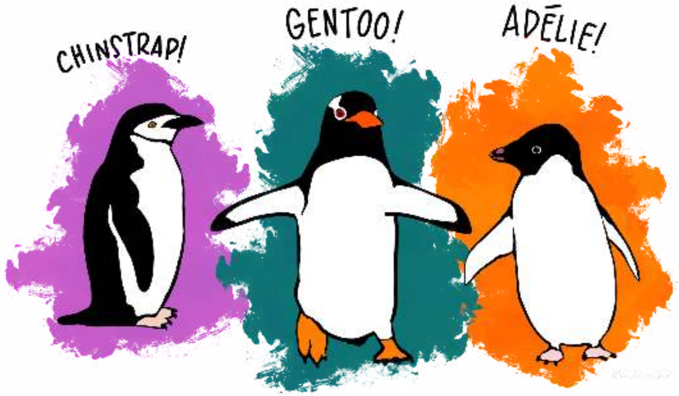
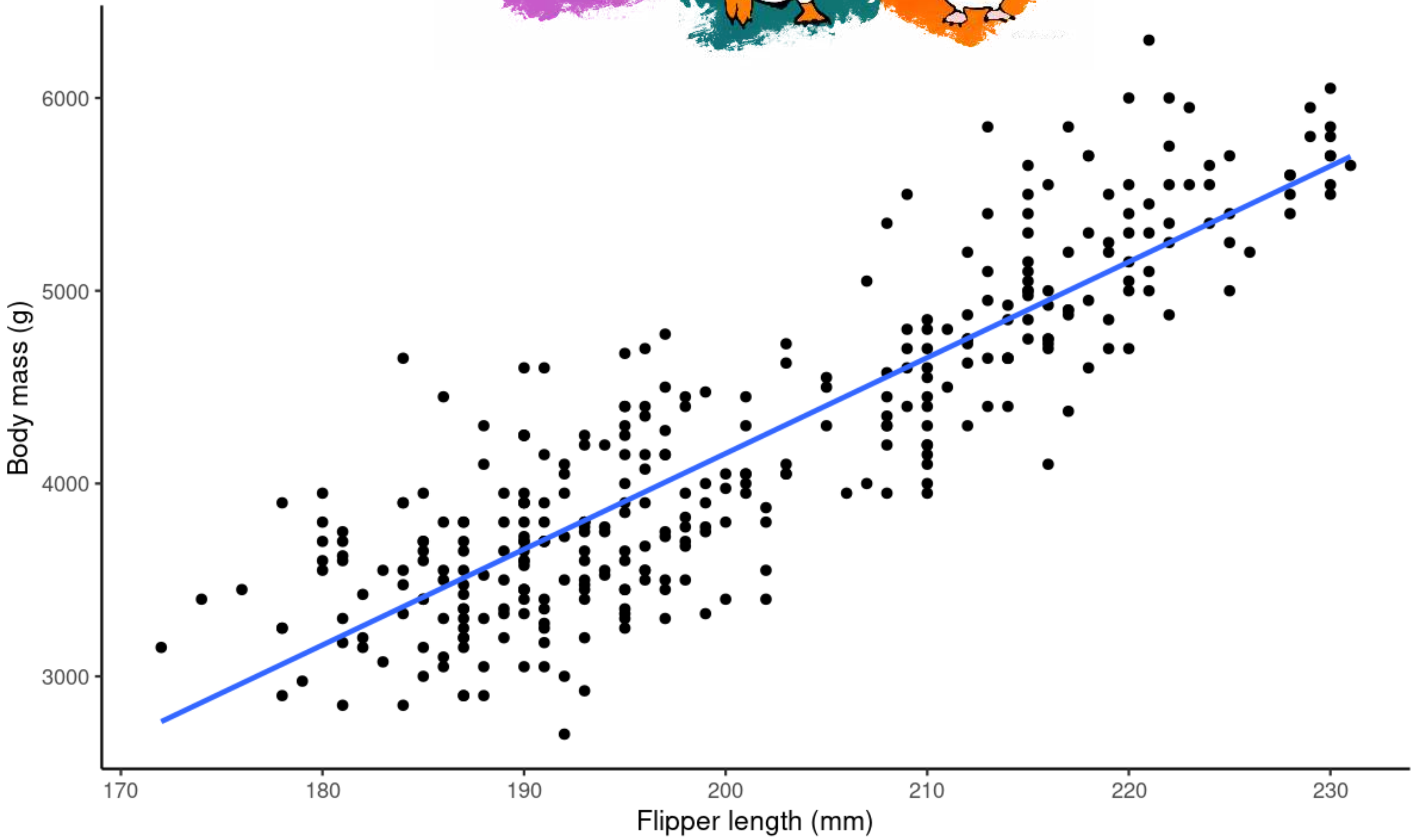
NRI 7350

Linear Models

Regressions,
ANOVAs, and
Model assumptions





Artwork by [@allison_horst]

Getting started (again)

Open RStudio

Open your NRI project

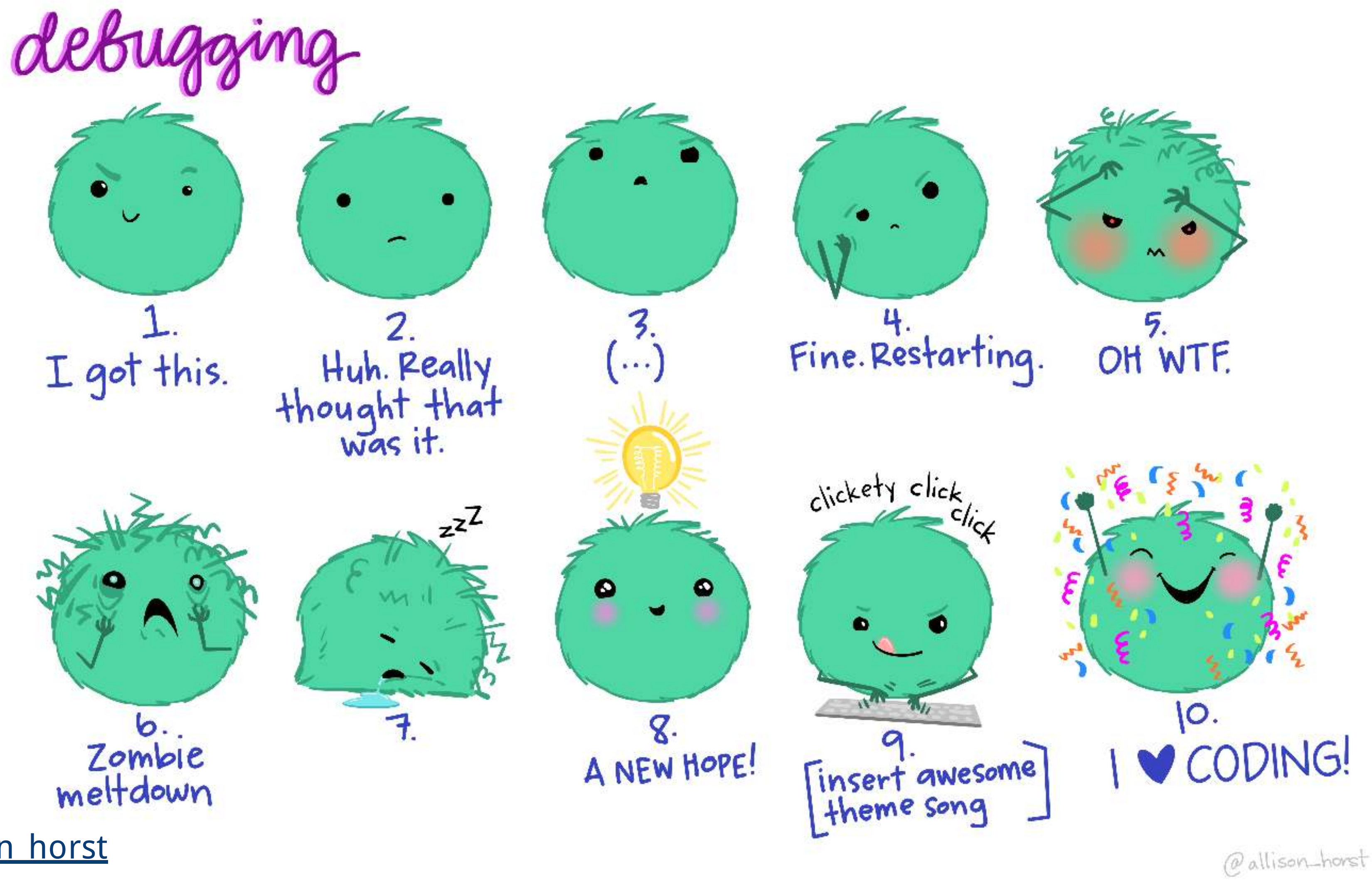
Open a **new** script for today:

File > New File > R Script

Make sure to load packages at the top:

```
library(tidyverse)
library(palmerpenguins)
```

How Are we Doing?



Artwork by [@allison_horst]

Linear Models

Linear Models

Running models in R

```
lm(y ~ x1 + x2, data = data)
```

- **y** is the **response** variable (**dependent**)
- **x** are the **explanatory** variables (**independent**, **predictor**)

Here we're assuming a **continuous y**

Linear Models

Running models in R

```
lm(y ~ x1 + x2, data = data)
```

- **y** is the **response** variable (**dependent**)
- **x** are the **explanatory** variables (**independent**, **predictor**)

Here we're assuming a **continuous y**

Different types of models

- If we only have one **x** which is continuous, this is a **simple linear regression**
- If both **x** are continuous, this is a **multiple linear regression**
- If both **x** are categorical, this is an **ANOVA**
- If **x1** is continuous and **x2** is categorical, this is an **ANCOVA**

Linear Models

Running models in R

```
lm(y ~ x1 + x2, data = data)
```

- **y** is the **response** variable (**dependent**)
- **x** are the **explanatory** variables (**independent**, **predictor**)

Here we're assuming a **continuous y**

Different types of models

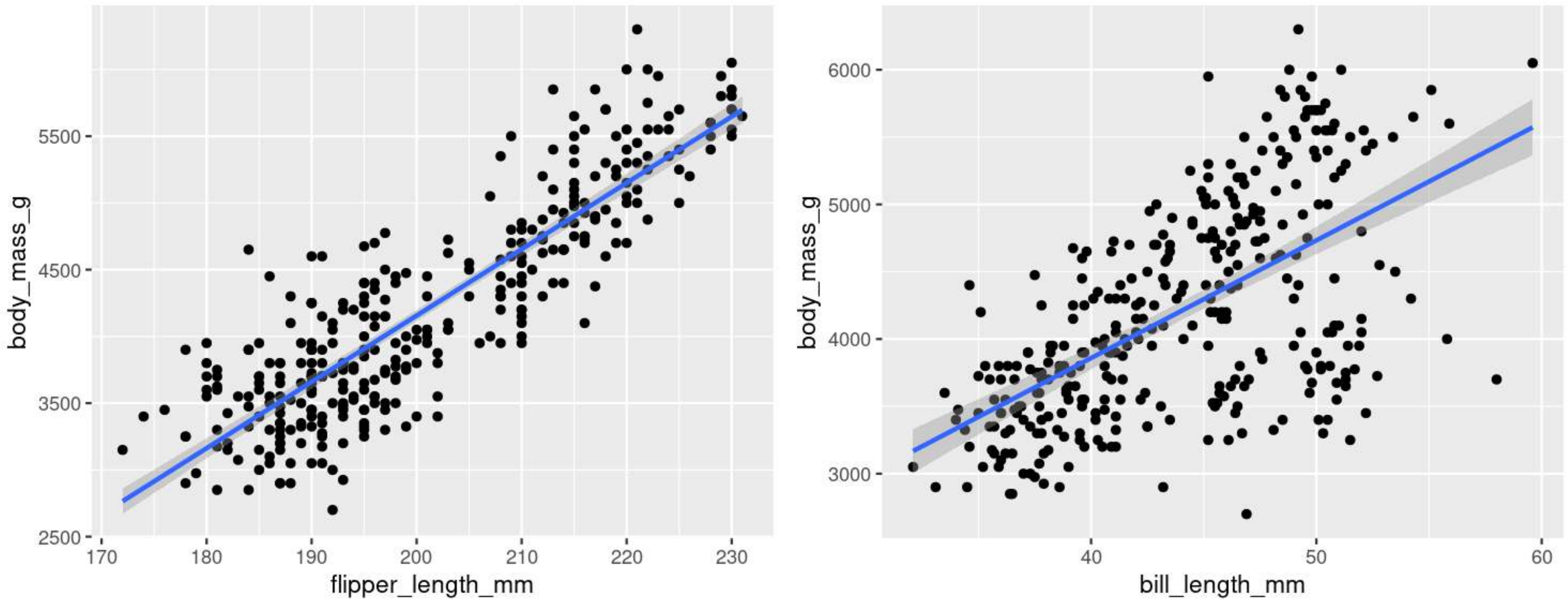
- If we only have one **x** which is continuous, this is a **simple linear regression**
- If both **x** are continuous, this is a **multiple linear regression**
- If both **x** are categorical, this is an **ANOVA**
- If **x1** is continuous and **x2** is categorical, this is an **ANCOVA**

R will figure it out for you

Regressions

Real example

- Is penguin body mass a function of skeletal size?
- Can it be predicted by flipper length and bill length?



