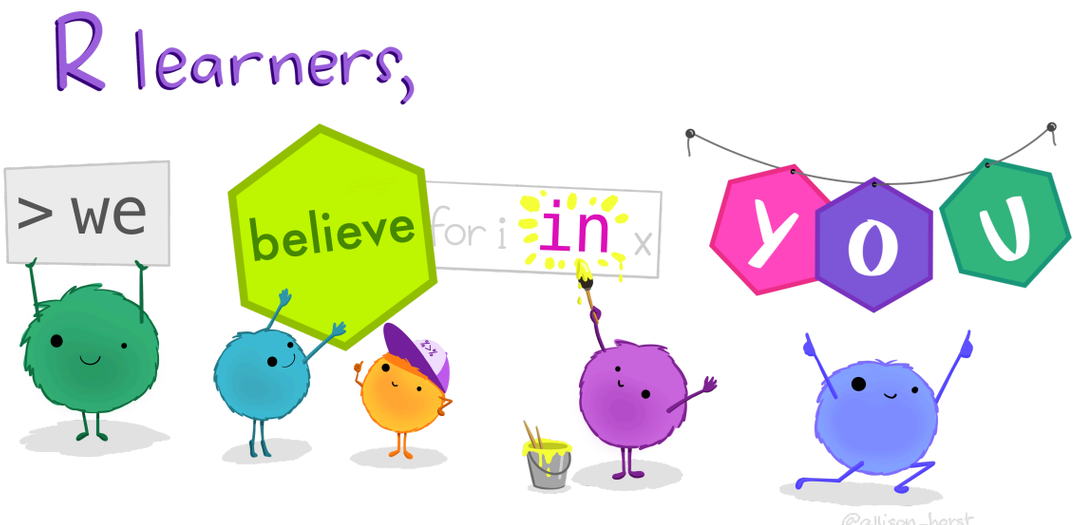
Intro to Stats in R

Linear Regression



alex-koiter
@Alex_Koiter@sciencemastodon.com
@alex_koiter
alexkoiter.ca

Compiled: 2025-03-28

You are not alone!

Getting started

1. Create and name a new folder
2. Open RStudio
3. Create R project
    - File -> New project -> Existing Directory
    - Browse to your folder
    - Select Create Project
4. Create new R script
    - File -> New File -> R Script

Getting started

Make sure to load packages at the top:

```
1 library(tidyverse)
2 library(palmerpenguins)
```

Getting started

We need one additional package:

`ggfortify` provides plotting tools for commonly used statistics

```
1 install.packages("ggfortify")
```

Add it to the top with the others

```
1 library(tidyverse)
2 library(palmerpenguins)
3 library(ggfortify)
```

Getting started with stats in R

1. Always start with a picture!

- Stats should support your figures
- Do you see the expected pattern?

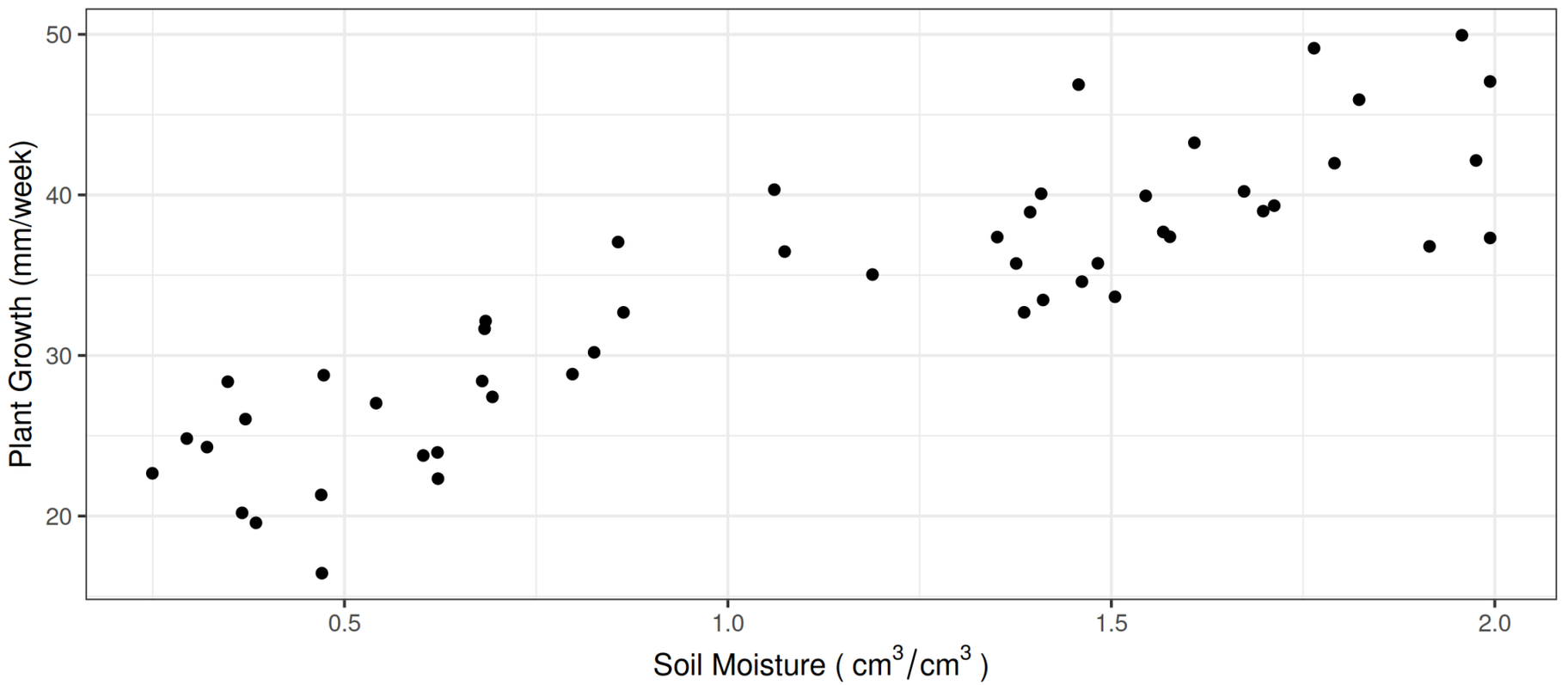
2. Translate your hypothesis into a statistical model

- Becomes easier with experience

3. Translate your statistical model in R language

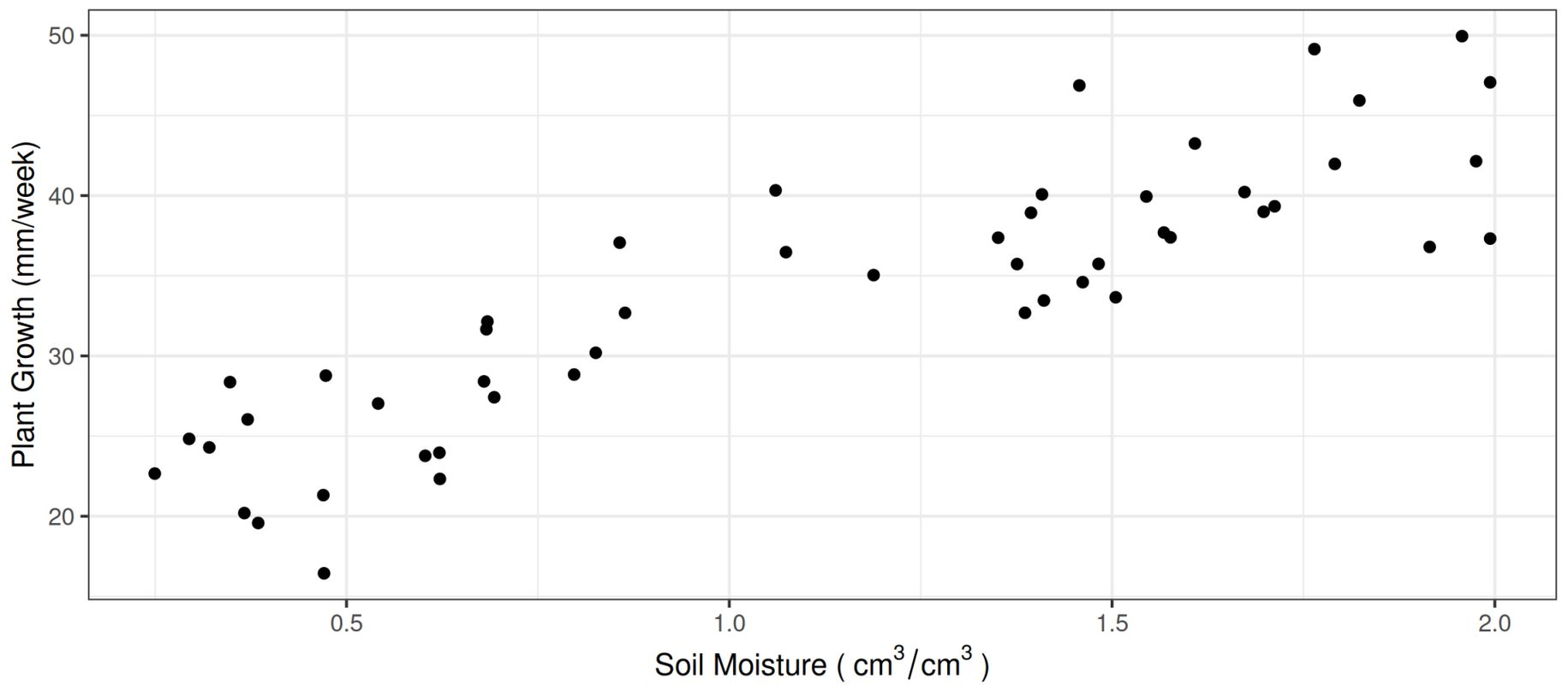
Picture

- You are all familiar with how to use `ggplot2` to create a figure
  - Go back to your previous scripts and notes



