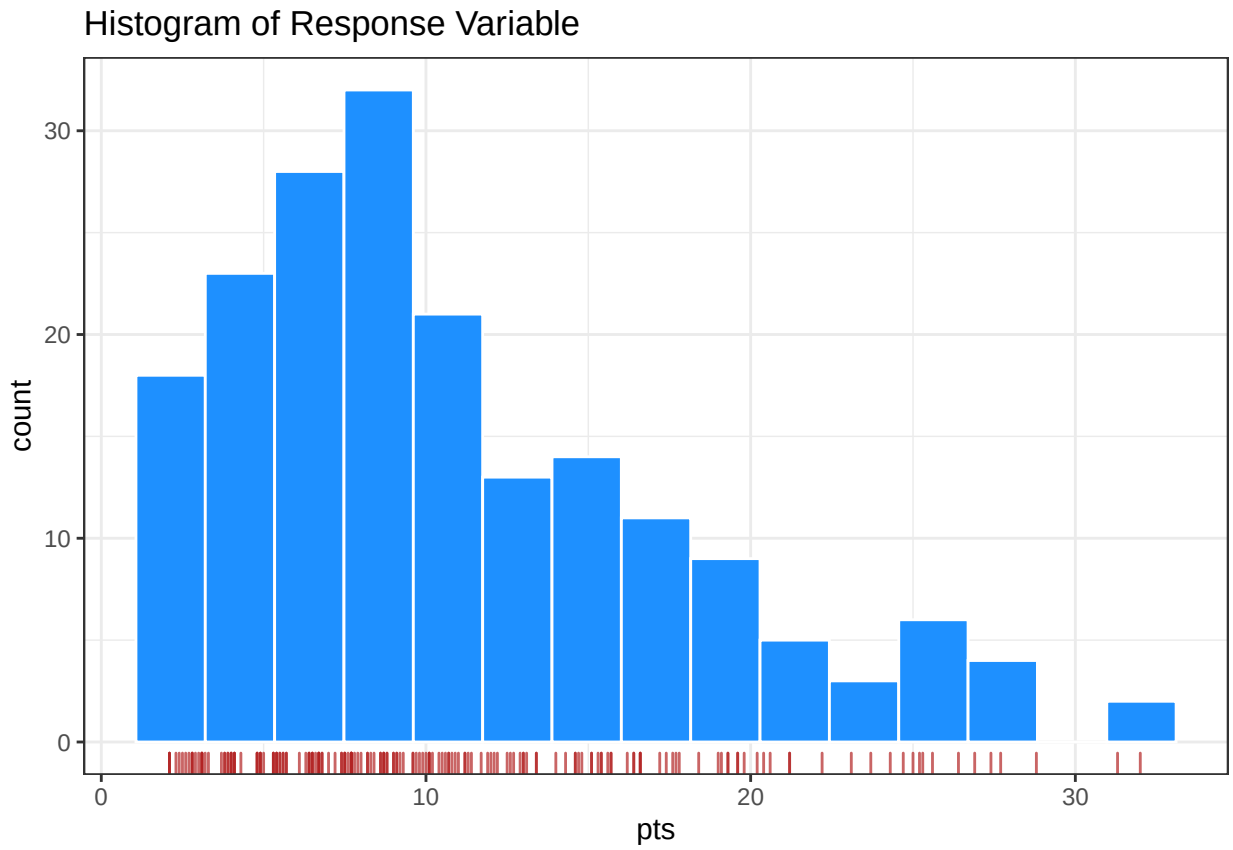
- Can use the acronym **L.I.N.E.** to help you remember.