Regressions

Real example

```r
lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Coefficients:
##       (Intercept)  flipper_length_mm     bill_length_mm
##         -5736.897             48.145              6.047
```

As we have **two continuous** predictors, this is technically a *multiple* regression

Regressions

Real example

```
lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Coefficients:
##       (Intercept)  flipper_length_mm     bill_length_mm
##         -5736.897             48.145              6.047
```

Hmm, not a lot of detail...
Only **`Intercept`** and Slopes
(**`flipper_length_mm`** and **`bill_length_mm`**)

As we have **two** predictors, this is technically a *multiple* regression

Regressions

Assign model to `m` (or any other name you want to give it)

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

`m` is a model object

```
class(m)
```

```
## [1] "lm"
```

This contains all the information about the model

Regressions

Use **`summary()`** to show summary table:

```r
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
```

```
                                                          ***
                                                          ***
## bill_length_mm        6.047       5.180   1.168     0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Your turn!

Create a model with your response variable by two of your *continuous* predictors.

Look at the output of **`summary()`**

Regressions

Use **`summary()`** to show summary table:

```
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
                                                              ***
                                                              ***
## bill_length_mm        6.047       5.180    1.168     0.244
## ---
#                                                  '.' 0.1 ' ' 1
#
#                                                  freedom
#
#                                                     0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Your turn!
Create a model with your response variable by two of your *continuous* predictors.

Look at the output of **`summary()`**

Wait!
Shouldn't interpret until we know the model is solid

Model Diagnostics

Model Assumptions

- Normality (of residuals)
- Constant Variance (no heteroscedasticity)

Other cautions

- Influential observations (Cook's D)
- Multiple collinearity (with more than one **x** or explanatory variables)

There are other assumptions (independence, etc.) but they reflect experimental design, not patterns in the data

Model Diagnostics

First let's get our relevant variables into a diagnostic data frame:

- `residuals` (regular and standardized)
- `fitted` **values**
- `cooks` **distance**
- `obs` **number**

```
d <- data.frame(residuals = residuals(m),
                std_residuals = rstudent(m),
                fitted = fitted(m),
                cooks = cooks.distance(m))

d <- mutate(d, obs = 1:n())
```

Remember! `n()` only works *inside* `summarize()`/`mutate()`

Model Diagnostics

First let's get our relevant variables into a diagnostic data frame:

- **`residuals`** (regular and standardized)
- **`fitted` values**
- **`cooks` distance**
- **`obs` number**

```
d <- data.frame(residuals = residuals(m),
                std_residuals = rstudent(m),
                fitted = fitted(m),
                cooks = cooks.distance(m))

d <- mutate(d, obs = 1:n())
```

```
head(d)
```

```
##     residuals std_residuals   fitted        cooks obs
## 1  536.220898   1.368529806 3213.779 5.539153e-03   1
## 2  343.077607   0.873050231 3456.922 1.609402e-03   2
## 3 -645.064115  -1.644516798 3895.064 3.797384e-03   3
## 5 -327.003441  -0.833002736 3777.003 1.992863e-03   4
## 6    1.707668   0.004338503 3648.292 3.338060e-08   5
## 7  412.430396   1.051400111 3212.570 3.272886e-03   6
```

Remember! **`n()`** only works *inside* **`summarize()`/`mutate()`**

Side Note: `tidyverse` functions

- From **`dplyr`** package (part of **`tidyverse`**)

```
d <- mutate(d, obs = 1:n())
```

`mutate()`

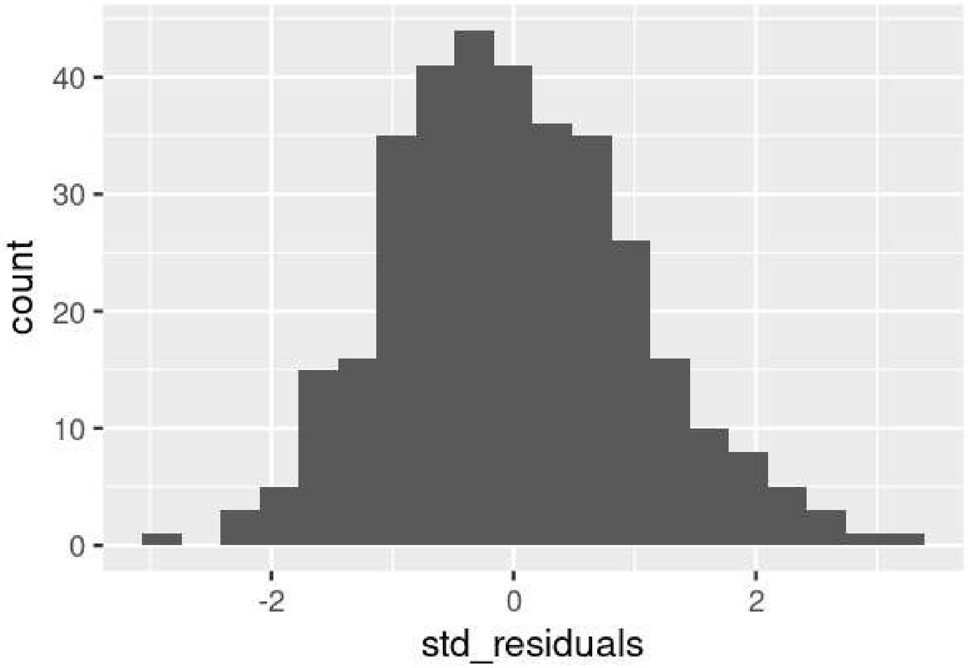
- **`tidyverse`** functions always start with the **data**, followed by **other arguments**
- **`mutate()`** adds **new columns** to your data
- Also note: **`1:5`** is the same as **`c(1,2,3,4,5)`**

Remember! **`n()`** only works *inside* **`summarize()`**/**`mutate()`**

Normality

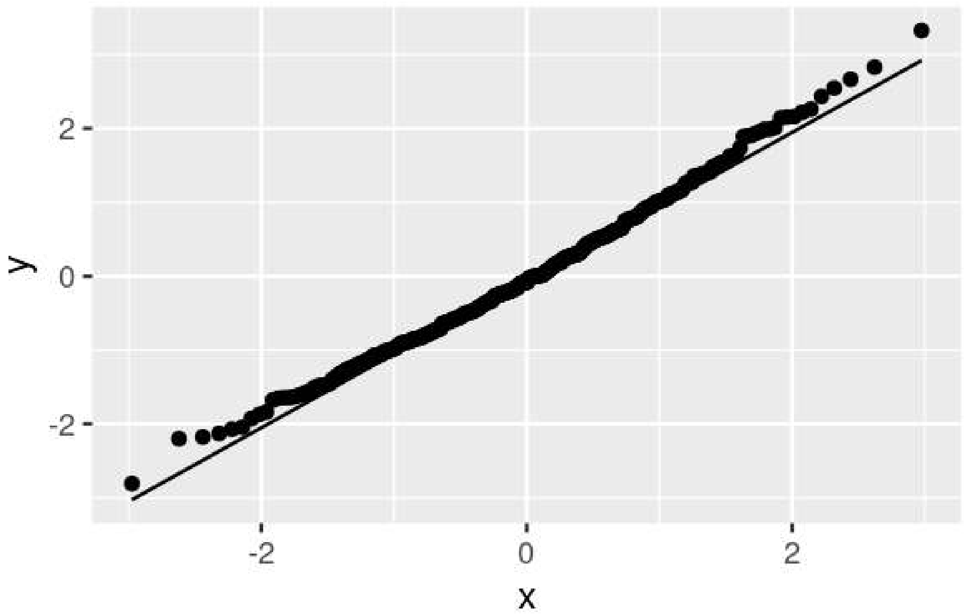
Histogram of residuals

```
ggplot(data = d, aes(x = std_residuals)) +
  geom_histogram(bins = 20)
```



QQ Normality plot of residuals

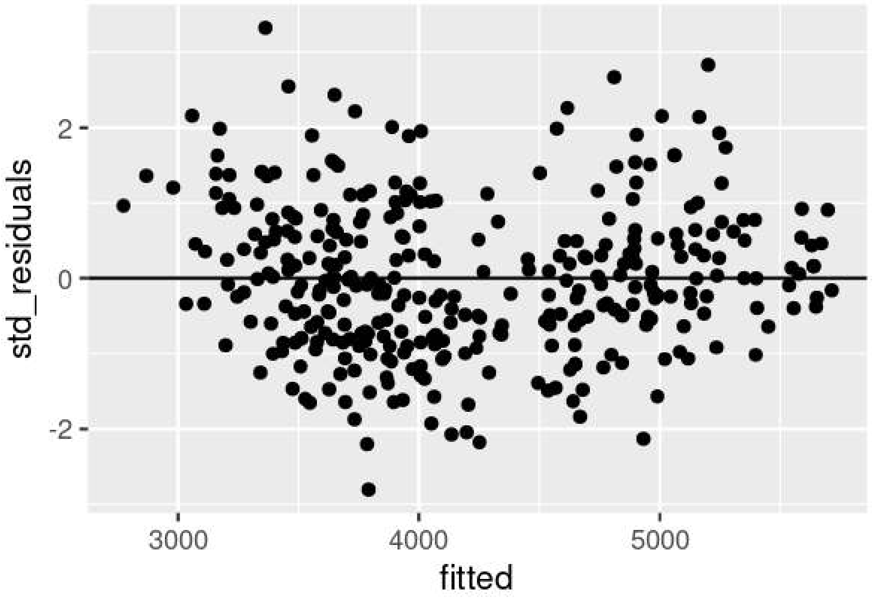
```
ggplot(data = d, aes(sample = std_residuals)) +
  stat_qq() +
  stat_qq_line()
```



Variance and Influence

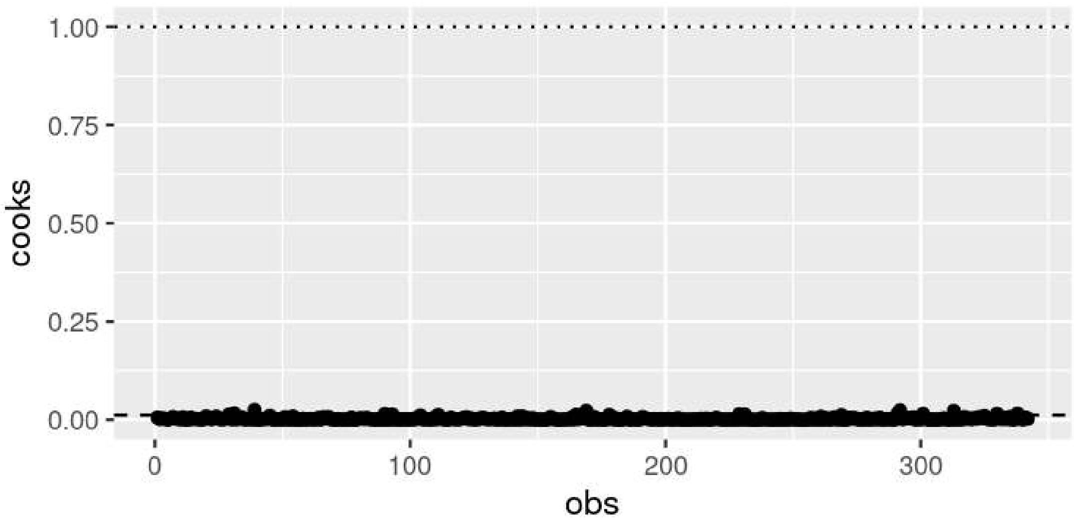
Check heteroscedasticity

```
ggplot(d, aes(x = fitted, y = std_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Cook's D

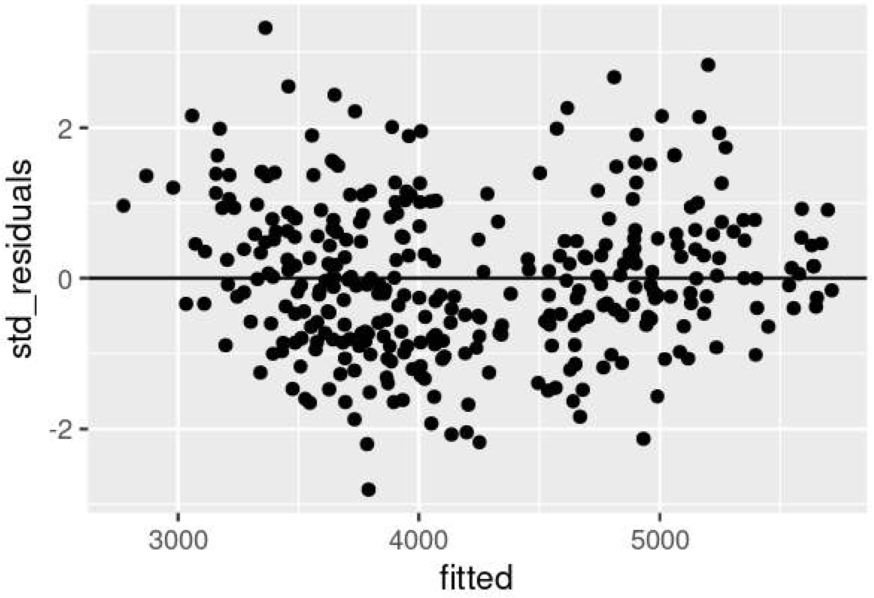
```
ggplot(d, aes(x = obs, y = cooks)) +
  geom_point() +
  geom_hline(yintercept = 1, linetype = "dotted") +
  geom_hline(yintercept = 4/nrow(penguins),
             linetype = "dashed")
```



Variance and Influence

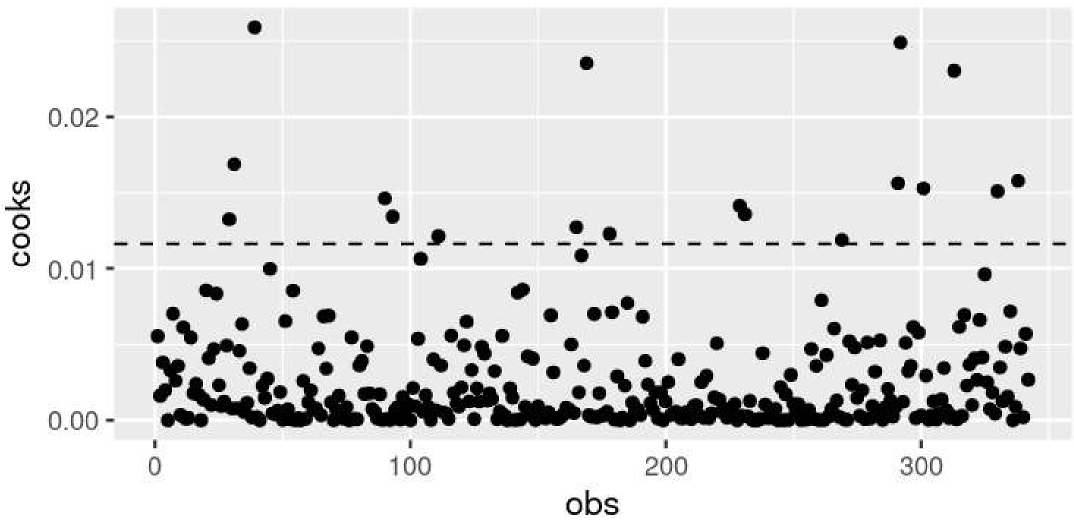
Check heteroscedasticity