Picture

- We see a positive relationship (slope)
  - More soil moisture results in higher plant growth (as expected)



Hypothesis into a statistical model

Hypothesis is that more moisture allows for higher growth rates

- Need to identify
  - Response (dependent, y) variable
  - Explanatory (independent, x) variable
  - Continuous or discreet/categorical

Hypothesis into a statistical model

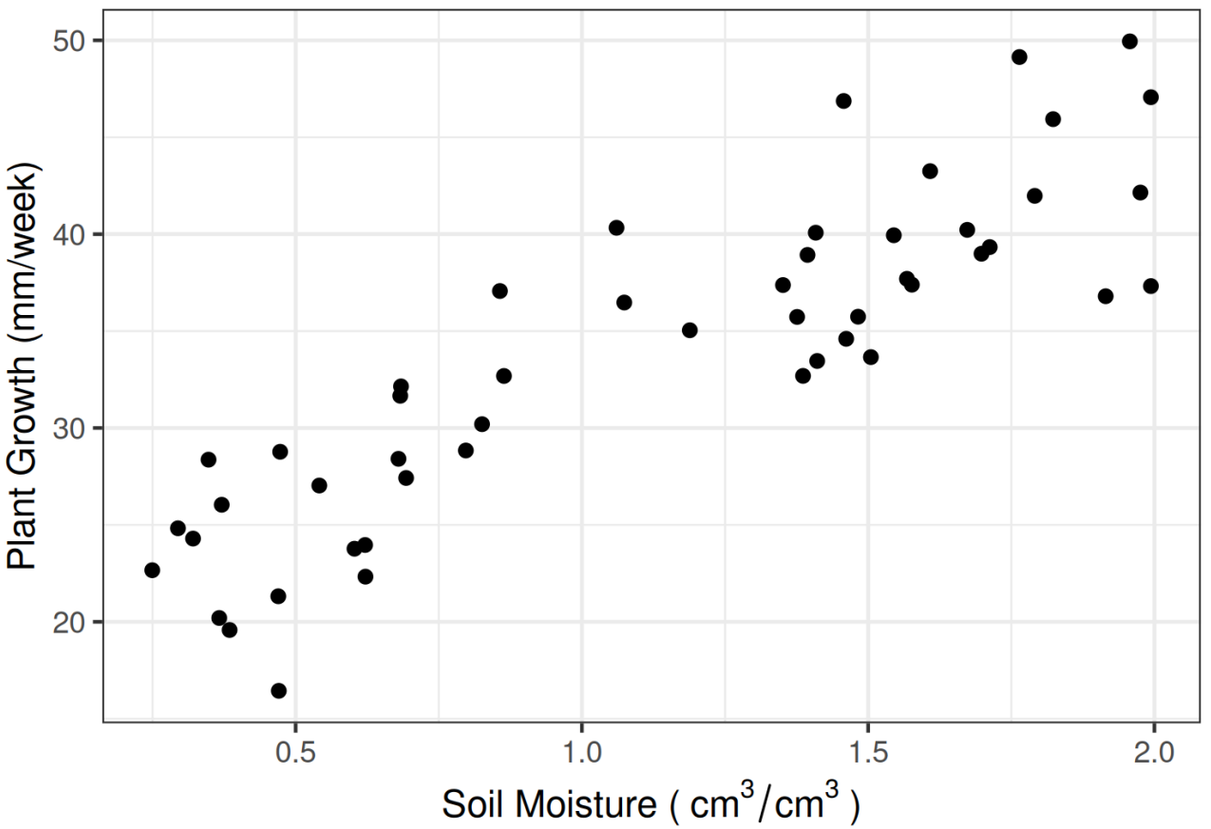
When both x and y are continuous we have a simple linear regression

Response variable (y)

- Plant growth
- Continuous

Explanatory variable (x)

- Soil Moisture
- Continuous



Linear regression

Running model in R

```
1  lm(y ~ x, data = data)
```

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

Linear regression

Running models in R

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

- You can have multiple explanatory variables
- Can be all continuous, all discreet, or a mix
  - But these are often called: ANOVA, ANCOVA, Multiple regression, etc.

Linear regression

Running models in R

```
1 lm(plant.growth.rate ~ soil.moisture.content, data = plant_gr)
```

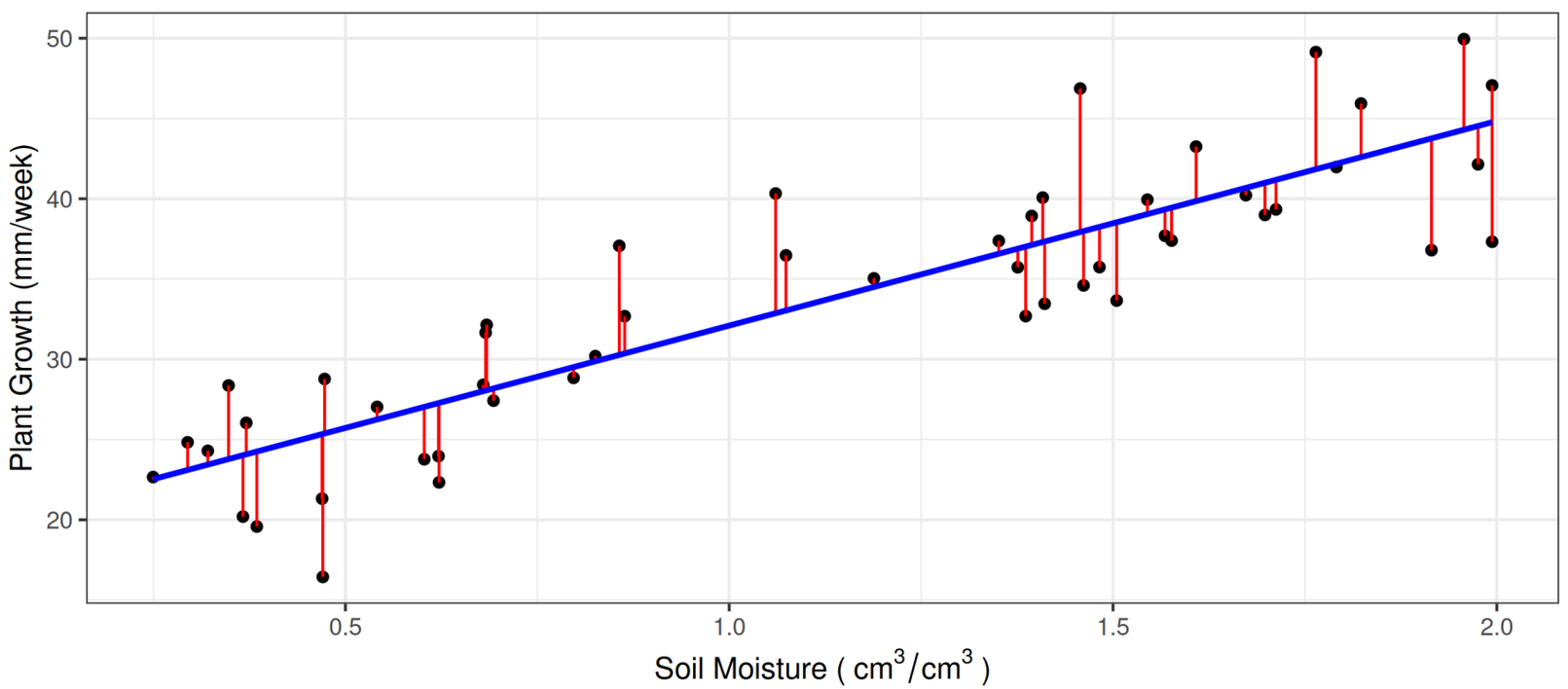
- This reads as: Fit a linear model, where we hypothesize that plant growth rate is a function of of soil moisture content, using data from the the `plant_gr` data frame.

Assumptions first!

1. Assumption of equal variance
2. Assumption of normality of residuals
3. Evaluate leverage (influential data points) - technically not an assumption

What are residuals?

- Residuals are difference between the observed value of the dependent variable (y) and the value predicted by the model
  - A measure of how well the model fits the data



Assumptions first!