Model 1: Not Satisfying Assumptions

```
nba_data1 <- NBAPerGameStatistics(season = 2021)

nba_data <- distinct(nba_data1, player, .keep_all = TRUE)  %>%
  filter(pts > 2) %>%
  filter(g > 10) %>%
  filter(pos %in% c("PG", "SG")) %>%
  mutate(pos = factor(pos))
```

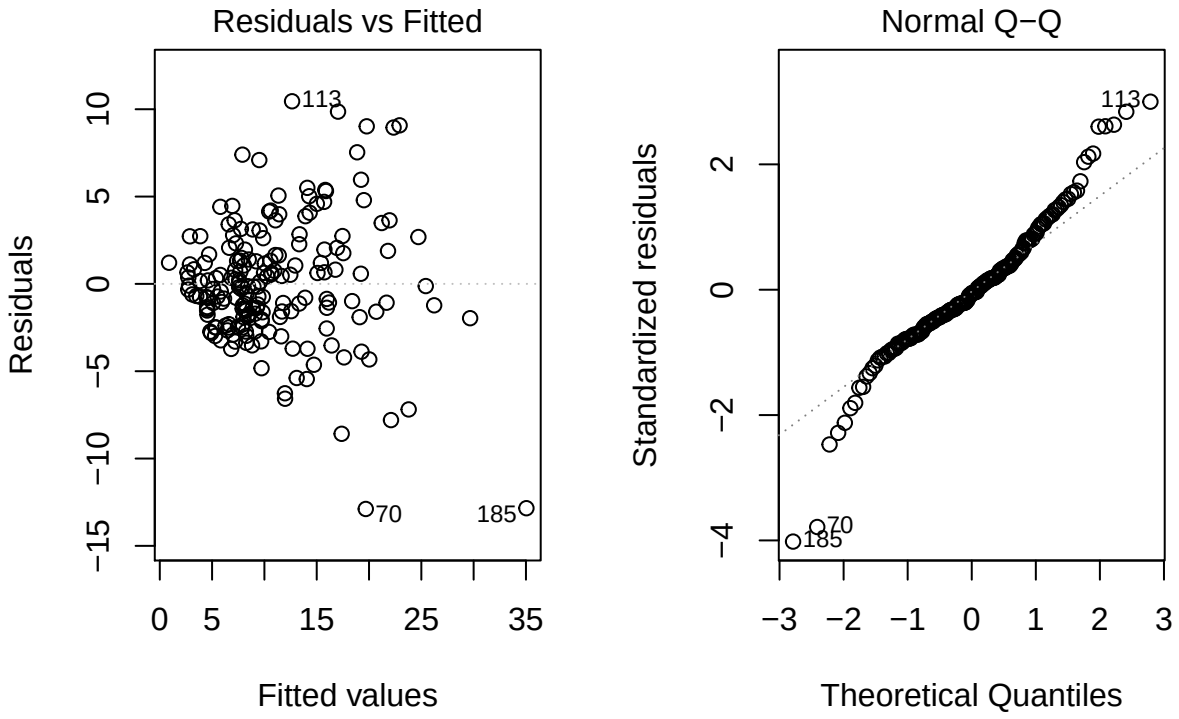
```
ggplot(data = nba_data,
       aes(x = pts)) +
  geom_histogram(fill = "dodgerblue", color = "white",
                 bins = 15) +
  geom_rug(color = "firebrick", alpha = 0.7) +
  labs(title = "Histogram of Response Variable") +
  theme_bw()
```




```
model_nba1 <- lm(pts ~ pos + tov + trb, data = nba_data)
summary(model_nba1)
```

```
##
## Call:
## lm(formula = pts ~ pos + tov + trb, data = nba_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8875  -1.8923  -0.1926   1.6874  10.4546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5012     0.6360   0.788  0.43163
## posSG         1.6473     0.5384   3.059  0.00255 **
## tov           5.3431     0.4299  12.430  < 2e-16 ***
## trb           0.7735     0.2467   3.135  0.00200 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.507 on 185 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.7319
## F-statistic: 172.1 on 3 and 185 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,2))
plot(model_nba1, which = 1, add.smooth = F) # resid vs fit
plot(model_nba1, which = 2) # qqplot
```




```
shapiro.test(resid(model_nba1))  # resid vs fit
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_nba1)
## W = 0.96241, p-value = 6.074e-05
```

```
bptest(model_nba1) # test for homoscedasticity
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_nba1
## BP = 42.618, df = 3, p-value = 2.966e-09
```

```
dwtest(model_nba1) # test for autocorrelation
```

```
##
##  Durbin-Watson test
##
## data:  model_nba1
## DW = 1.8996, p-value = 0.241
## alternative hypothesis: true autocorrelation is greater than 0
```