```
ggplot(d, aes(x = fitted, y = std_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Cook's D

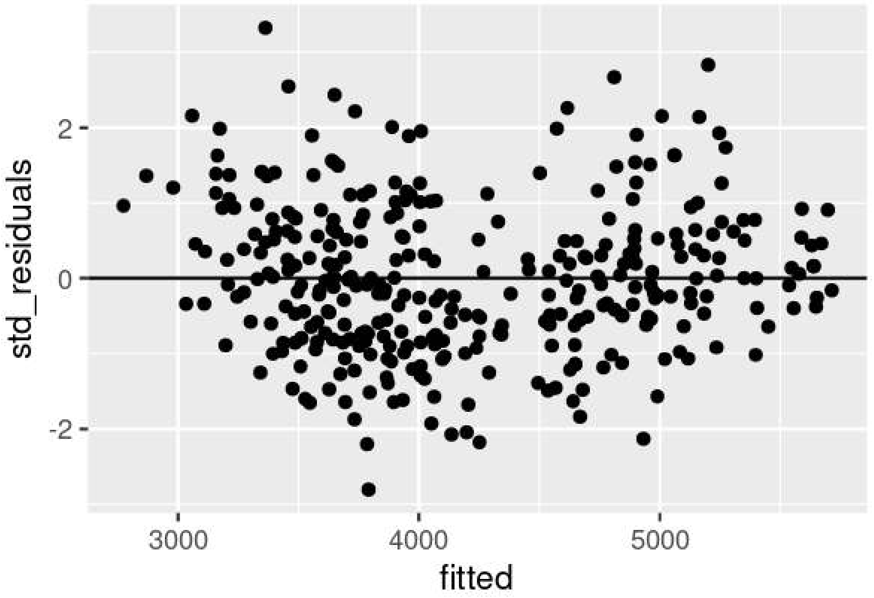
```
ggplot(d, aes(x = obs, y = cooks)) +
  geom_point() +
  geom_hline(yintercept = 4/nrow(penguins),
             linetype = "dashed")
```