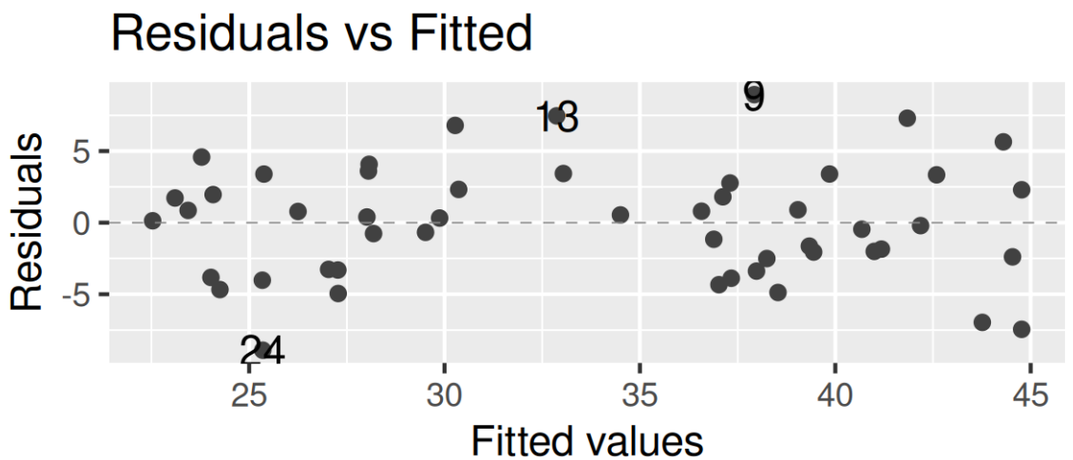
```
model_pgr <- lm(plant.growth.rate ~ soil.moisture.content,
                data = plant_gr)
```

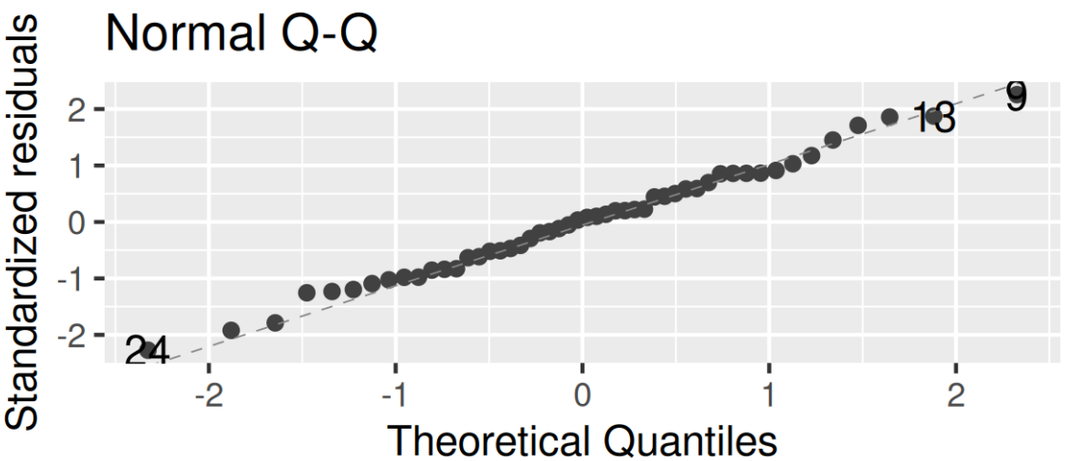
```
autoplot(model_pgr, smooth.colour = NA)
```

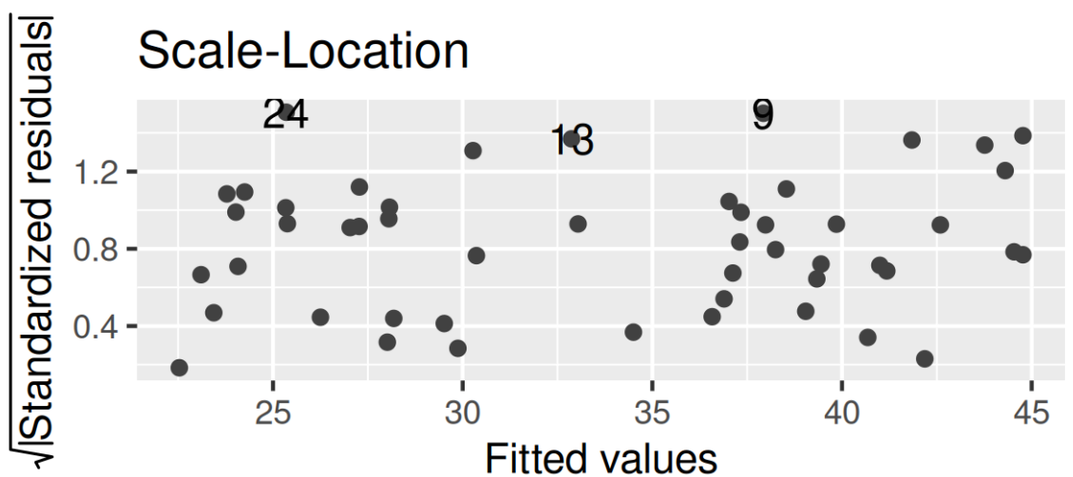
- The `autoplot()` function is part of the `ggfortify` package
- The `smooth.colour = NA` argument suppresses the “wiggly” line that is a locally weighted regression line.
  - Some argue the line can be unhelpful as one tends to look at the line and not the data

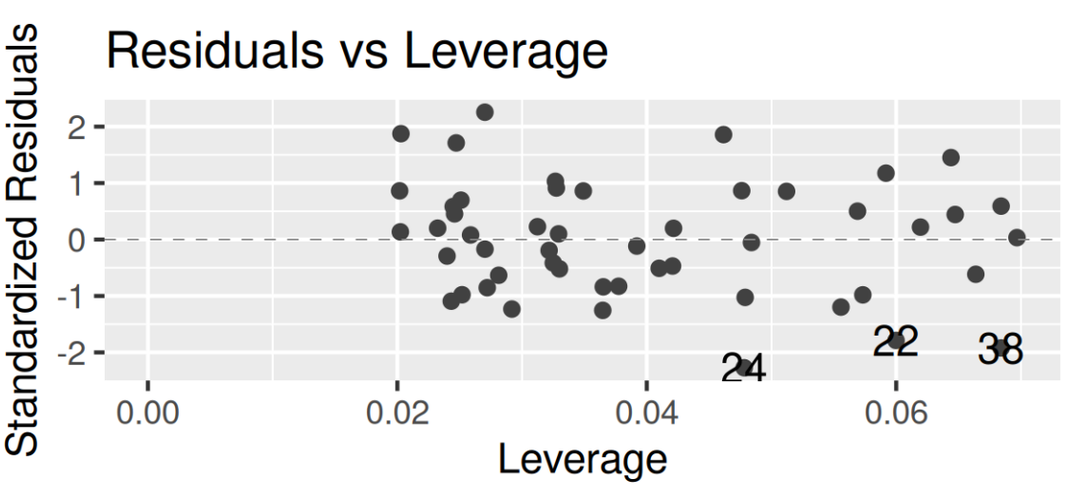
Assumptions first!

```
1 autoplot(model_pgr, smooth.colour = NA)
```