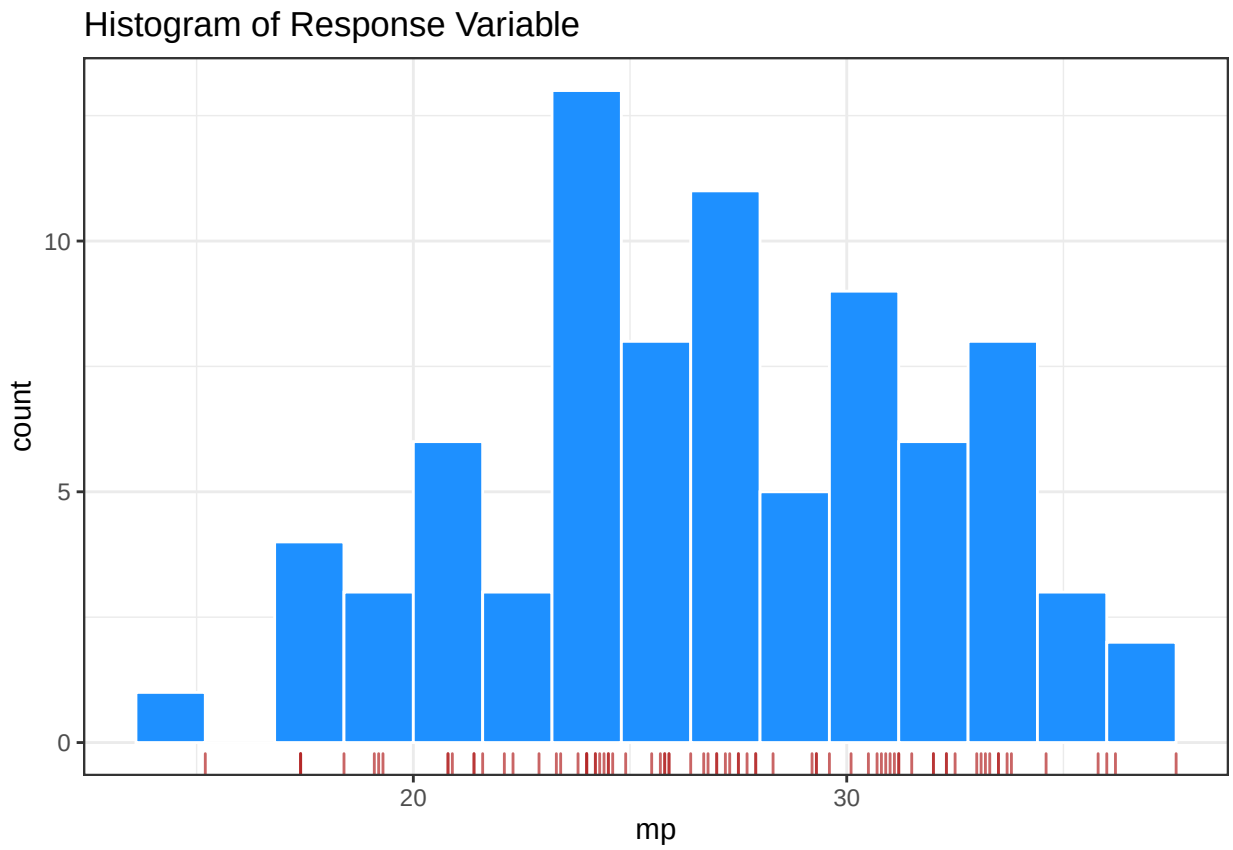
```
gqtest(model_nba1) # test for homoscedasticity
```

```
##
##  Goldfeld-Quandt test
##
## data:  model_nba1
## GQ = 1.1463, df1 = 91, df2 = 90, p-value = 0.2588
## alternative hypothesis: variance increases from segment 1 to 2
```

Model 2: Satisfying Assumptions

```
nba_data <- distinct(nba_data1, player, .keep_all = TRUE)  %>%
  filter(pts > 8) %>%
  filter(g > 20) %>%
  filter(pos %in% c("PF", "C")) %>%
  mutate(pos = factor(pos))
```

```
par(mfrow = c(1,1))
ggplot(data = nba_data,
       aes(x = mp)) +
  geom_histogram(fill = "dodgerblue", color = "white",
                 bins = 15) +
  geom_rug(color = "firebrick", alpha = 0.7) +
  labs(title = "Histogram of Response Variable") +
  theme_bw()
```