Variance and Influence

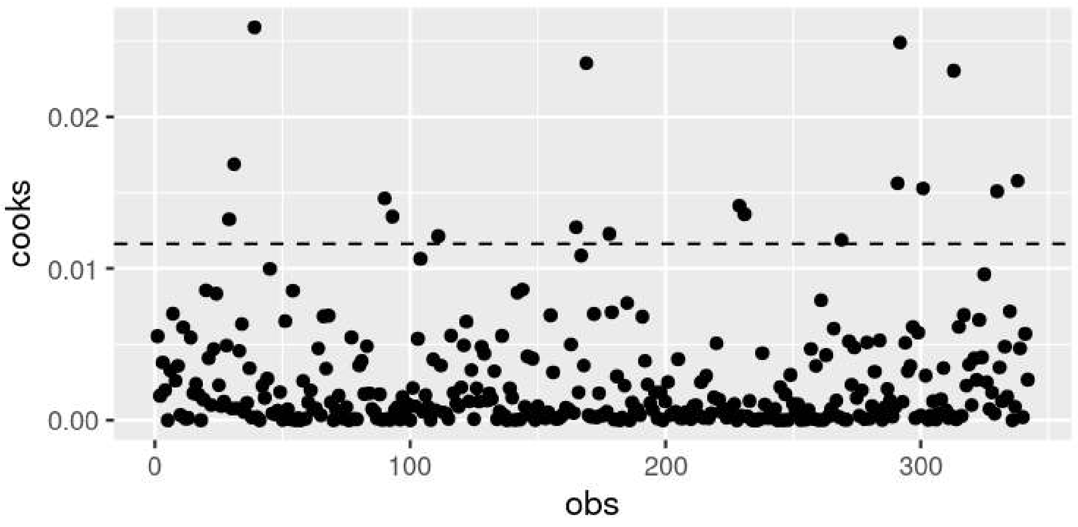
Check heteroscedasticity

```
ggplot(d, aes(x = fitted, y = std_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Cook's D

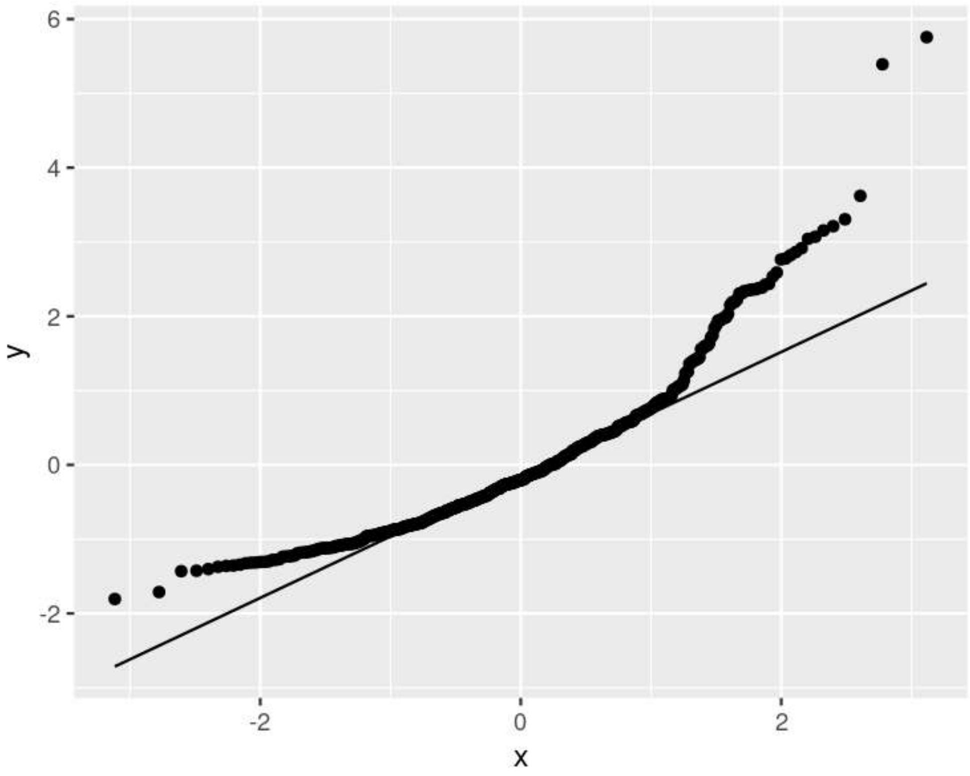
```
ggplot(d, aes(x = obs, y = cooks)) +
  geom_point() +
  geom_hline(yintercept = 4/nrow(penguins),
             linetype = "dashed")
```

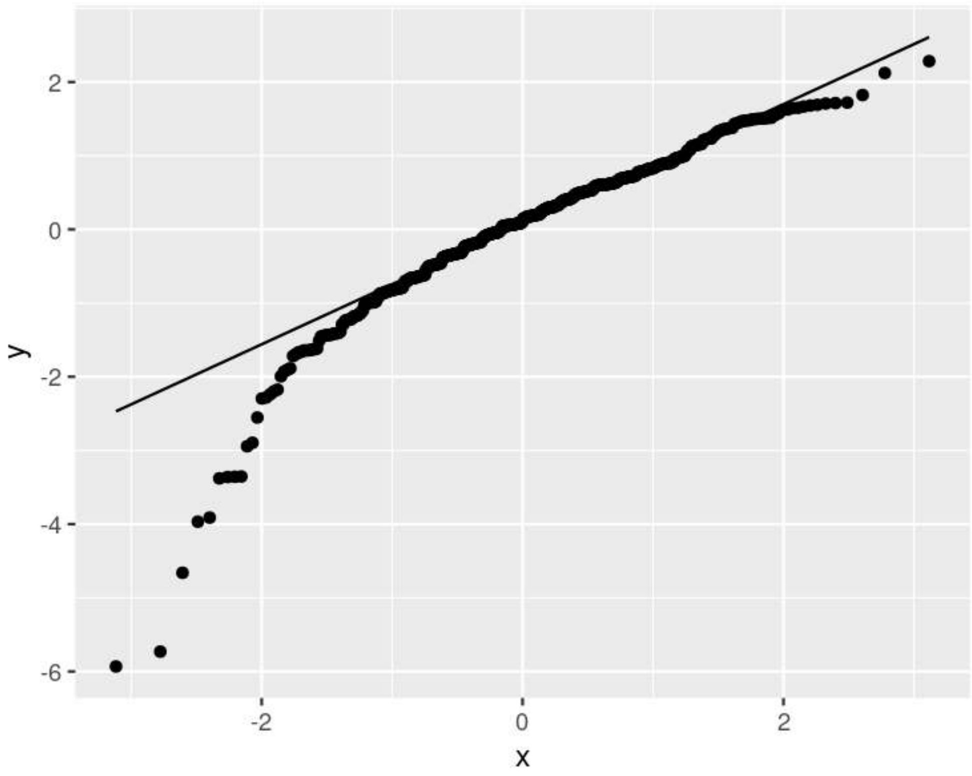


Pretty good!

What is a 'Good' Normality Plot?

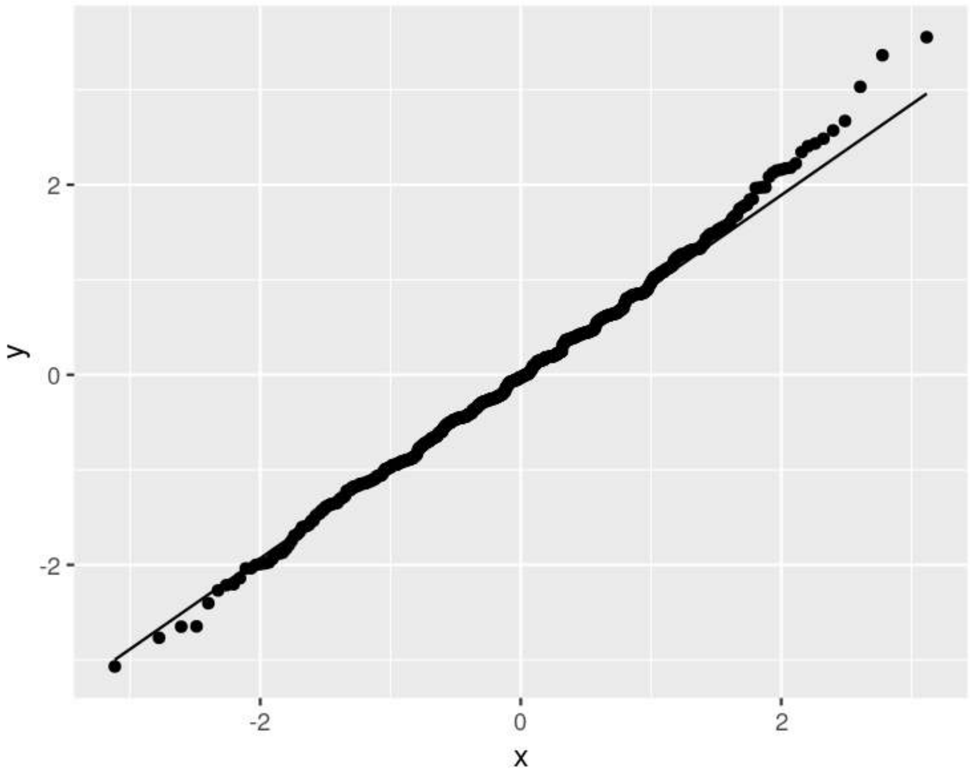
Problematic

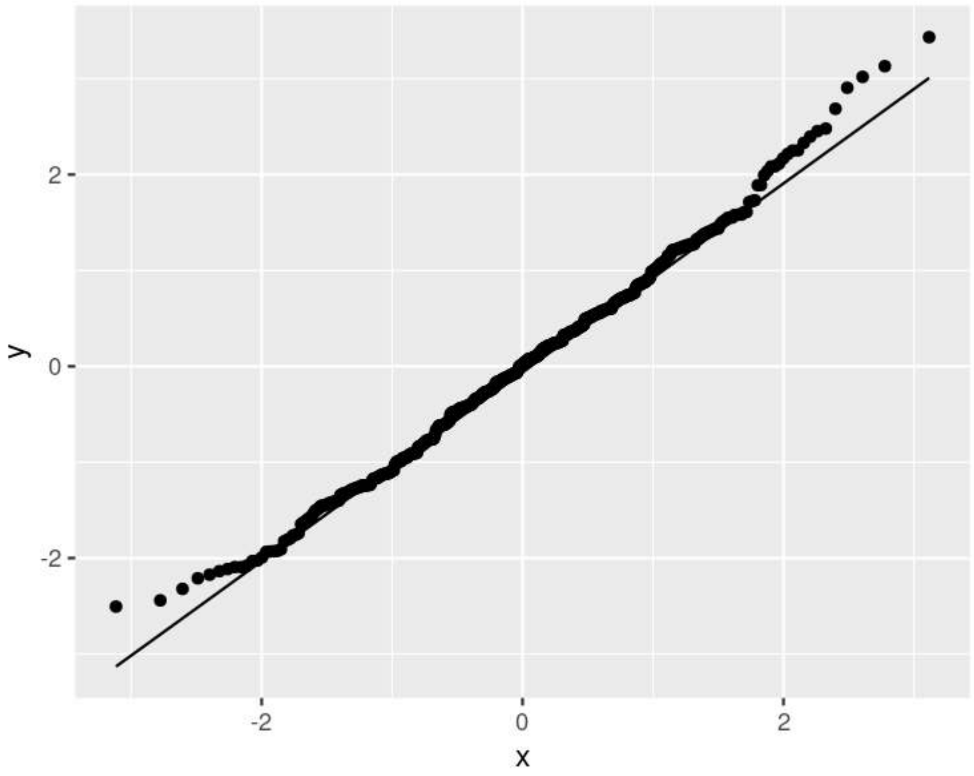




What is a 'Good' Normality Plot?

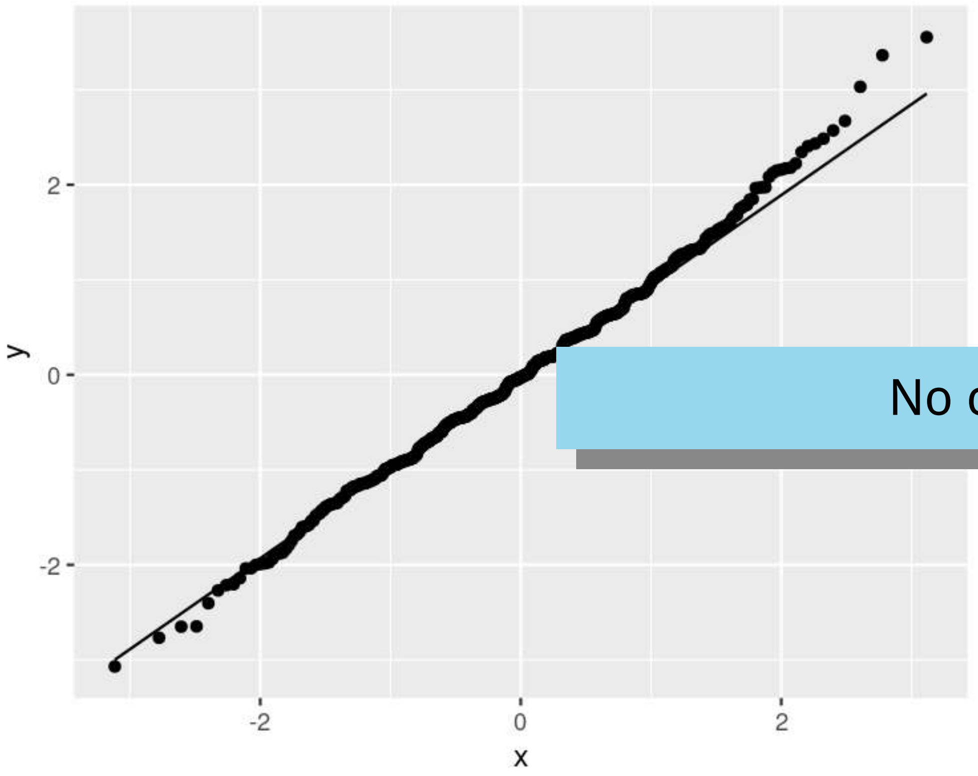
Good

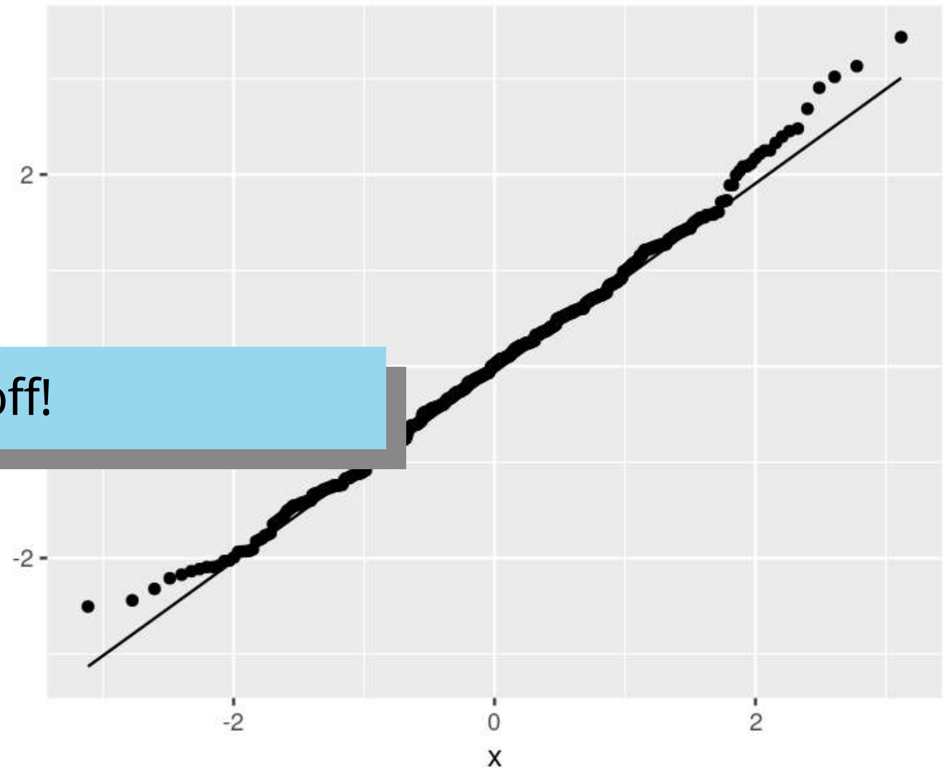




What is a 'Good' Normality Plot?

Good

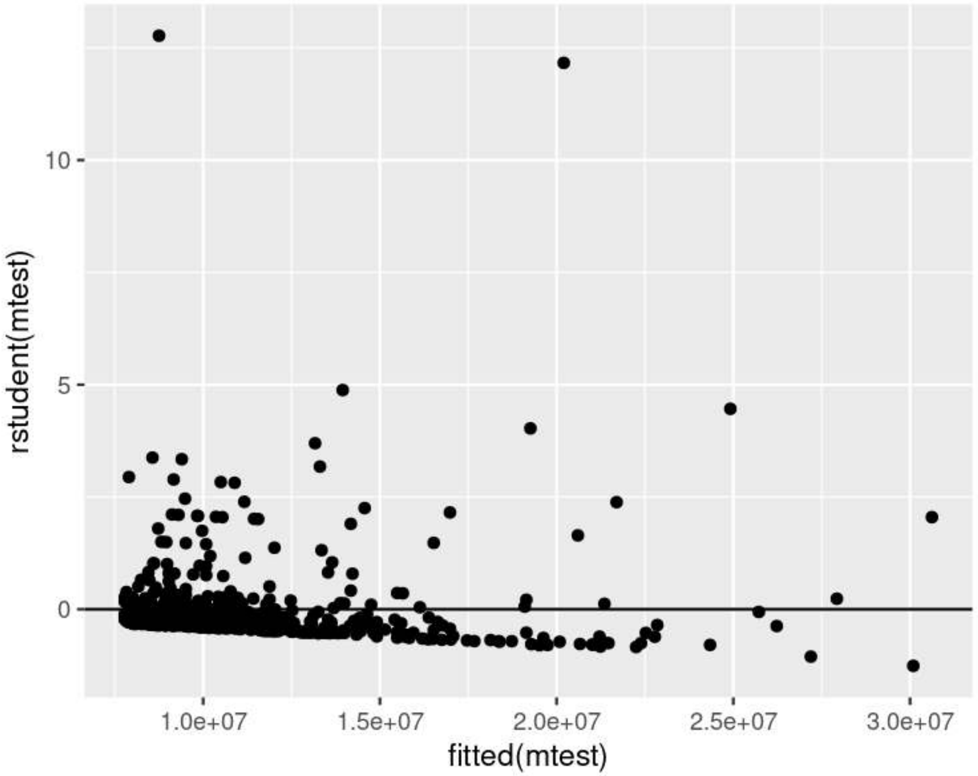


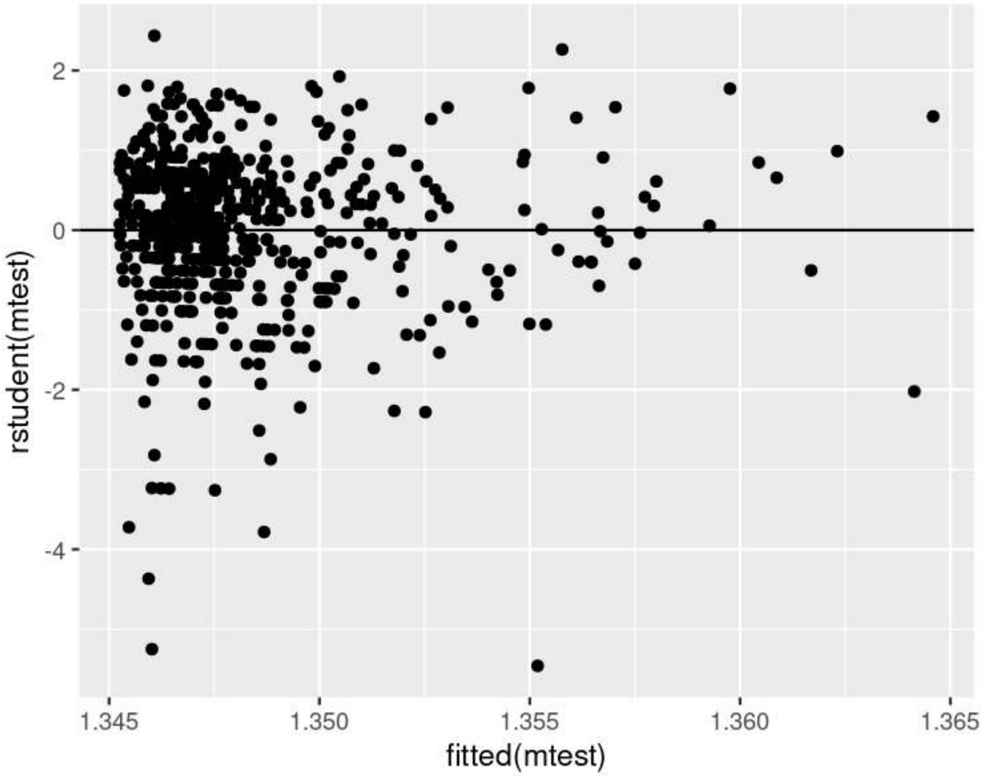


No clear cutoff!

What is a 'Good' Heteroscedasticity Plot?

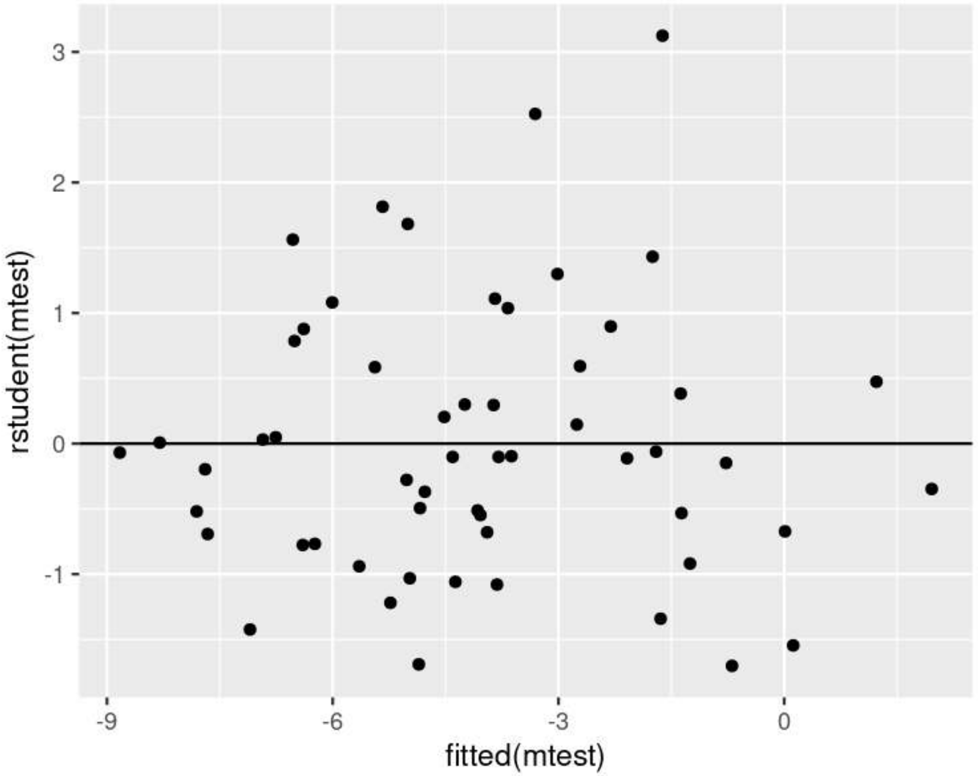
Problematic

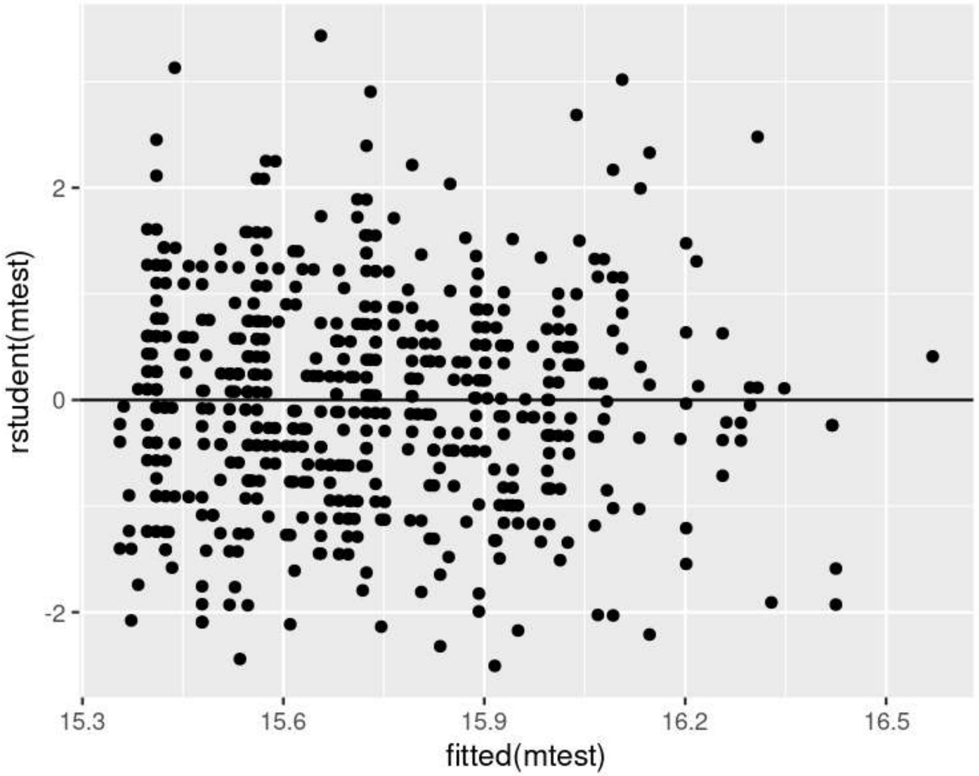




What is a 'Good' Heteroscedasticity Plot?

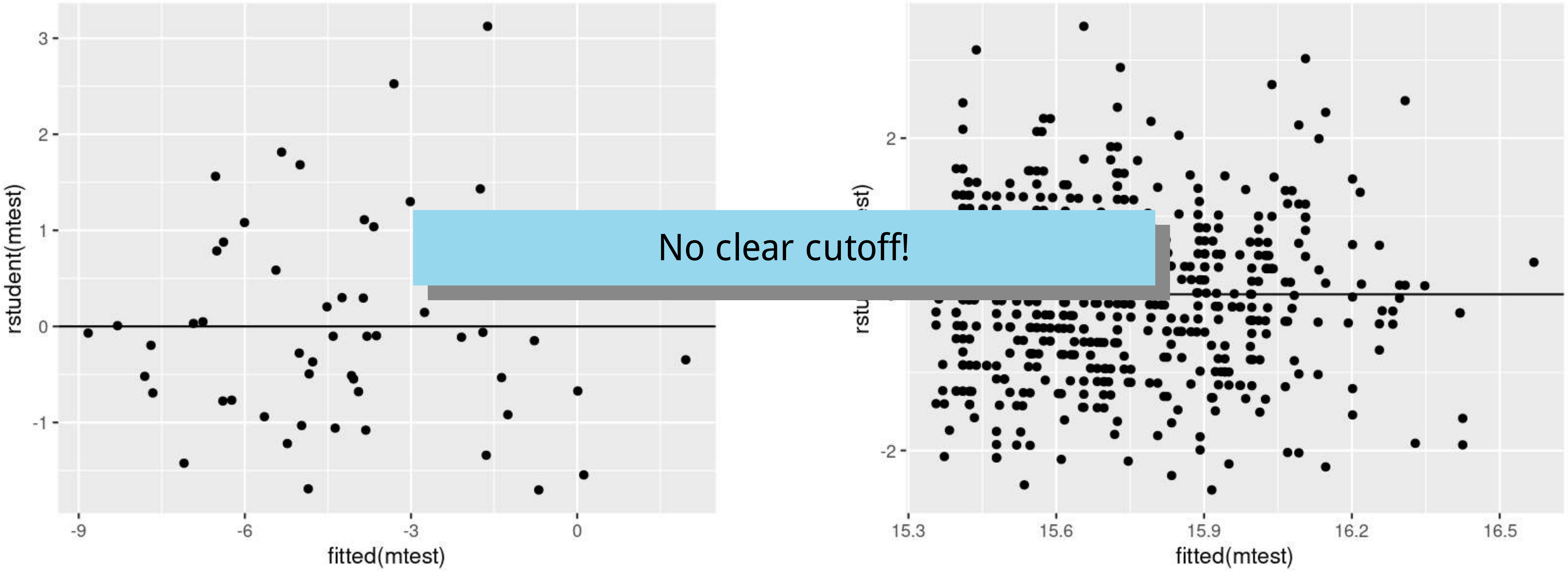
Good





What is a 'Good' Heteroscedasticity Plot?

Good



Multicollinearity (collinearity)

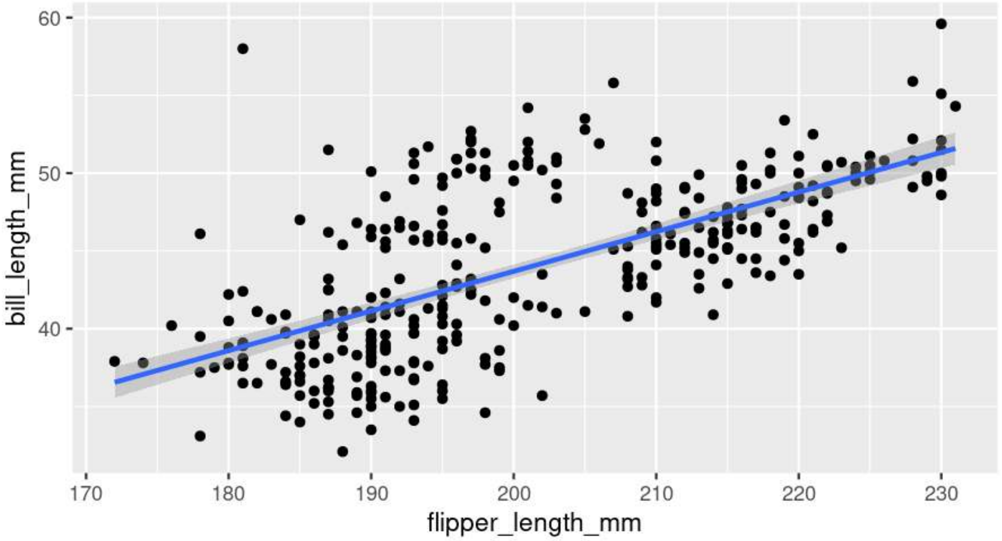
- Only relevant with **more than one explanatory variable**
- If explanatory variables too correlated, can interfere with model interpretation

Multicollinearity (collinearity)

- Only relevant with **more than one explanatory variable**
- If explanatory variables too correlated, can interfere with model interpretation

Look at our two explanatory variables