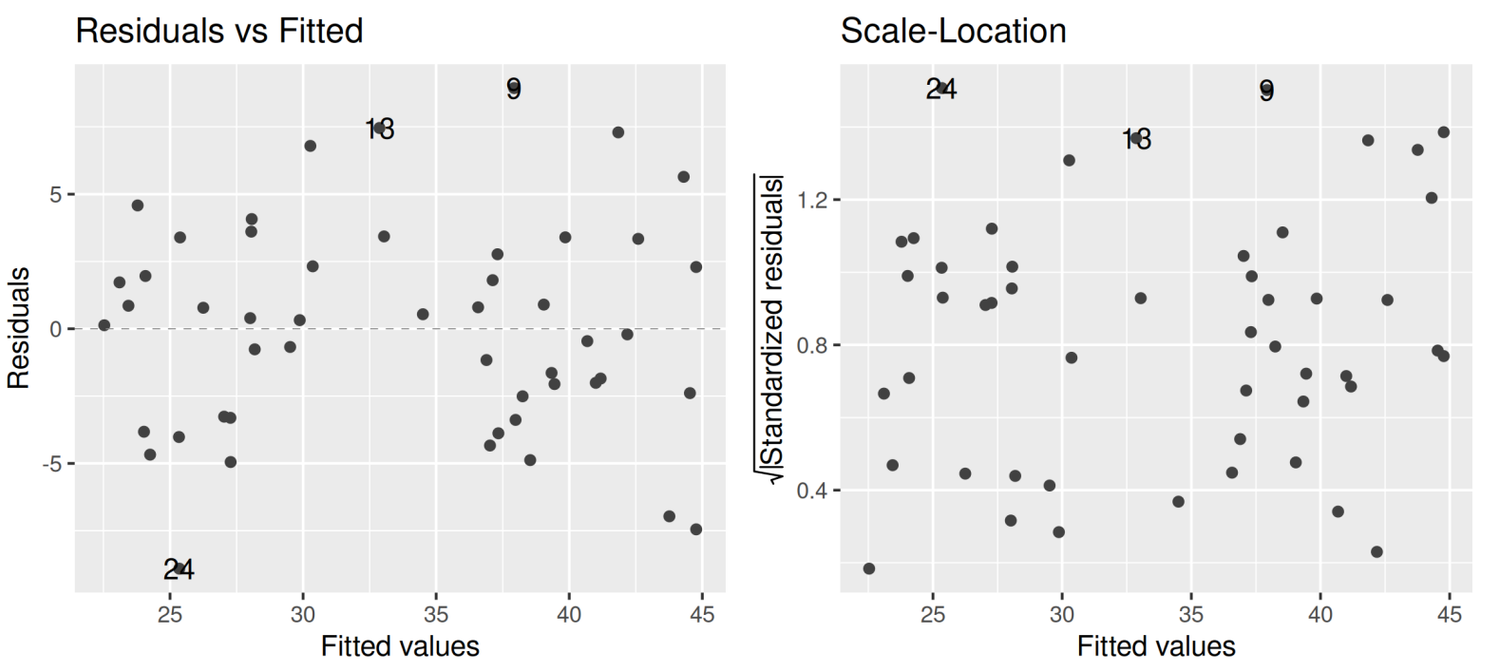




Equal variance

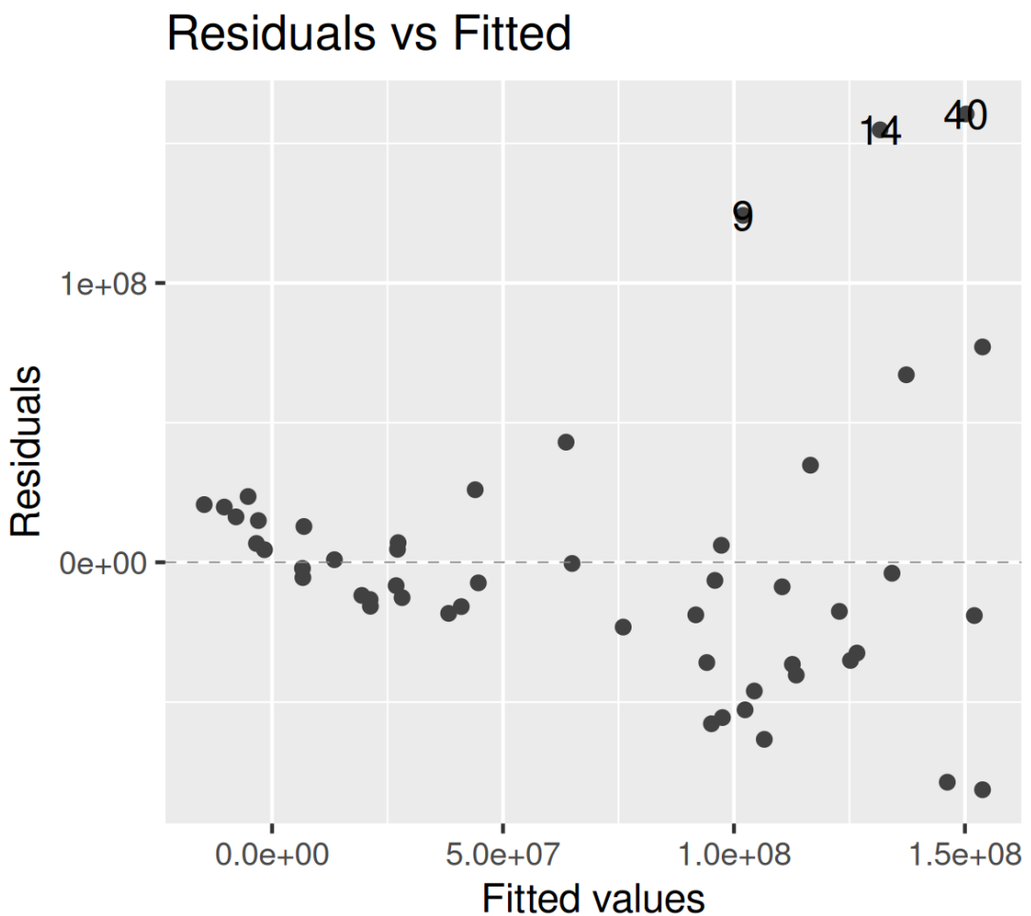
Homoscedasticity

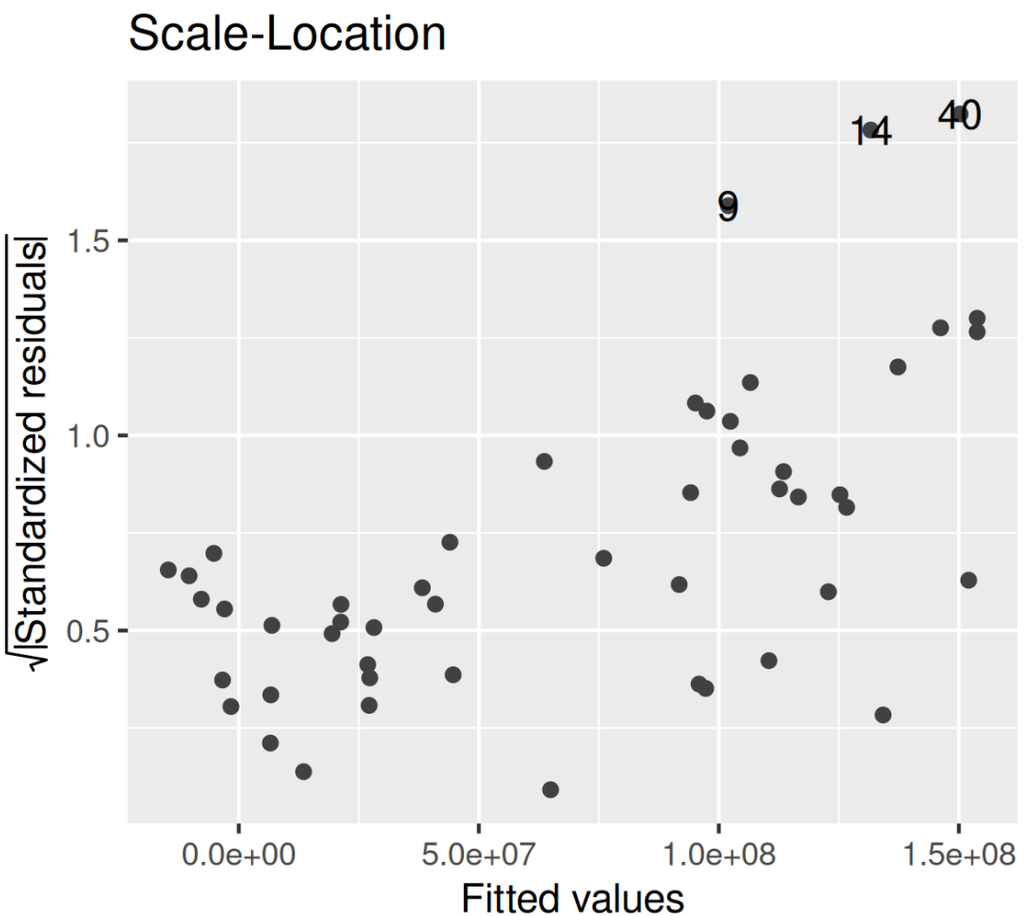
- The spread of residuals does not systematically increase or decrease as the predictor values change