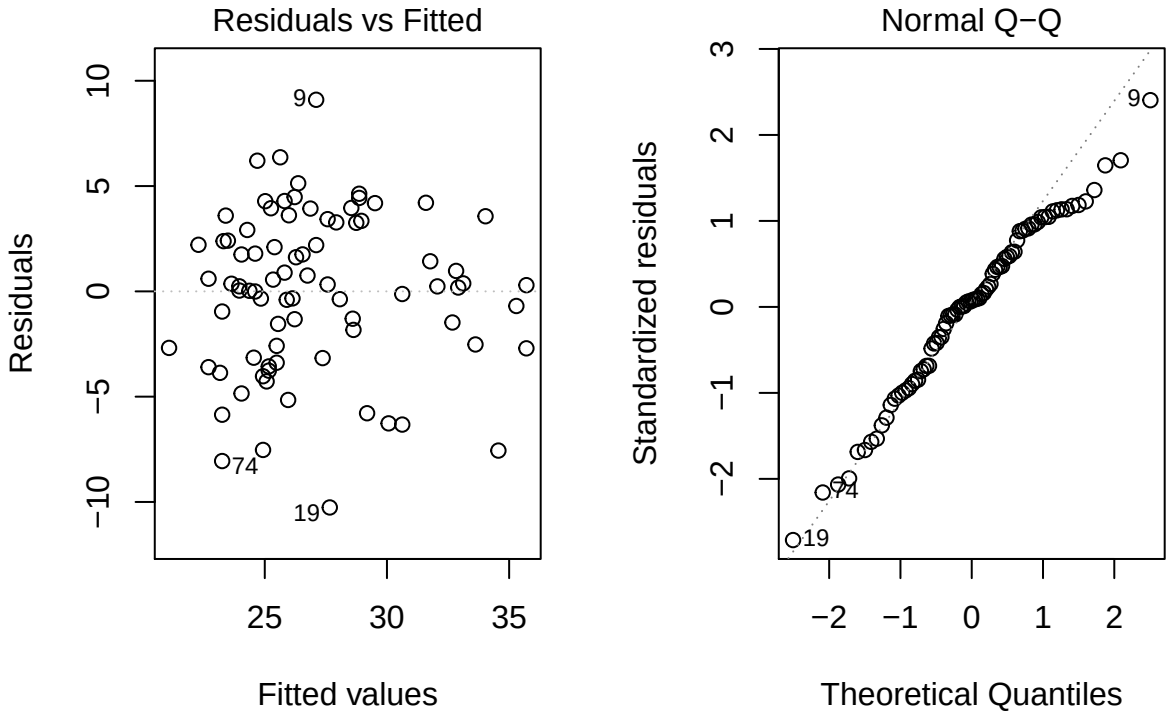



```
model_nba2 <- lm(mp ~ stl + tov, data = nba_data)
summary(model_nba2)
```

```
##
## Call:
## lm(formula = mp ~ stl + tov, data = nba_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2604  -2.6602   0.2668   3.1706   9.0972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.0310     1.2109  14.891  < 2e-16 ***
## stl           5.5758     1.5829   3.522 0.000714 ***
## tov           3.2305     0.6505   4.966 3.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.808 on 79 degrees of freedom
## Multiple R-squared:  0.4605, Adjusted R-squared:  0.4468
## F-statistic: 33.71 on 2 and 79 DF,  p-value: 2.597e-11
```

```
par(mfrow = c(1,2))
plot(model_nba2, which = 1, add.smooth = F) # resid vs fit
plot(model_nba2, which = 2) # qqplot
```




```
shapiro.test(resid(model_nba2)) # resid vs fit
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_nba2)
## W = 0.9811, p-value = 0.269
```

```
bptest(model_nba2) # test for homoscedasticity
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  model_nba2
## BP = 0.021789, df = 2, p-value = 0.9892
```

```
dwtest(model_nba2) # test for autocorrelation
```

```
##
##  Durbin-Watson test
##
## data:  model_nba2
## DW = 2.4024, p-value = 0.9682
## alternative hypothesis: true autocorrelation is greater than 0
```

```
gqtest(model_nba2) # test for homoscedasticity
```

```
##
##  Goldfeld-Quandt test
##
## data:  model_nba2
## GQ = 1.0151, df1 = 38, df2 = 38, p-value = 0.4817
## alternative hypothesis: variance increases from segment 1 to 2
```

Generalized Linear Models: Logistic Regression

- Last week we used species from the penguins dataset as a categorical predictor variable. Today let's use it as our response variable in a logistic regression example.