```
ggplot(data = penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point() +
  stat_smooth(method = "lm")
```

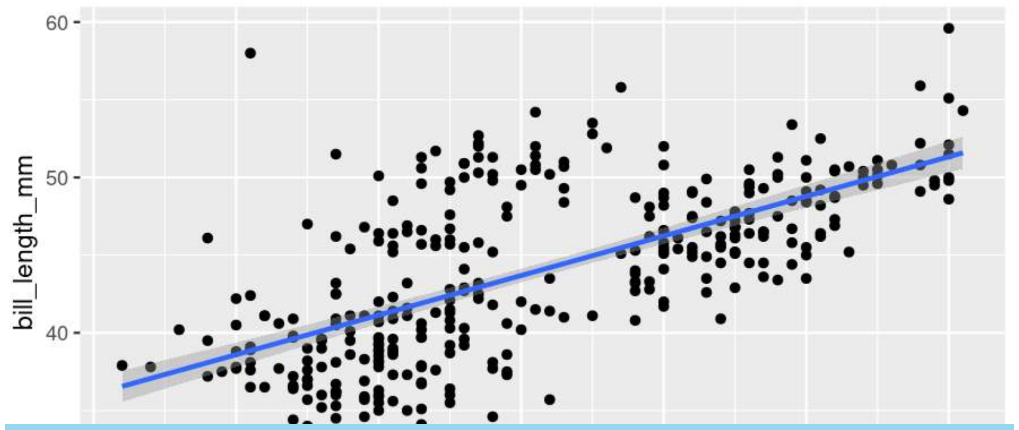


Multicollinearity (collinearity)

- Only relevant with **more than one explanatory variable**
- If explanatory variables too correlated, can interfere with model interpretation

Look at our two explanatory variables

```
ggplot(data = penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point() +
  stat_smooth(method = "lm")
```



Correlated, but not necessarily a problem

Multicollinearity (collinearity)

- Only relevant with **more than one explanatory variable**
- If explanatory variables too correlated, can interfere with model interpretation
- Correlations between variables *might* be problematic (but not necessarily)

Use `vif()` function from `car` package (vif = variance inflation factor*)

```
library(car)
vif(m)
```

```
## flipper_length_mm    bill_length_mm
##          1.756154          1.756154
```

Hmm, that's pretty good (looking for < 10)

* Can be interpreted as how much influence the variable has on the model

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -1090.5   -285.7    -32.1    244.2   1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Model

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -1090.5   -285.7   -32.1    244.2   1287.5
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm     48.145      2.011  23.939   <2e-16 ***
## bill_length_mm         6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting Regressions

```r
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```r
summary(m)
```

Effects

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,   Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Missing observations

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

R^2 and adjusted R^2

- Adjusted for the number of parameters

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Specific Details

- **`Estimate`**
  - Slope of the effect
- **`Std. Error`**
  - Variability in the estimates
- **`t value`**
  - Test statistic
  - Think of it as a holistic combination of estimate and variability
- **`Pr(>|t|)`**
  - **P-value**, significance of the results
  - Probability of getting **`t-value`** by chance

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Specific Details

Intercept

- Significant ($P < 2e^{-16}$*)
- Penguins with a flipper length of 0 mm are predicted to have a body mass of -5736.9g
    - Not useful!

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,	Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Specific Details

Effect of **Flipper Length**

- Significant (P < $2e^{-16}$*)
- For each 1 mm increase in flipper length, body mass is predicted to increase by 48.14g

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,	Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Specific Details

Effect of **Flipper Length**

- Significant (P < $2e^{-16}$*)
- For each 1 mm increase in flipper length, body mass is predicted to increase by 48.14g

Effect of **Bill Length**

- Non-significant (P = 0.244, i.e. P < 0.05)
- Therefore no effect (in this model) (and no interpretation of estimate)

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

Interpreting Regressions