Unequal variance

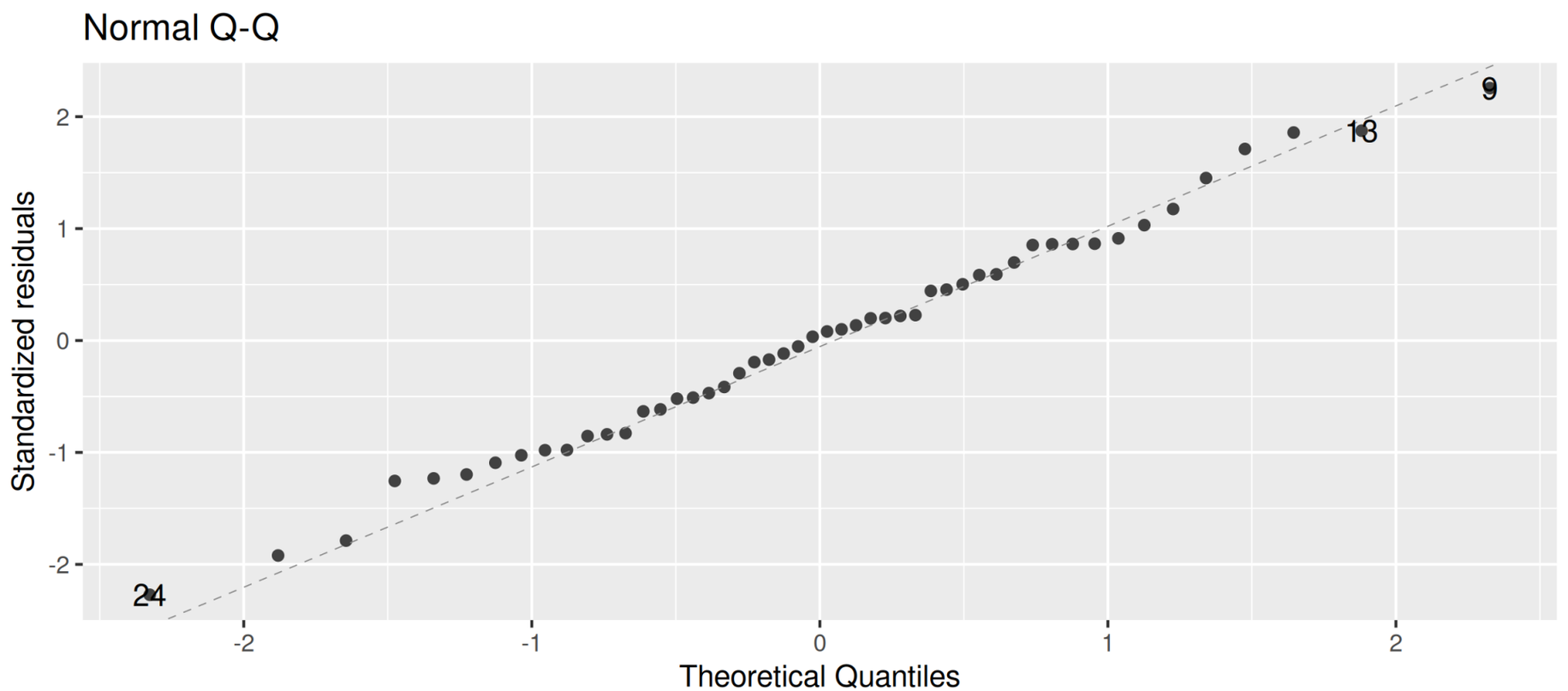
Heteroscedasticity





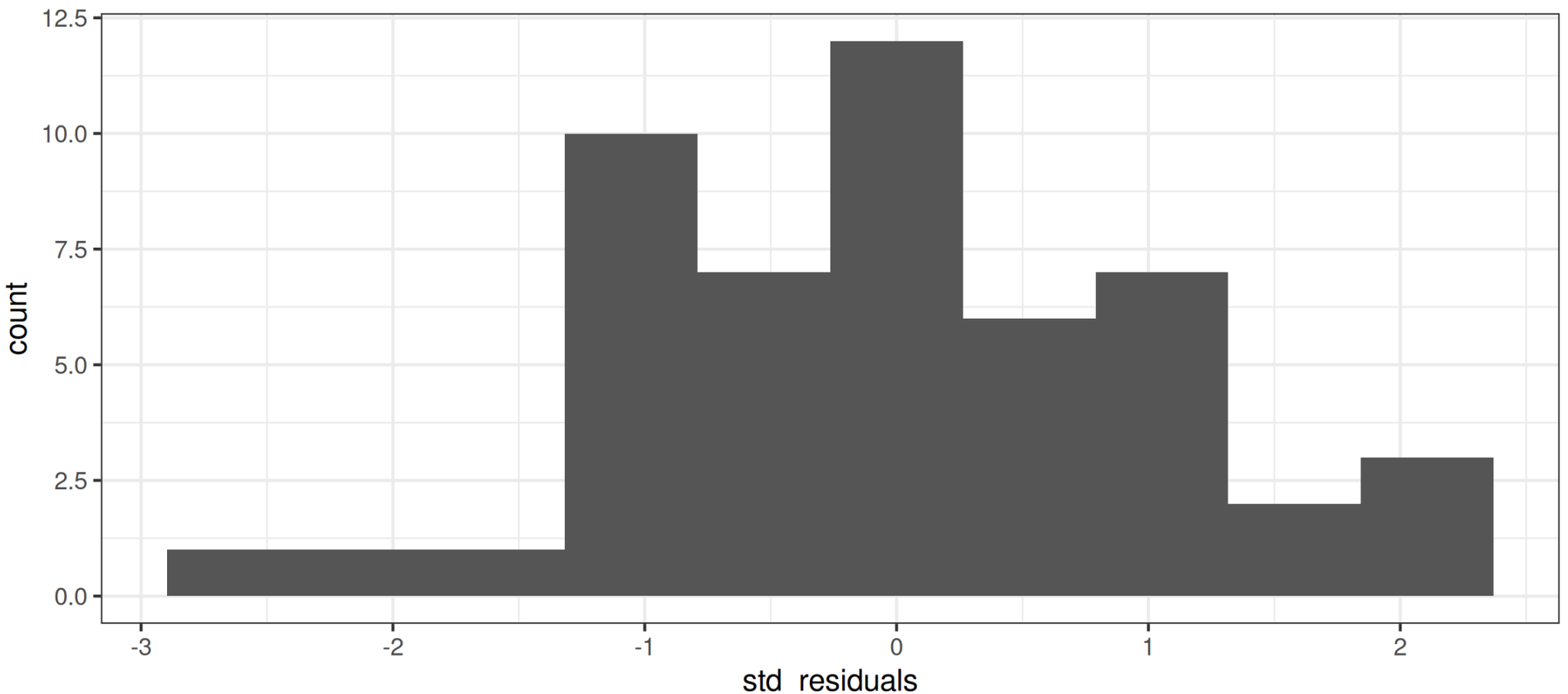
Normal distribution of residuals

- Assess the distribution of residuals - looking for a normal distribution
  - Ensures the validity of conclusions made