Data

```
penguins_noNA_no_Gentoo <- penguins %>%
  drop_na() %>%
  filter(species != "Gentoo" )%>%
  mutate(species = fct_drop(species))

slice_sample(penguins_noNA_no_Gentoo, n = 5)
```

```
## # A tibble: 5 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Dream            40.6          17.2              187        3475 male
## 2 Chinst~ Dream            51.9          19.5              206        3950 male
## 3 Adelie  Biscoe           38.6          17.2              199        3750 fema~
## 4 Adelie  Dream            41.1          18.1              205        4300 male
## 5 Adelie  Biscoe           38.8          17.2              180        3800 male
## # ... with 1 more variable: year <int>
```

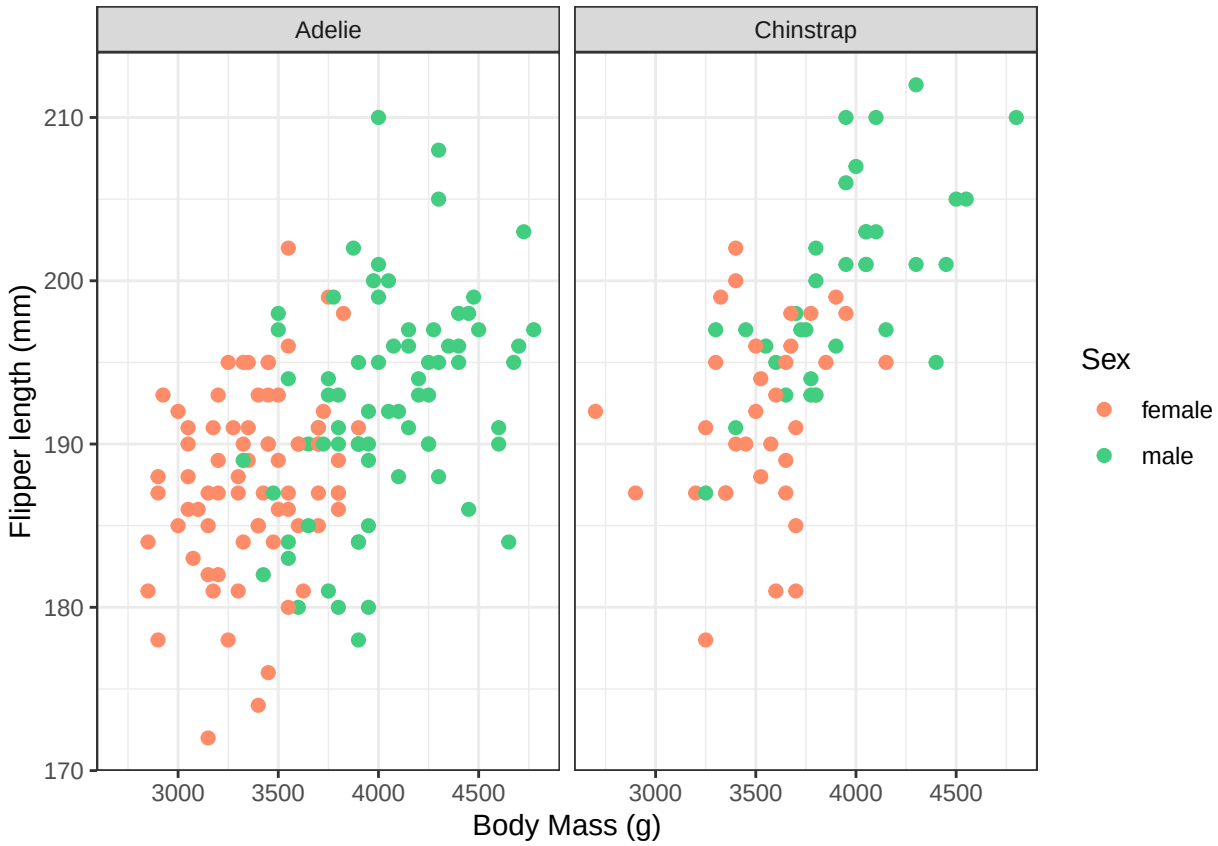
```
levels(penguins_noNA_no_Gentoo$species)
```

```
## [1] "Adelie"    "Chinstrap"
```