```
m <- lm(bo
```

```
summary(m)
```

Therefore
There is a significant relationship between flipper length and body mass
But not between bill length and body mass (when including flipper length)

Specific Details

Effect of **Flipper Length**

- Significant (P < $2e^{-16}$*)
- For each 1 mm increase in flipper length, body mass is predicted to increase by 48.14g

Effect of **Bill Length**

- Non-significant (P = 0.244, i.e. P < 0.05)
- Therefore no effect (in this model) (and no interpretation of estimate)

```
_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Specific Details

Effect of **Flipper Length**

- Significant (P < $2e^{-16}$*)
- For each 1 mm increase in flipper length, body mass is predicted to increase by 48.14g

Effect of **Bill Length**

- Non-significant (P = 0.244, i.e. P < 0.05)
- Therefore no effect (and no interpretation of estimate)

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

$y = mx + b$

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,    Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

$$y = m_1x_1 + m_2x_2 + b$$

Specific Details

Effect of **Flipper Length**

- Significant (P < $2e^{-16}$*)
- For each 1 mm increase in flipper length, body mass is predicted to increase by 48.14g

Effect of **Bill Length**

- Non-significant (P = 0.244, i.e. P < 0.05)
- Therefore no effect (and no interpretation of estimate)

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

Interpreting Regressions

```
m <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
```

```
summary(m)
```

Specific Details

Effect of **Flipper Length**

- Significant (P < $2e^{-16}$*)
- For each 1 mm increase in flipper length, body mass is predicted to increase by 48.14g

Effect of **Bill Length**

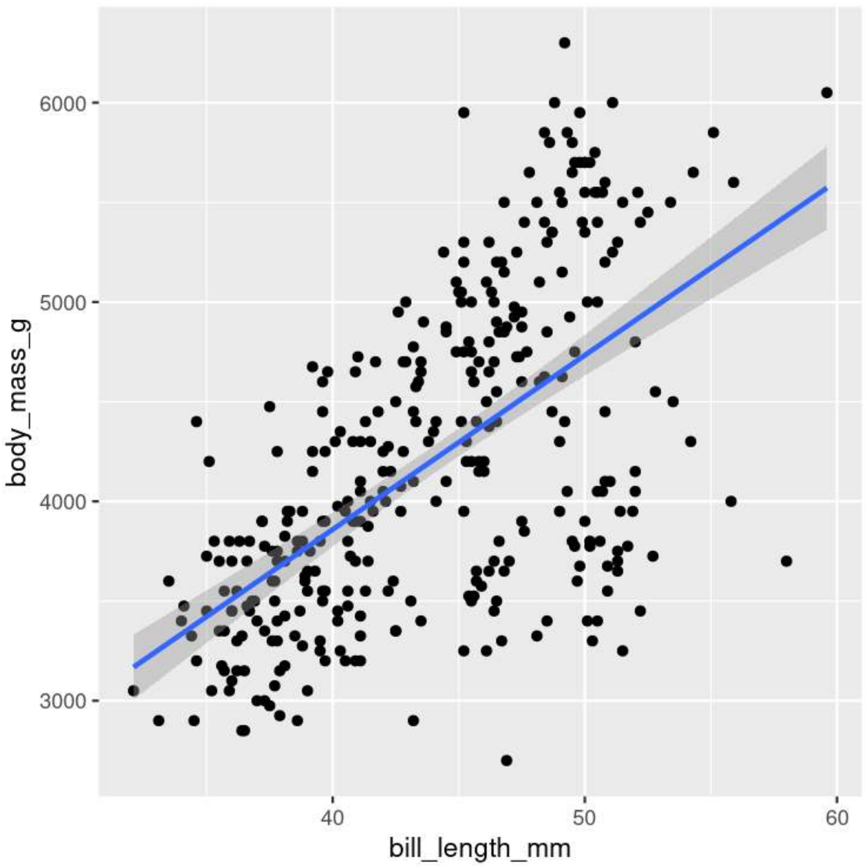
- Non-significant (P = 0.244, i.e. P < 0.05)
- Therefore no effect (and no interpretation of estimate)

* $2e^{-16}$ = 0.0000000000000002, R uses this as the smallest number

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.1 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.76,   Adjusted R-squared:  0.7585
## F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

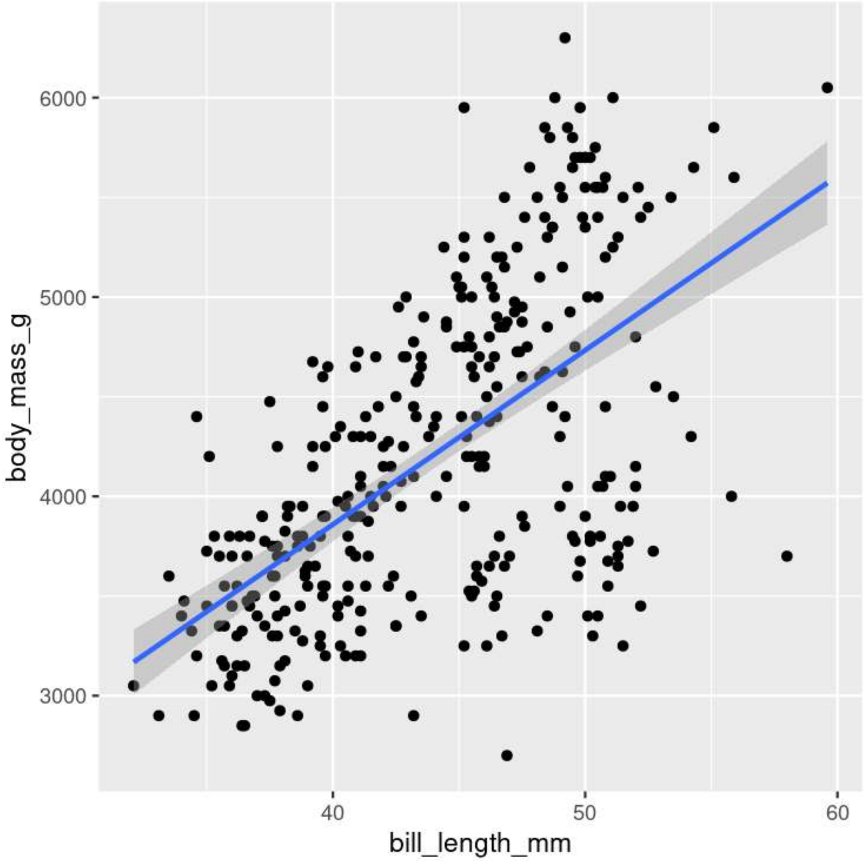
$$y = 48.14x_1 + 6.05x_2 + (-5736.9)$$

Why no effect of Bill Length?



Extra

Why no effect of Bill Length?




```r
m <- lm(body_mass_g ~ bill_length_mm, data = penguins)
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_length_mm, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1762.08  -446.98    32.59   462.31  1636.86
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     362.307    283.345   1.279    0.202
## bill_length_mm   87.415      6.402  13.654   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.4 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3542,    Adjusted R-squared:  0.3523
## F-statistic: 186.4 on 1 and 340 DF,  p-value: < 2.2e-16
```

Extra

Why no effect of Bill Length?

- Hypothesis of *causation* but really just correlation
- Flipper length is the 'better' predictor of body mass
- When flipper length in the model, no extra variation explained by bill length
- When flipper length *not* in the model, some variation left to be explained

```
m <- lm(body_mass_g ~ bill_length_mm, data = penguins)
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_length_mm, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1762.08  -446.98    32.59   462.31  1636.86
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     362.307    283.345   1.279    0.202
## bill_length_mm   87.415      6.402  13.654   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.4 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3542,    Adjusted R-squared:  0.3523
## F-statistic: 186.4 on 1 and 340 DF,  p-value: < 2.2e-16
```

Homework (Practice)*

Consider **bill depth** your response and **bill length** your predictor

1. Plot the relationship
2. Create a linear regression model
3. Check your model diagnostics
    - Normality
    - Heteroscedasticity
    - Influential variables (i.e. Cook's distance)
4. Interpret the results of your model

* Not to be handed in, answers posted in these slides next week

ANOVAs

Linear Models

Running models in R

```
lm(y ~ x1 + x2, data = data)
```

- **y** is the **response** variable (**dependent**)
- **x** are the **explanatory** variables (**independent**, **predictor**)

Here we're assuming a **continuous y**

Linear Models

Running models in R

```
lm(y ~ x1 + x2, data = data)
```

- **y** is the **response** variable (**dependent**)
- **x** are the **explanatory** variables (**independent**, **predictor**)

Here we're assuming a **continuous y**

Different types of models

- If we only have one **x** which is continuous, this is a **simple linear regression**
- If both **x** are continuous, this is a **multiple linear regression**
- If both **x** are categorical, this is an **ANOVA**
- If **x1** is continuous and **x2** is categorical, this is an **ANCOVA**

Linear Models

Running models in R

```
lm(y ~ x1 + x2, data = data)
```

- **y** is the **response** variable (**dependent**)
- **x** are the **explanatory** variables (**independent**, **predictor**)

Here we're assuming a **continuous y**

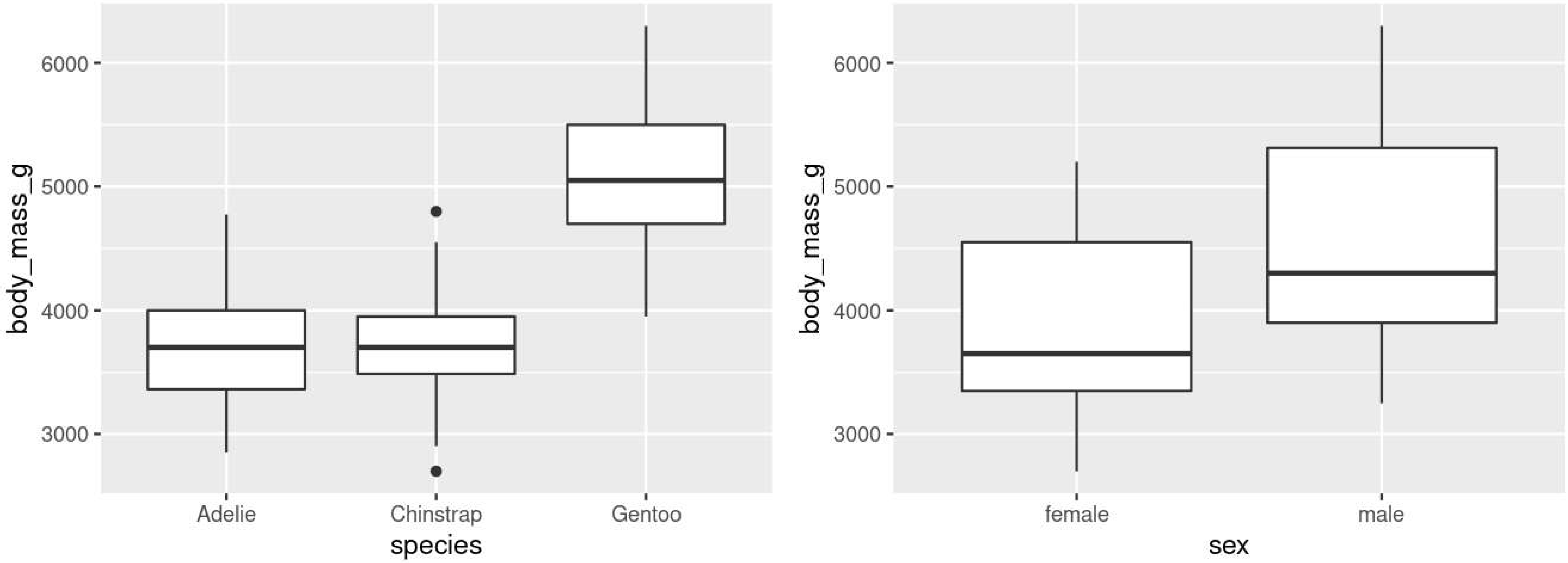
Different types of models

- If we only have one **x** which is continuous, this is a **simple linear regression**
- If both **x** are continuous, this is a **multiple linear regression**
- If both **x** are categorical, this is an **ANOVA**
- If **x1** is continuous and **x2** is categorical, this is an **ANCOVA**

R will figure it out for you

Real example

- Are male penguins larger than female penguins?
- Are different species different sizes?
- Can body mass be predicted by species and sex?



ANOVAs

Real example

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

As we have two **categorical** predictors, this is an ANOVA

Your turn!
Create a model with your response variable by your *one categorical* predictor.

Look at the output of `summary()` *and* `anova()`

ANOVAs

Real example

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

As we have two **categorical** predictors, this is an ANOVA

Your turn!
Create a model with your response variable by your *one categorical* predictor.