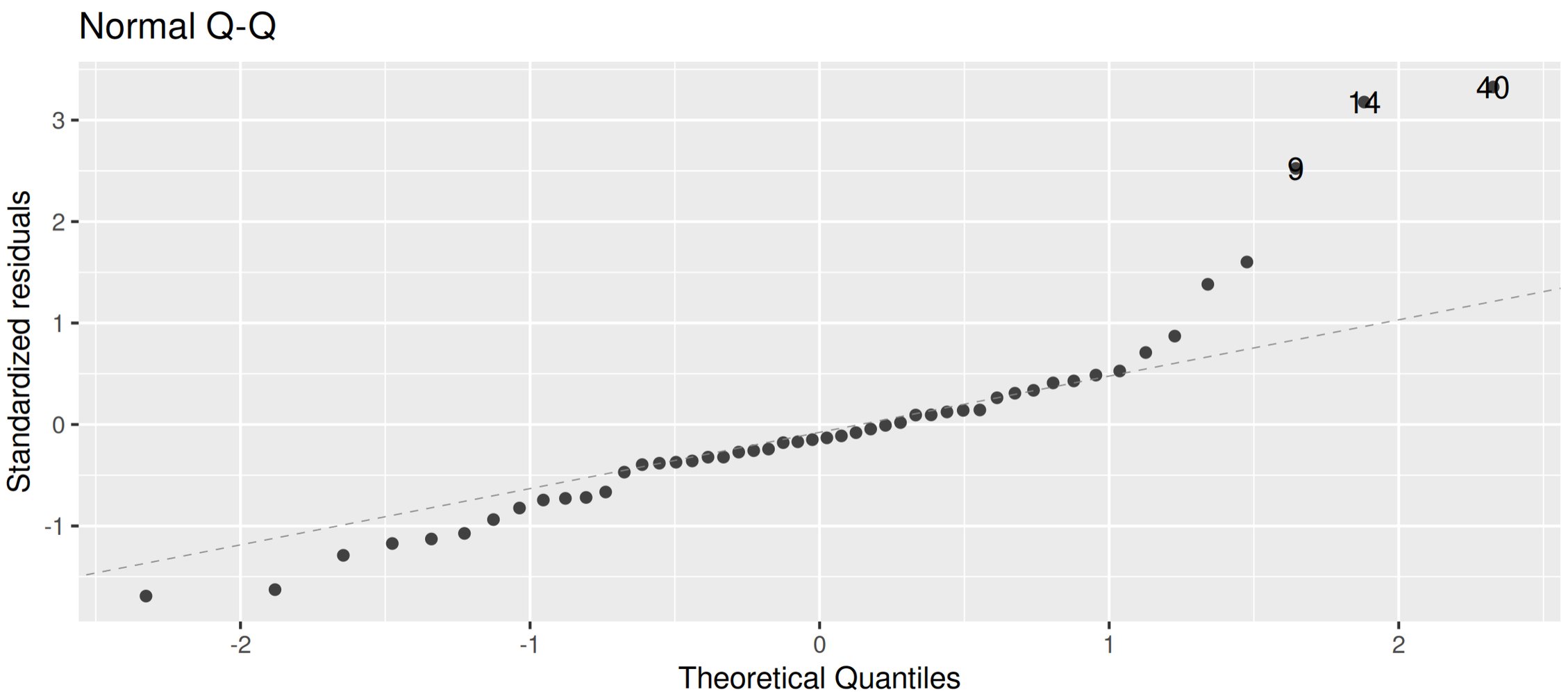


Normal distribution of residuals

- Q-Q plots are a better tool when you have a sample size < 100



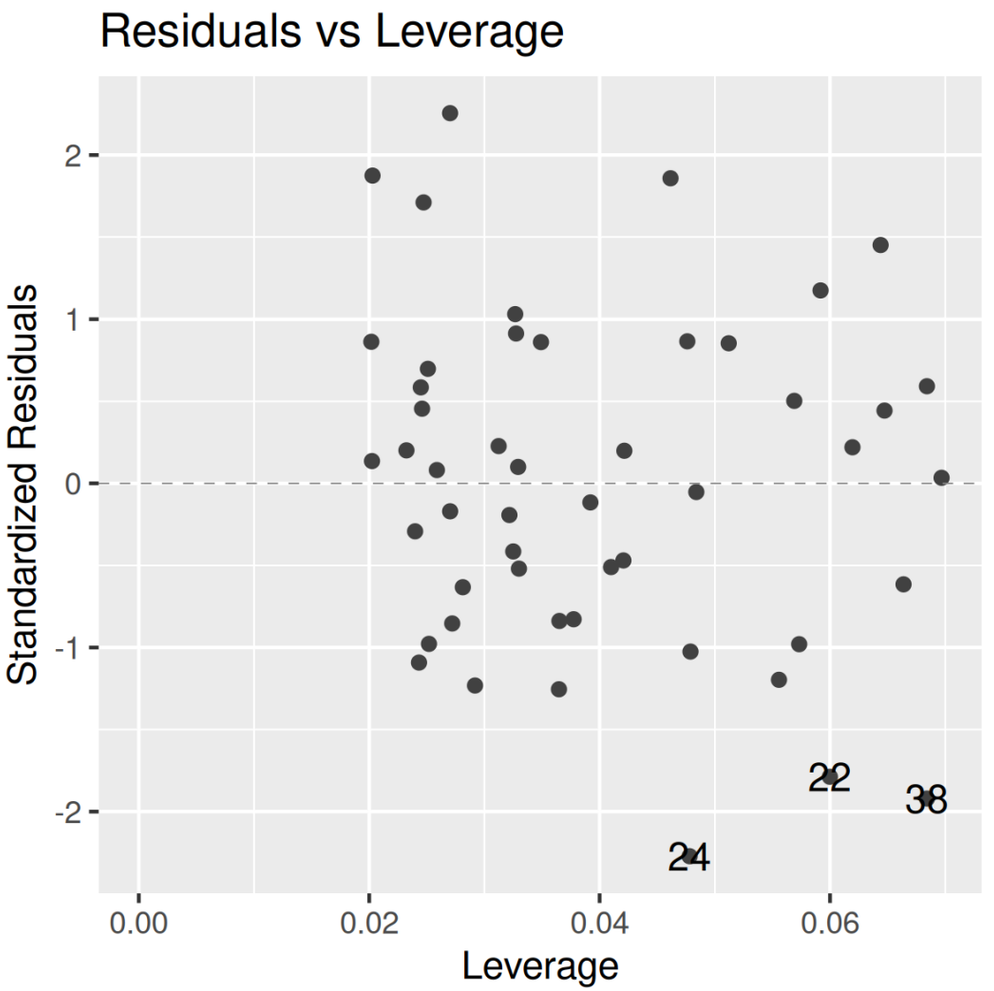
Non-normal distribution



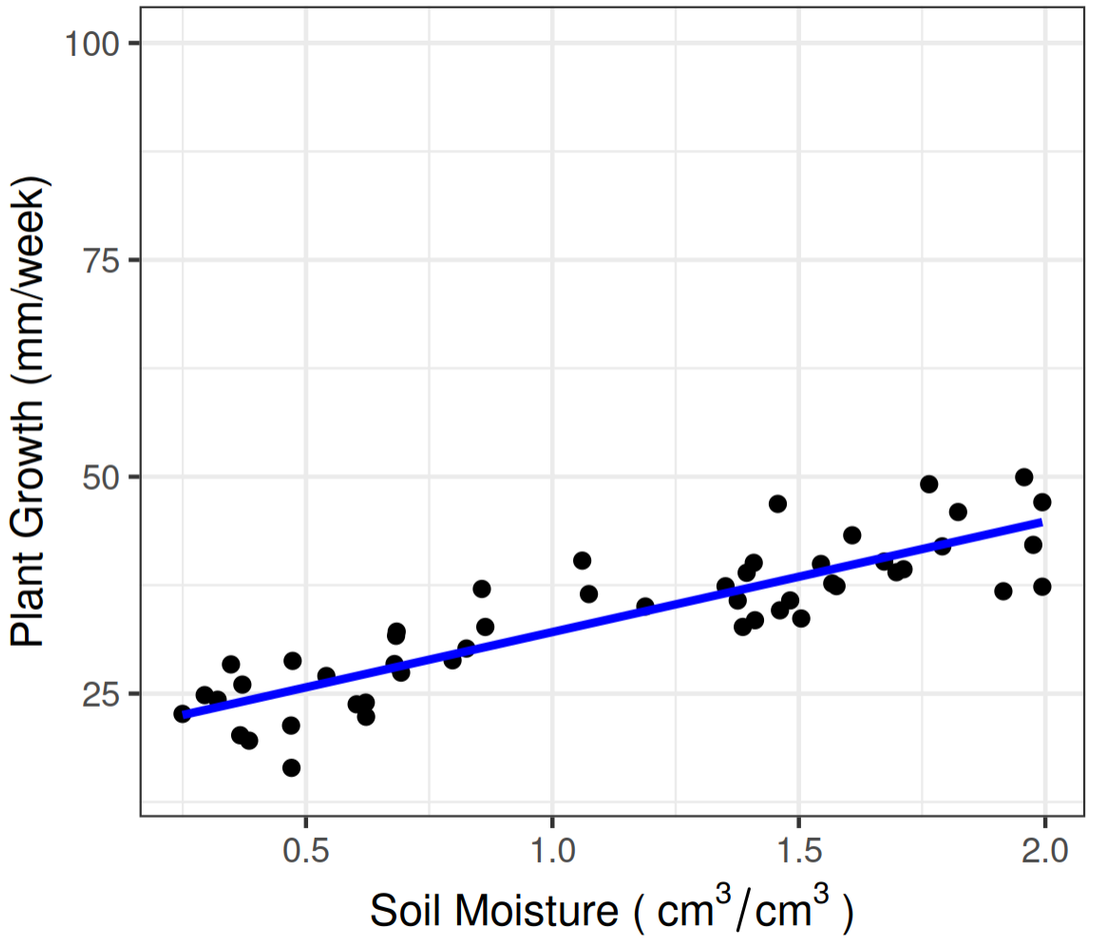
Leverage

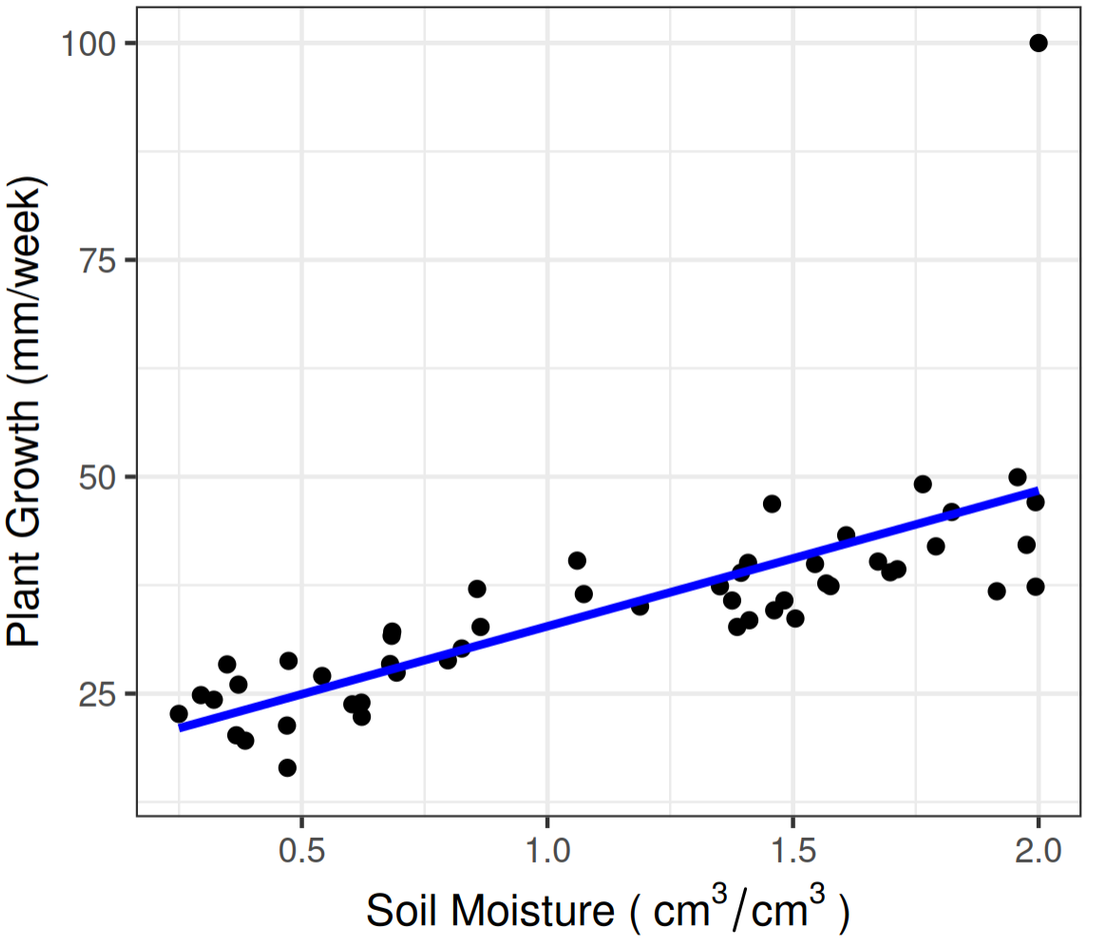
Be very careful here!

- Data points whose response variable (y) does not follow the general trend
  - Can skew results
- Do not throw out data because you don’t like it!
- Outlier, or meaningful aspect of your data?



Leverage


100
75
50
25
Plant Growth (mm/week)
0.5
1.0
1.5
2.0
Soil Moisture (cm^3/cm^3)


100
75
50
25
Plant Growth (mm/week)
0.5
1.0
1.5
2.0
Soil Moisture (cm^3/cm^3)

Model assumptions

You can use formal tests for assumptions

(e.g., Bartlett, Levene, Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling…)

I generally recommend **NOT** doing this

- Small meaningless differences may lead to rejection
- Just look at your data!

Model interpertation

Everything is looking good!