EDA

```
mycolors <- c("salmon1", "seagreen3")

par(mfrow = c(1,1))
ggplot(data = penguins_noNA_no_Gentoo,
       aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(aes(color = sex), size = 2) +
  scale_color_manual(values = mycolors) +
  facet_wrap(~ species) +
  labs(x = "Body Mass (g)",
       y = "Flipper length (mm)",
       color = "Sex") +
  theme_bw()
```



Logistic Regression Model

```
model_species <- glm(species ~ body_mass_g + flipper_length_mm + sex,
                     data = penguins_noNA_no_Gentoo,
                     family = "binomial")

summary(model_species)
```

```
##
## Call:
## glm(formula = species ~ body_mass_g + flipper_length_mm + sex,
##     family = "binomial", data = penguins_noNA_no_Gentoo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9456  -0.8195  -0.5581   1.0065   2.3698
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.161e+01  5.784e+00  -5.465 4.63e-08 ***
## body_mass_g       -8.576e-04  5.428e-04  -1.580    0.114
## flipper_length_mm  1.778e-01  3.236e-02   5.495 3.91e-08 ***
## sexmale           -5.716e-01  4.499e-01  -1.270    0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 267.57  on 213  degrees of freedom
## Residual deviance: 226.45  on 210  degrees of freedom
## AIC: 234.45
##
## Number of Fisher Scoring iterations: 4
```

```
tidy(model_species)
```

```
## # A tibble: 4 x 5
##   term                estimate std.error statistic      p.value
##   <chr>                  <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)       -31.6       5.78         -5.47 0.0000000463
## 2 body_mass_g        -0.000858  0.000543     -1.58 0.114
## 3 flipper_length_mm   0.178     0.0324        5.49 0.0000000391
## 4 sexmale            -0.572     0.450        -1.27 0.204
```

```
coef(model_species)
```

```
##       (Intercept)       body_mass_g flipper_length_mm           sexmale
##     -3.160824e+01     -8.576326e-04      1.778080e-01     -5.715776e-01
```

Likelihood Ratio Test

```
model_species_smaller <- glm(species ~ flipper_length_mm + body_mass_g,
                     data = penguins_noNA_no_Gentoo,
                     family = "binomial")

model_species_larger <- glm(species ~ body_mass_g + flipper_length_mm + sex + bill_depth_mm,
                     data = penguins_noNA_no_Gentoo,
                     family = "binomial")

anova(model_species_smaller,
      model_species_larger,
      test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: species ~ flipper_length_mm + body_mass_g
## Model 2: species ~ body_mass_g + flipper_length_mm + sex + bill_depth_mm
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       211     228.09
## 2       209     226.43  2   1.6609   0.4359
```

Prediction

What's the probabilty that a male penguin weighing 4150 grams with a flipper length of 205 mm is a Chinstrap?

```
new_data <- data.frame(
  body_mass_g = 4150,
  flipper_length_mm = 205,
  sex = "male")

predict(model_species, newdata = new_data, se.fit = TRUE,
        type = "response")$fit
```

```
##         1
## 0.6707634
```

What's the probability that a female penguin weighing 4950 grams with a flipper length of 190 mm is a Chinstrap?

```
new_data2 <- data.frame(
  body_mass_g = 4950,
  flipper_length_mm = 190,
  sex = "female")

predict(model_species, newdata = new_data2, se.fit = TRUE,
        type = "response")$fit
```