Look at the output of **`summary()`** *and* **`anova()`**

Wait!
Shouldn't interpret until we know the model is solid

Model Diagnostics

Same as before! Let's get our relevant variables into a diagnostic data frame:

- **`residuals`** (regular and standardized)
- **`fitted` values**
- **`cooks` distance**
- **`obs` number**

```
d <- data.frame(residuals = residuals(m),
                std_residuals = rstudent(m),
                fitted = fitted(m),
                cooks = cooks.distance(m))

d <- mutate(d, obs = 1:n())
```

Remember! **`n()`** only works *inside* **`summarize()`**/**`mutate()`**

Model Diagnostics

Same as before! Let's get our relevant variables into a diagnostic data frame:

- **`residuals`** (regular and standardized)
- **`fitted` values**
- **`cooks` distance**
- **`obs` number**

```
d <- data.frame(residuals = residuals(m),
                std_residuals = rstudent(m),
                fitted = fitted(m),
                cooks = cooks.distance(m))

d <- mutate(d, obs = 1:n())
```

```
head(d)
```

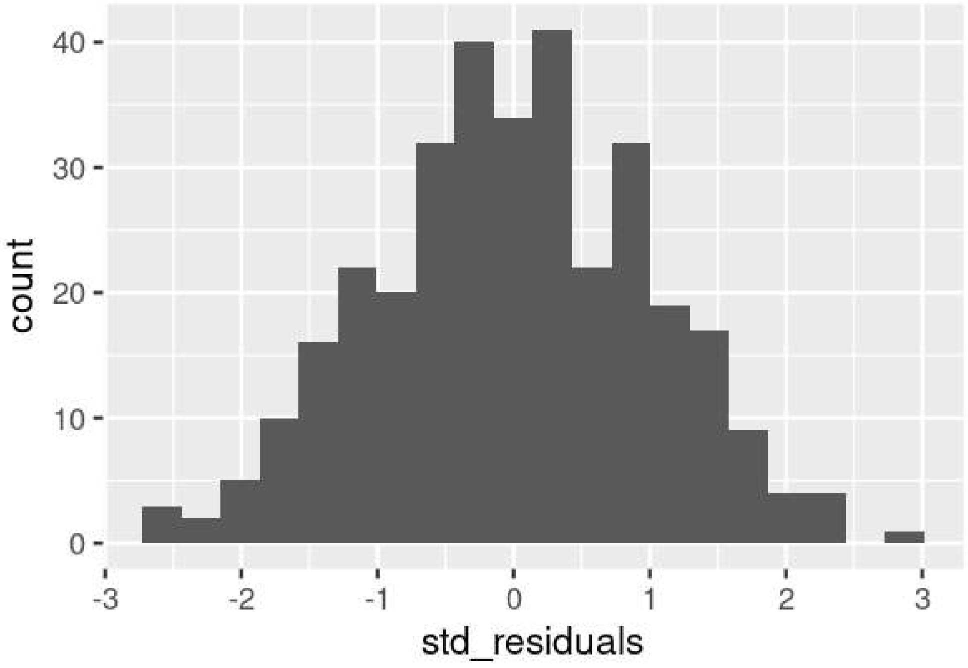
```
##    residuals std_residuals   fitted        cooks obs
## 1 -289.94196    -0.9201062 4039.942 0.0021071103   1
## 2  427.61319     1.3590567 3372.387 0.0045831843   2
## 3 -122.38681    -0.3879729 3372.387 0.0003754346   3
## 5   77.61319     0.2460043 3372.387 0.0001509858   4
## 6 -389.94196    -1.2387413 4039.942 0.0038112292   5
## 7  252.61319     0.8013974 3372.387 0.0015994740   6
```

Remember! **`n()`** only works *inside* **`summarize()`/`mutate()`**

Normality

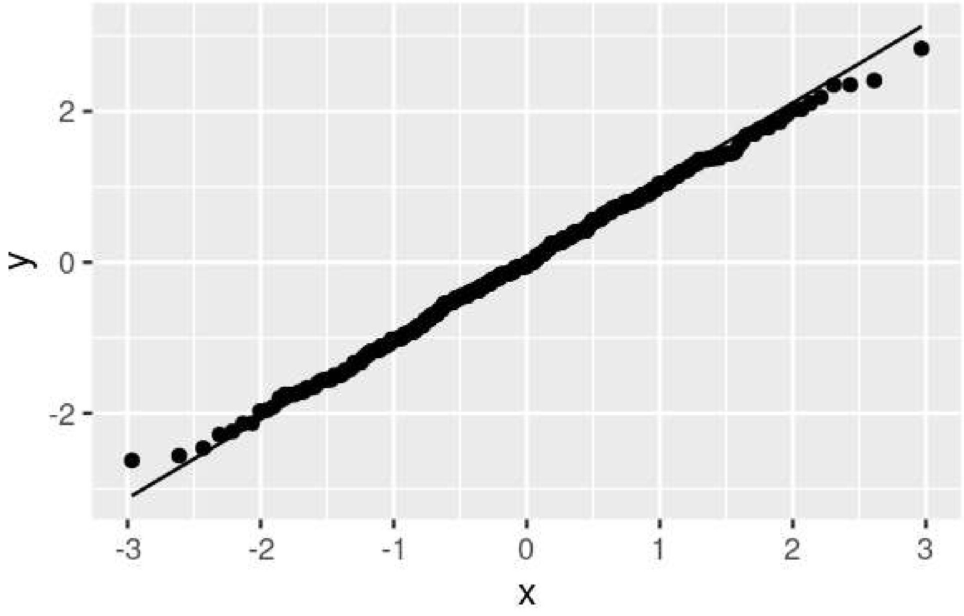
Histogram of residuals

```r
ggplot(data = d, aes(x = std_residuals)) +
  geom_histogram(bins = 20)
```



QQ Normality plot of residuals

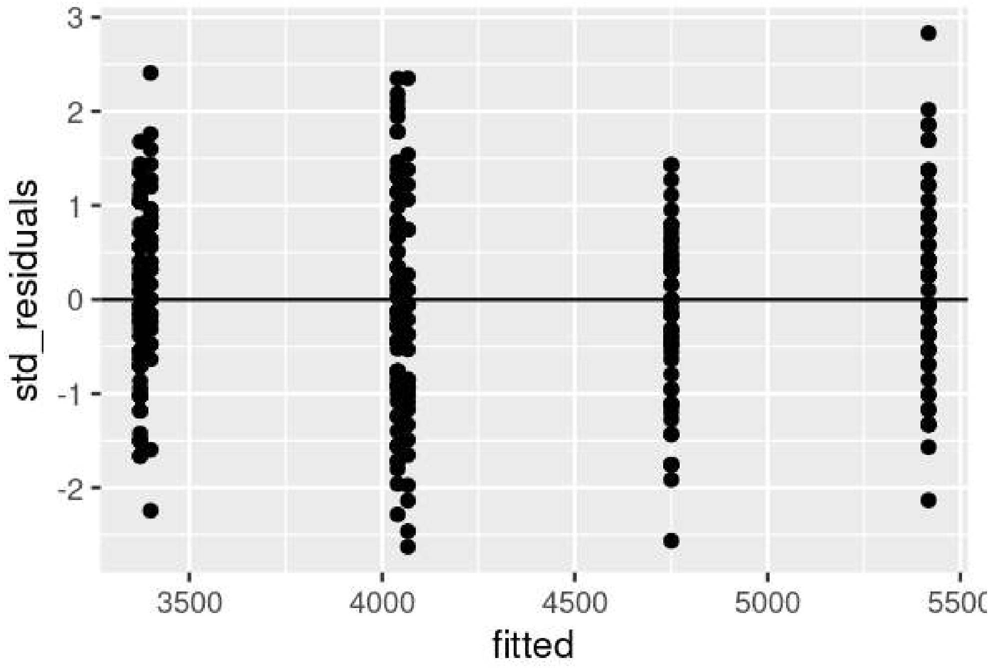
```r
ggplot(data = d, aes(sample = std_residuals)) +
  stat_qq() +
  stat_qq_line()
```



Variance and Influence

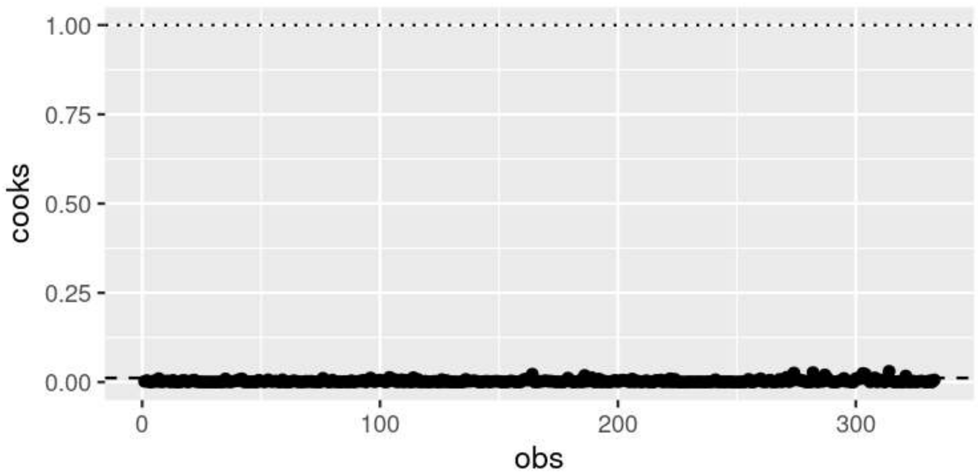
Check heteroscedasticity

```
ggplot(d, aes(x = fitted, y = std_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Cook's D

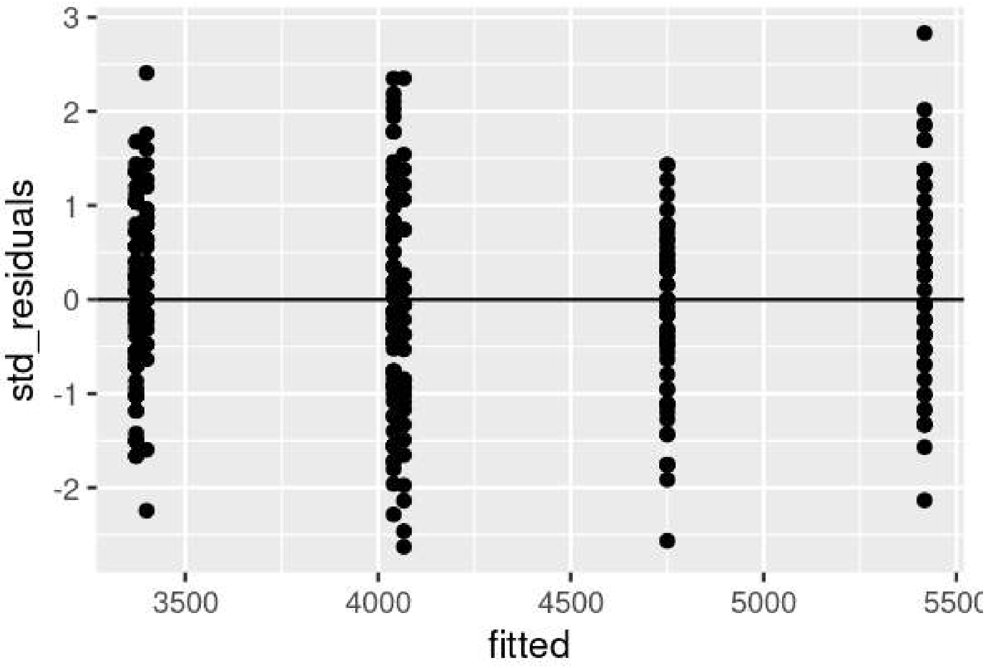
```
ggplot(d, aes(x = obs, y = cooks)) +
  geom_point() +
  geom_hline(yintercept = 1, linetype = "dotted") +
  geom_hline(yintercept = 4/nrow(penguins),
             linetype = "dashed")
```



Variance and Influence

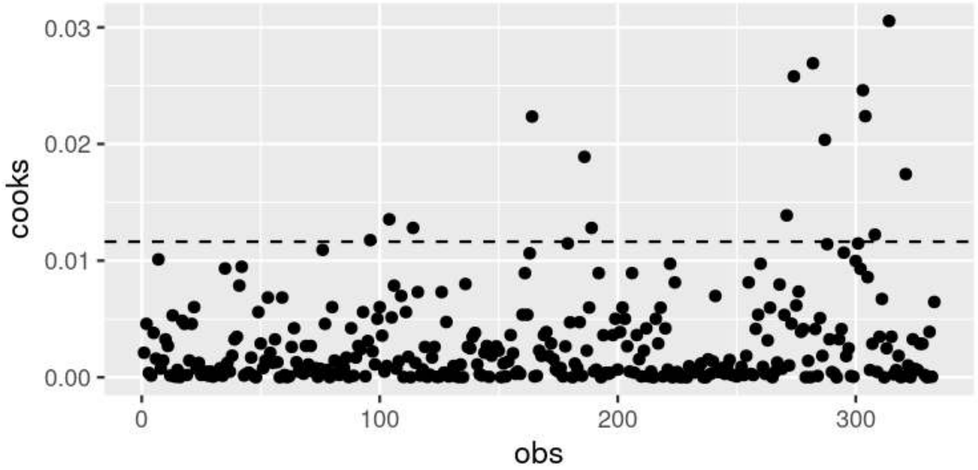
Check heteroscedasticity