- Lets see if we can reject the null hypothesis
  - Soil moisture has no effect on plant growth rate
- We are going to use the `summary()` function to help us out

Model interpertation

```r
model_pgr <- lm(plant.growth.rate ~ soil.moisture.content, data = plant_gr)
```

```r
summary(model_pgr)
```

```
Call:
lm(formula = plant.growth.rate ~ soil.moisture.content, data =
plant_gr)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9089 -3.0747  0.2261  2.6567  8.9406

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             19.348      1.283   15.08   <2e-16 ***
soil.moisture.content   12.750      1.021   12.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.019 on 48 degrees of freedom
Multiple R-squared:  0.7648,    Adjusted R-squared:  0.7599
F-statistic: 156.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

Model interpertation

Model

```

Call:
lm(formula = plant.growth.rate ~ soil.moisture.content, data = plant_gr)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9089 -3.0747  0.2261  2.6567  8.9406 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)             19.348      1.283   15.08   <2e-16 ***
soil.moisture.content   12.750      1.021   12.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.019 on 48 degrees of freedom
Multiple R-squared:  0.7648,    Adjusted R-squared:  0.7599 
F-statistic: 156.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

Model interpertation

Effects

```
1
2  Call:
3  lm(formula = plant.growth.rate ~ soil.moisture.content, data = plant_gr)
4
5  Residuals:
6      Min      1Q  Median      3Q     Max
7  -8.9089 -3.0747  0.2261  2.6567  8.9406
8
9  Coefficients:
10                       Estimate Std. Error t value Pr(>|t|)
11 (Intercept)             19.348      1.283   15.08   <2e-16 ***
12 soil.moisture.content   12.750      1.021   12.49   <2e-16 ***
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15
16 Residual standard error: 4.019 on 48 degrees of freedom
17 Multiple R-squared:  0.7648,    Adjusted R-squared:  0.7599
18 F-statistic: 156.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

Model interpertation

Model interpertation

Model interpertation

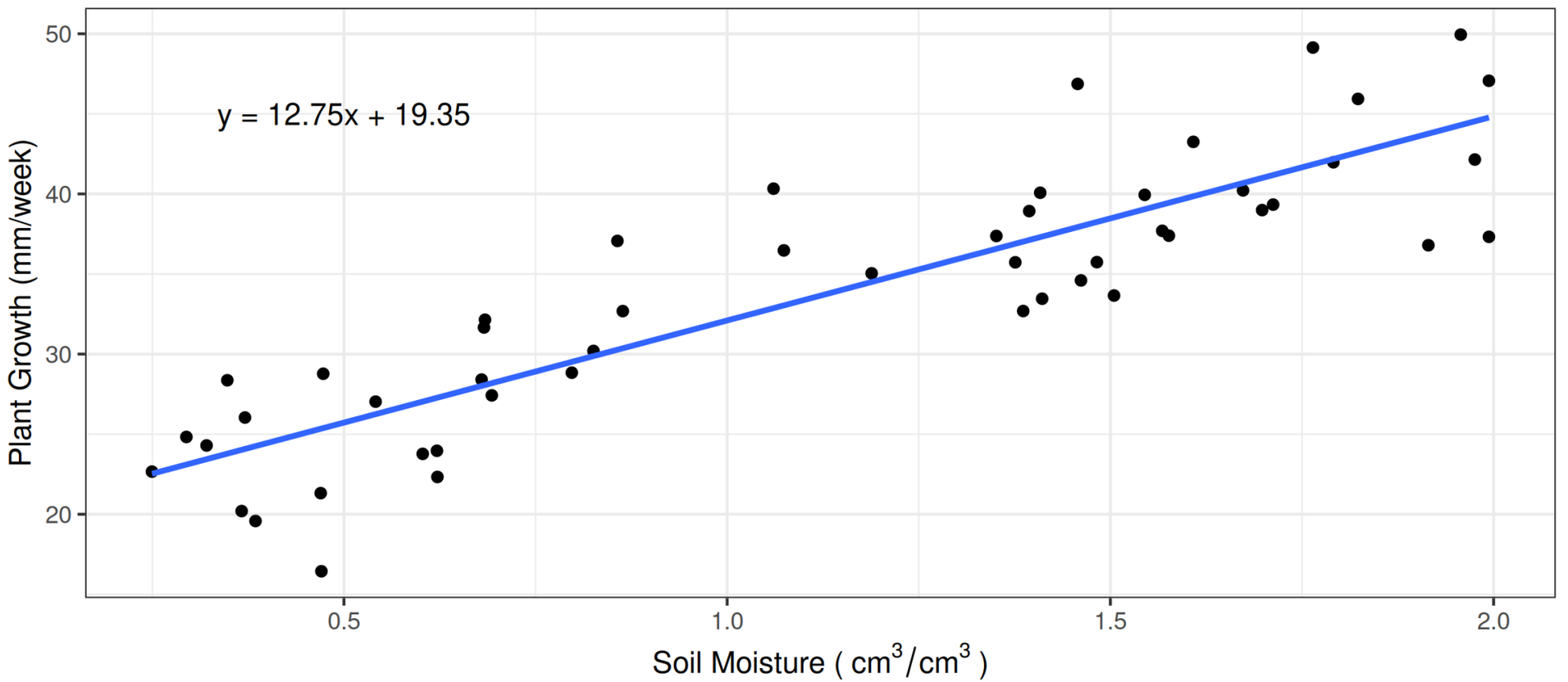
Model interpertation

Model interpertation

Back to the picture

$y = mx + b$

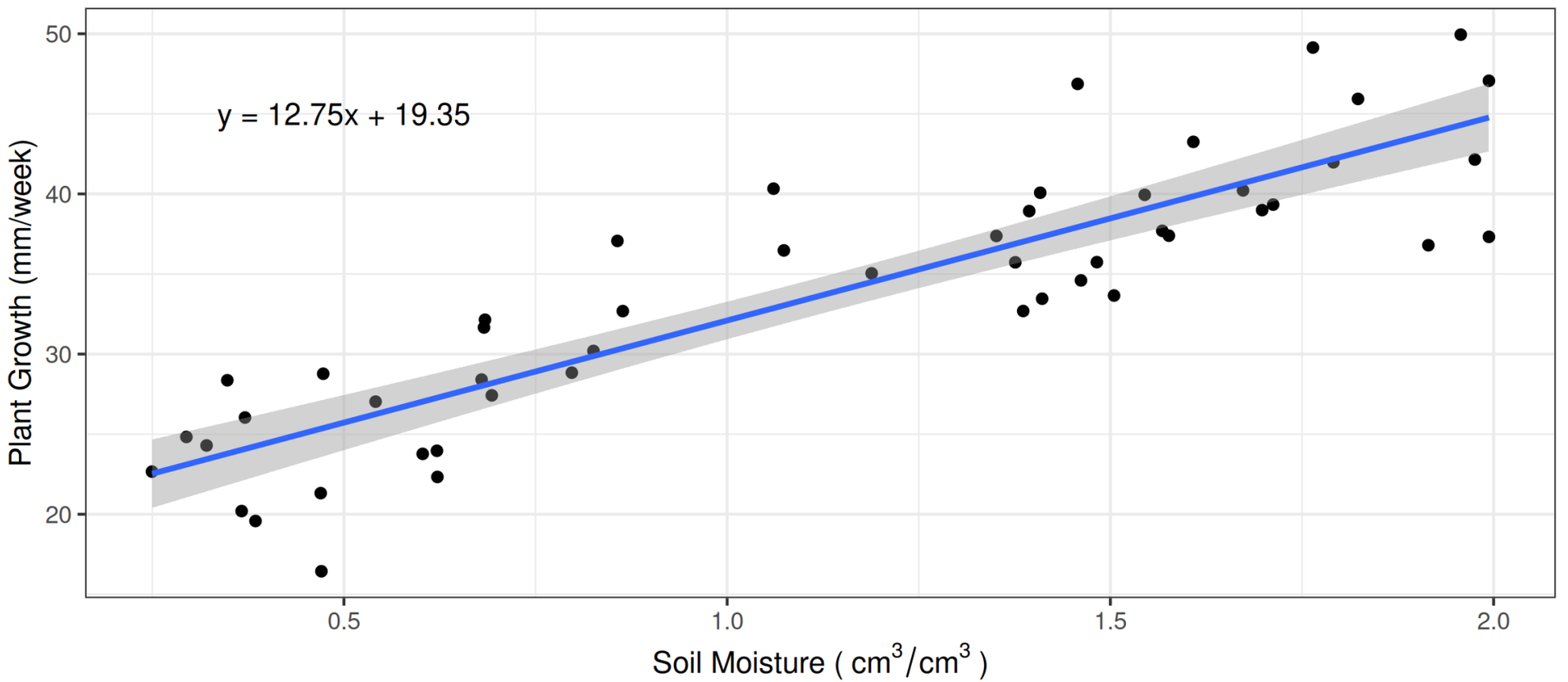
```
1 geom_smooth(method = "lm", se = FALSE)
```



Back to the picture

Confidence interval