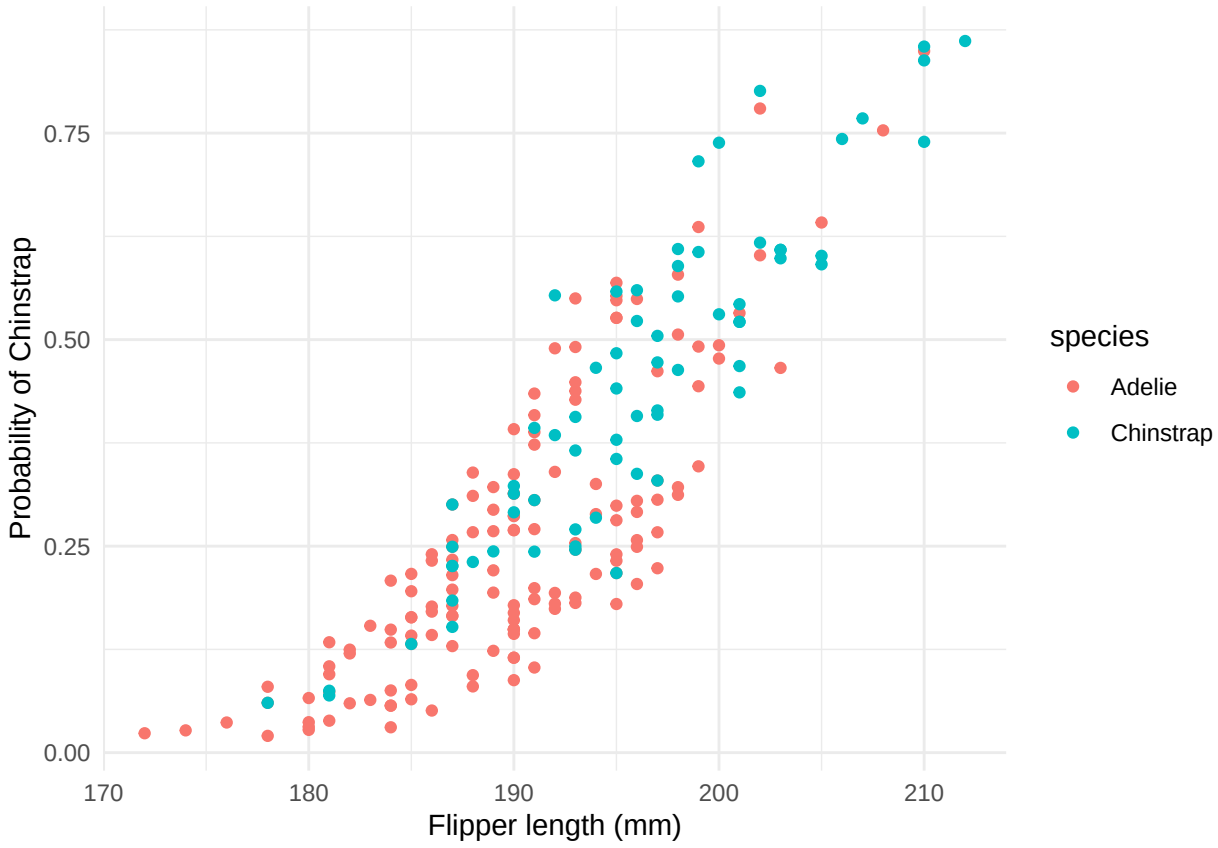
```
##         1
## 0.1120462
```

```
model_fitted <- augment(model_species, type.predict = "response")

slice_sample(model_fitted, n = 5)[1:5]
```

```
## # A tibble: 5 x 5
##   species   body_mass_g flipper_length_mm sex    .fitted
##   <fct>           <int>             <int> <fct>    <dbl>
## 1 Adelie           4350               196 male     0.257
## 2 Adelie           3950               190 male     0.144
## 3 Chinstrap        3400               191 male     0.243
## 4 Chinstrap        4400               195 male     0.217
## 5 Adelie           3700               187 female   0.178
```

```
ggplot(data = model_fitted, aes(x = flipper_length_mm, y = .fitted)) +
  geom_point(aes(color = species)) +
  labs(x = "Flipper length (mm)",
       y = "Probability of Chinstrap") +
  theme_minimal()
```


Probability of Chinstrap
0.00
0.25
0.50
0.75
170
180
190
200
210
Flipper length (mm)
species
Adelie
Chinstrap