```
ggplot(d, aes(x = fitted, y = std_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Cook's D

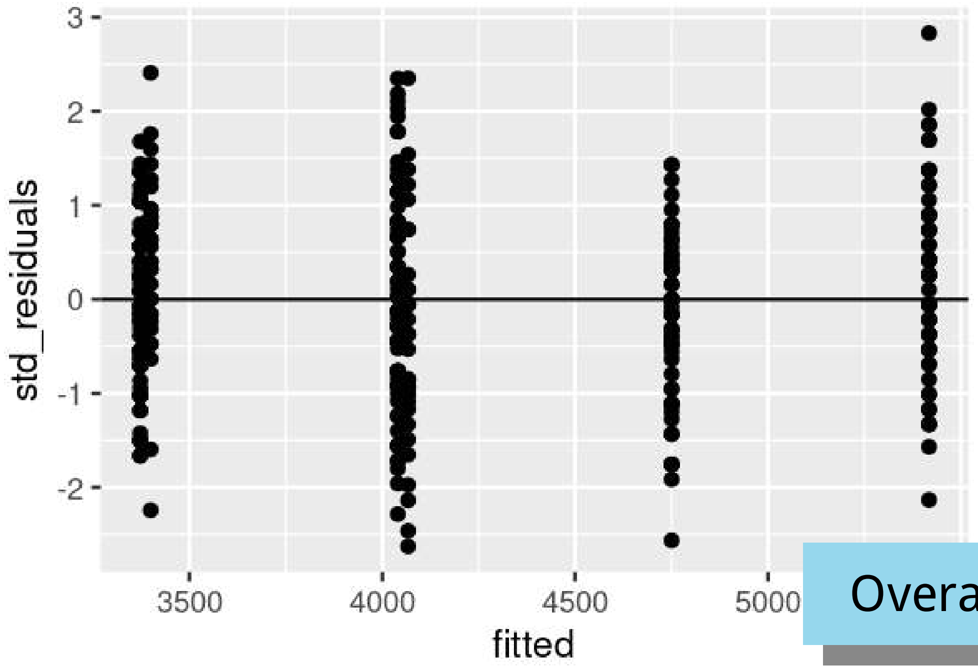
```
ggplot(d, aes(x = obs, y = cooks)) +
  geom_point() +
  geom_hline(yintercept = 4/nrow(penguins),
             linetype = "dashed")
```



Variance and Influence

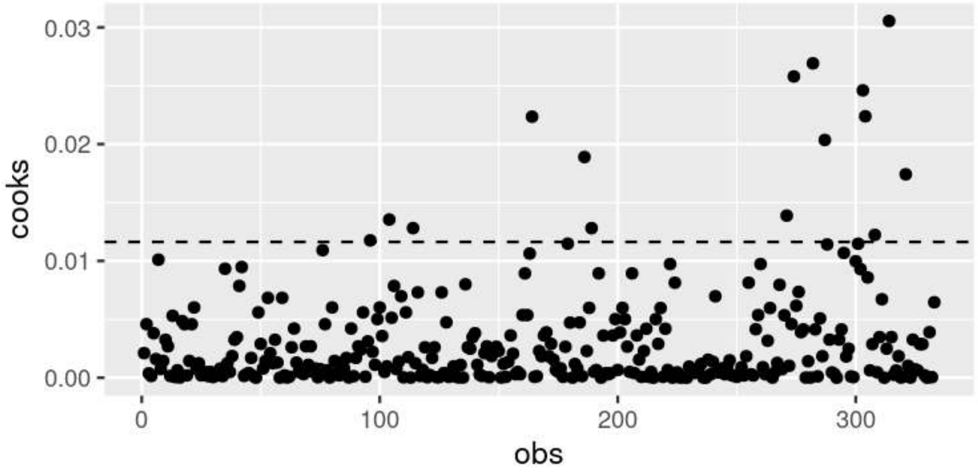
Check heteroscedasticity

```
ggplot(d, aes(x = fitted, y = std_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Cook's D

```
ggplot(d, aes(x = obs, y = cooks)) +
  geom_point() +
  geom_hline(yintercept = 4/nrow(penguins),
             linetype = "dashed")
```



Overall not too bad

Multicollinearity (collinearity)

`vif()` function from `car` package

```
library(car)
vif(m)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## species 1.000146  2        1.000036
## sex     1.000146  1        1.000073
```

Here we consider the **`GVIF^(1/2*Df))`** value*

Looks good!

* See **`?vif`** and the reference therein: Fox, J. and Monette, G. (1992) Generalized collinearity diagnostics. JASA, 87, 178–183.

Interpreting ANOVA Summaries

```r
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```r
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries

```r
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```r
summary(m)
```

Model

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
summary(m)
```

Effects

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
summary(m)
```

Missing observations

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
summary(m)
```

R^2 and adjusted R^2

- Adjusted for the number of parameters

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries

```r
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```r
summary(m)
```

Specific Details

- **`Estimate`**
  - Treatment contrasts
  - Average *differences* among categories compared to the base category
- **`Std. Error`**
  - Variability in the estimates
- **`t value`**
  - Test statistic
- **`Pr(>|t|)`**
  - **P-value**, significance of the *differences*
  - Probability of getting **`t-value`** by chance

```
## 
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -816.87 -217.80  -16.87  227.61  882.20 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563    
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,	Adjusted R-squared:  0.8454 
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
summary(m)
```

Specific Details

- **`Estimate`**
  - Treatment contrasts
  - Average *diffe*
    compared to t
- **`Std. Error`**
  - Variability in the estimates
- **`t value`**
  - Test statistic
- **`Pr(>|t|)`**
  - **P-value**, significance of the *differences*
  - Probability of getting **`t-value`** by chance

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap     26.92      46.48   0.579    0.563
## speciesGentoo      1377.86      39.10  35.236   <2e-16 ***
## sexmale             667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Easier to interpret estimates if we consider a simpler model

Interpreting ANOVA Summaries

```
m <- lm(body_mass_g ~ species, data = penguins)
```

```
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1126.02  -333.09   -33.09   316.91  1223.98
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3700.66      37.62   98.37   <2e-16 ***
## speciesChinstrap    32.43      67.51    0.48    0.631
## speciesGentoo     1375.35      56.15   24.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462.3 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6697,    Adjusted R-squared:  0.6677
## F-statistic: 343.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries


```
m <- lm(body_mass_g ~ species, data = penguins)
```

```
summary(m)
```

Effect of **Species**

- **`(Intercept)`** represents base category (i.e. Adelie penguins)
- Adelie have mean body mass of 3700.66 g
- On average, Chinstrap penguins are 32.43 g heavier than Adelie penguins
- On average, Gentoo penguins are 1375.35 g heavier than Adelie penguins

```
##
## Call:
## lm(formula = body_mass_g ~ species, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1126.02  -333.09   -33.09   316.91  1223.98
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3700.66      37.62   98.37   <2e-16 ***
## speciesChinstrap    32.43      67.51    0.48    0.631
## speciesGentoo     1375.35      56.15   24.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462.3 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6697,    Adjusted R-squared:  0.6677
## F-statistic: 343.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Summaries


```r
m <- lm(body_mass_g ~ species, data = penguins)
```

```r
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species, data = penguins)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1126.02  -333.09   -33.09    316.91   1223.98
```

Back to original model

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3700.66      37.62   98.37   <2e-16 ***
## speciesChinstrap    32.43      67.51    0.48    0.631
## speciesGentoo     1375.35      56.15   24.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462.3 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6697,    Adjusted R-squared:  0.6677
## F-statistic: 343.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Effect of **Species**

- **`(Intercept)`** represents base category (i.e. Adelie penguins)
- Adelie have mean body mass of 3700.
- On average, Chinstrap penguins are 32.43 g heavier than Adelie penguins
- On average, Gentoo penguins are 1375.35 g heavier than Adelie penguins

Interpreting ANOVA Summaries

```r
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```r
summary(m)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Effect of **Species** and **Sex**

- **`(Intercept)`** represents base category but is a combination of factors
- Much more complicated to interpret
- Comparisons are often not of interest anyway (unless you've set up contrasts, which are advanced stats but awesome!)

Interpreting ANOVA Summaries

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
summary(m)
```

Effect of **Species** and **Sex**

- **`(Intercept)`** represents base category but is a combination of factors
- Much more complicated to interpret
- Comparisons are often not of interest anyway (unless you've set up contrasts, which are advanced stats but awesome!)

So let's look at ANOVA tables instead

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816.87 -217.80  -16.87  227.61  882.20
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3372.39      31.43 107.308   <2e-16 ***
## speciesChinstrap    26.92      46.48   0.579    0.563
## speciesGentoo     1377.86      39.10  35.236   <2e-16 ***
## sexmale            667.56      34.70  19.236   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.6 on 329 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8468,    Adjusted R-squared:  0.8454
## F-statistic: 606.1 on 3 and 329 DF,  p-value: < 2.2e-16
```

Interpreting ANOVA Tables

Type I

```r
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```r
anova(m)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 145190219 72595110  724.21 < 2.2e-16
***
## sex         1  37090262 37090262  370.01 < 2.2e-16
***
## Residuals 329  32979185   100241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

Interpreting ANOVA Tables

Type I

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
anova(m)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 145190219 72595110  724.21 < 2.2e-16
***
## sex         1  37090262 37090262  370.01 < 2.2e-16
***
## Residuals 329  32979185   100241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

Overall effects of **Species** and **Sex**

- Yes there are differences among **Species** ($P < 2.2e^{-16}$)
- Yes there are differences between **Sexes** ($P < 2.2e^{-16}$)

Interpreting ANOVA Tables

Type I

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
anova(m)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 145190219 72595110  724.21 < 2.2e-16
***
## sex         1  37090262 37090262  370.01 < 2.2e-16
***
## Residuals 329  32979185   100241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

Overall effects of **Species** and **Sex**

- Yes there are differences among **Species** ($P < 2.2e^{-16}$)
- Yes there are differences between **Sexes** ($P < 2.2e^{-16}$)

Not a whole lot of information...
Stay tuned for **Post-Hoc** tests next week!

Interpreting ANOVA Tables

Type I

```
m1 <- lm(body_mass_g ~ species + sex, data =
penguins)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##              Df    Sum Sq  Mean Sq F value    Pr(>F)
## species       2 145190219 72595110  724.21 < 2.2e-16
***
## sex           1  37090262 37090262  370.01 < 2.2e-16
***
## Residuals 329  32979185   100241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

```
m2 <- lm(body_mass_g ~ sex + species, data =
penguins)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##              Df    Sum Sq  Mean Sq F value    Pr(>F)
## sex           1  38878897 38878897  387.86 < 2.2e-16
***
## species       2 143401584 71700792  715.29 < 2.2e-16
***
## Residuals 329  32979185   100241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

- For Type I ANOVAs, order matters with unbalanced samples
  - See that **`Sum sq`**, **`Mean Sq`** and **`F value`** all differ between the models
- Here, pretty minor, but important to remember with greater unbalances

Interpreting ANOVA Tables

Type III

```
m <- lm(body_mass_g ~ species + sex, data = penguins)
```

```
library(car)
Anova(m, type = "3")
```

```
## Anova Table (Type III tests)
##
## Response: body_mass_g
##                 Sum Sq  Df  F value    Pr(>F)
## (Intercept) 1154266972   1 11514.96 < 2.2e-16 ***
## species      143401584   2   715.29 < 2.2e-16 ***
## sex           37090262   1   370.01 < 2.2e-16 ***
## Residuals     32979185 329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

Type II ANOVAs do exist as well, but generally we use Type III in natural sciences

Interpreting ANOVA Tables

Type III

```r
m1 <- lm(body_mass_g ~ species + sex, data =
penguins)
Anova(m1, type = "3")
```

```
## Anova Table (Type III tests)
##
## Response: body_mass_g
##                 Sum Sq  Df  F value    Pr(>F)
## (Intercept) 1154266972   1 11514.96 < 2.2e-16 ***
## species      143401584   2   715.29 < 2.2e-16 ***
## sex           37090262   1   370.01 < 2.2e-16 ***
## Residuals     32979185 329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

```r
m2 <- lm(body_mass_g ~ sex + species, data =
penguins)
Anova(m2, type = "3")
```

```
## Anova Table (Type III tests)
##
## Response: body_mass_g
##                 Sum Sq  Df  F value    Pr(>F)
## (Intercept) 1154266972   1 11514.96 < 2.2e-16 ***
## sex           37090262   1   370.01 < 2.2e-16 ***
## species      143401584   2   715.29 < 2.2e-16 ***
## Residuals     32979185 329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

- Type III and unbalanced samples: Not dependent on variable order

Homework (Practice)*

Consider flipper length your response variable and species and sex your predictor variables

1. Plot the relationship between flipper length and species and between flipper length and sex
2. Create an ANOVA model of flipper length and species
3. Check diagnostics
4. Interpret the **summary table**
5. Interpret the **ANOVA Table**
6. Create an ANOVA model of flipper length and species and sex
7. Check diagnostics
8. Interpret the **ANOVA Table**

* Not to be handed in, answers posted in these slides next week