```
1 geom_smooth(method = "lm", se = TRUE)
```



Confidence interval

- The range in which the mean (expected value) of the response variable is likely to fall for a given explanatory variable value

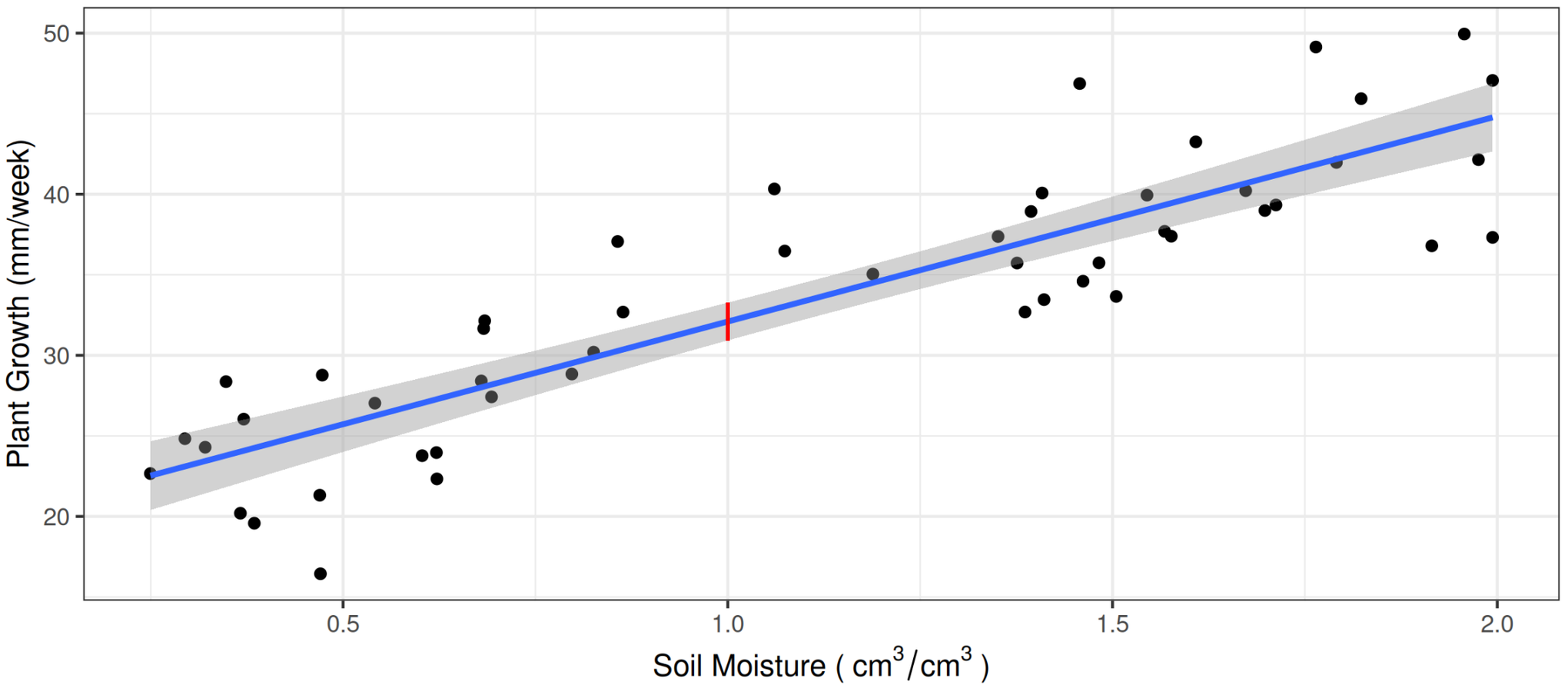
```r
new_data <- data.frame(soil.moisture.content = 1)
predict(model_pgr, newdata = new_data,
        interval = "confidence", level = 0.95)
```

```
    fit       lwr      upr
1 32.098 30.92568 33.27031
```

- A 95% confidence interval for plant growth rate at soil moisture content of 1 cm^3/cm^3 would be 30.93 to 33.27 mm/week
- This means we are 95% confident that the **mean plant growth rate** falls within this range

Confidence interval

- Double check



Prediction interval

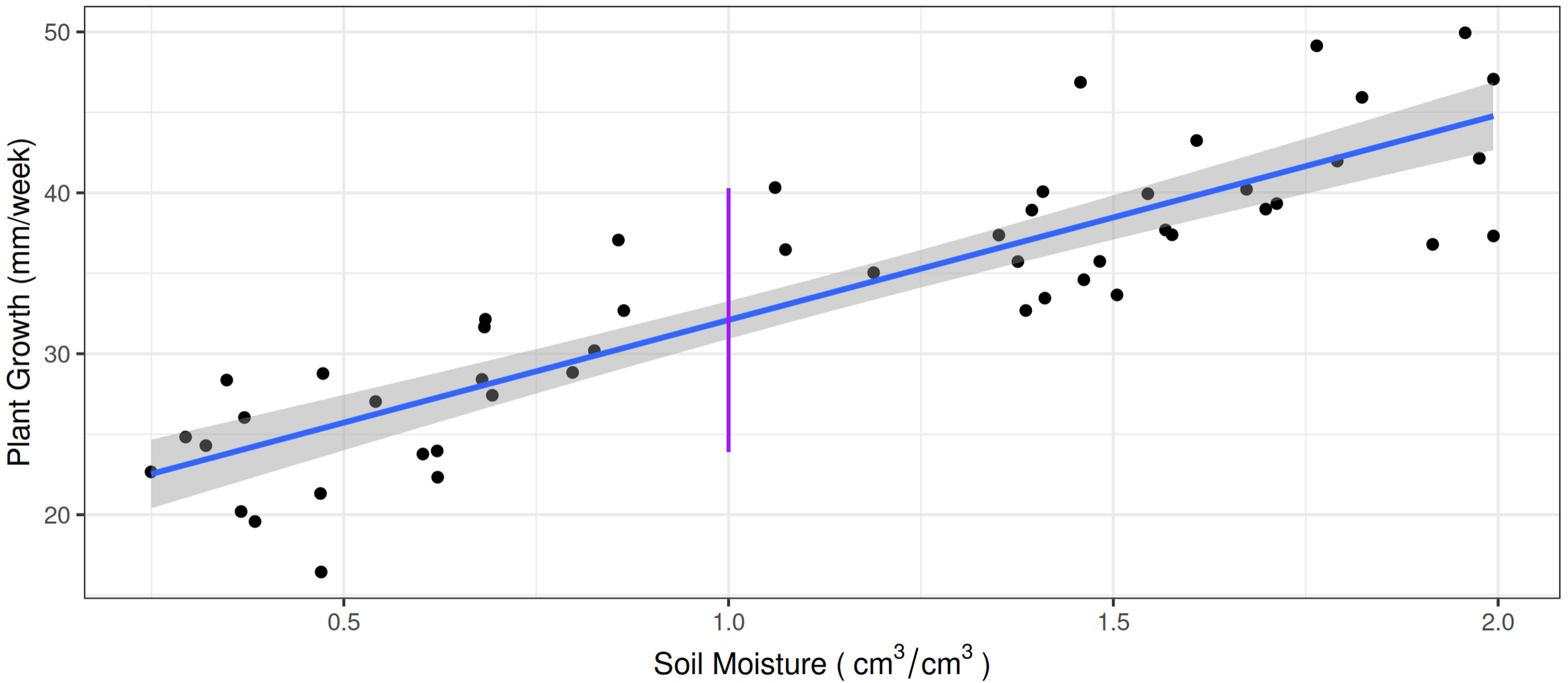
- The range in which an individual new observation is likely to fall for a given explanatory variable value

```r
new_data <- data.frame(soil.moisture.content = 1)
predict(model_pgr, newdata = new_data,
        interval = "predict", level = 0.95)
```

```
    fit       lwr      upr
1 32.098 23.93247 40.26352
```

- A 95% prediction interval for plant growth rate at soil moisture content of 1 cm^3/cm^3 would be 23.93 to 40.26 mm/week
- This means we are 95% confident that if we take a **random** plant at that moisture content it will fall within this range

Prediction interval



Your turn!

Does flipper length, bill depth, and bill length in penguins vary with body mass?

Each student is going to do something slightly different!

1. bill length (Adelie)
2. bill depth (Adelie)
3. flipper length (Adelie)
4. bill length (Gentoo)
5. bill depth (Gentoo)
6. flipper length (Gentoo)

Your turn!

We need to filter the data by penguin species

```
1 adelie_data <- filter(penguins, species == "Adelie")
```

or

```
1 gentoo_data <- filter(penguins, species == "Gentoo")
```

Create your picture

```
ggplot(data =  ____, aes(_____________________________________________)) +
  geom_____() +
  theme_____() +
  labs(y = ____, x = ____)
```

Linear regression

```
1 _______ <- lm(_______ ~ _______, data = _______)
```

Assess your assumptions

```
1 autoplot(_______, smooth.colour = _______)
```

Model interpertation

```
1 summary(_______)
```

Final picture

```
penguin_plot <- ggplot(data =  ____, aes(_____________________________________
  geom_____() +
  theme_____() +
  labs(y = ____, x = ____) +
  ____(method = ____, se = ____)
penguin_plot
```

Struggling to get your desired colours or text size? Please ask!

The plot preview may look different the the saved version.

```
ggsave(filename = ____, plot = ____, height = ____,
       width = ____, units = "mm", dpi =600)
```

Assignment

Complete the following and compile into a single well-formatted document

- Include identifying information (name, date, title, etc)
- PDF format
- Figures are numbered and have meaningful captions
- Pay attention to spelling and grammar

Additionally

- Submit your R script file so I can reproduce your work! (not the .Rproj file)

Assignment

1. Create a figure showing the relationship showing the relationship between body weight and flipper length, bill depth, or bill length for either Adelie or Chinstrap penguins. Include the standard error (95% confidence interval) about the regression line
2. Using the diagnostic plots for your linear regression model (do not include figure in the report), evaluate the assumptions of normality, equal variance, and influential points/outliers based on your output. Summarize any potential violations and their possible impact on the interpretation of the model

Assignment

3. Determine the regression model (intercept and slope) and express the equation
4. Determine the r^2 for the model
5. Test the significance of the modeled relationship (i.e., slope) using an of 0.05
6. Discuss the model's effectiveness in explaining bill depth, length, or flipper length using the coefficient of determination and significance test results

Assignment

7. Explain what your r^2 value indicates about the relationship between the independent and dependent variables in your model
8. If the r^2 value is less than 100%, suggest another potential independent variables that could added (i.e., multiple regression) to improve the model

Wrapping up

- Stats can be confusing at the best of times!
- Be kind to yourself
- It takes practice!
- Your stats should support your figures (not the other way around

Thank you!

alexkoiter.ca

koitera@brandonu.ca

Slides created with Quarto on 2025-03-28