# Checklist

1. Would you help me? If yes, unmute yourself, and let me know.
   1. Be my co-host, and admit people from the waiting room.
   2. If there is a question in the chat, please remind me.
   3. If you notice somebody raises their hand during lecture, please remind me.
2. Type your questions, if any, in the chat now.
3. I will set an alarm for 1hr 45min at 12:45pm.
4. I will record to the cloud.

# Unit 3: Basic regression

## 2. How useful is your linear model?

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

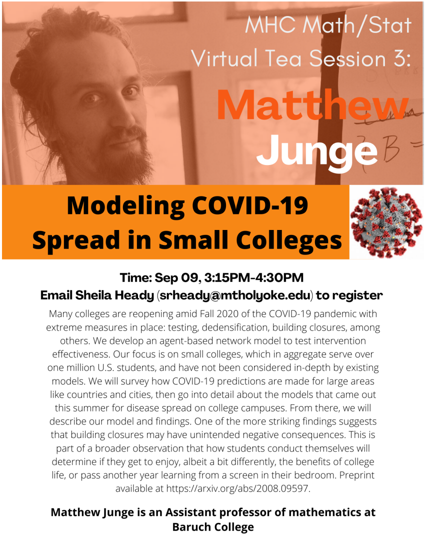
# Outline

1. Announcement

2. Today: How useful is your linear model
    1. Finding the least square line in R
    2. How useful is the model?

3. Summary

# Math/Stat Virtual Tea #3



MHC Math/Stat
Virtual Tea Session 3:
**Matthew Junge**

**Modeling COVID-19 Spread in Small Colleges**

**Time: Sep 09, 3:15PM-4:30PM**
**Email Sheila Heady (srheady@mtholyoke.edu) to register**

Many colleges are reopening amid Fall 2020 of the COVID-19 pandemic with extreme measures in place: testing, dedensification, building closures, among others. We develop an agent-based network model to test intervention effectiveness. Our focus is on small colleges, which in aggregate serve over one million U.S. students, and have not been considered in-depth by existing models. We will survey how COVID-19 predictions are made for large areas like countries and cities, then go into detail about the models that came out this summer for disease spread on college campuses. From there, we will describe our model and findings. One of the more striking findings suggests that building closures may have unintended negative consequences. This is part of a broader observation that how students conduct themselves will determine if they get to enjoy, albeit a bit differently, the benefits of college life, or pass another year learning from a screen in their bedroom. Preprint available at https://arxiv.org/abs/2008.09597.

**Matthew Junge is an Assistant professor of mathematics at Baruch College**

# How do you remove an outlier?

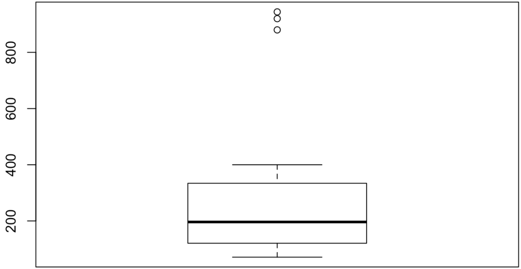
The extra credit problems are meant to be a challenge for you, meaning they often ask questions that are related to what we are doing in class, but require you to learning something on your own before you can solve them.

The process involves

- asking good questions
- finding information
- decision making
- learning
- application

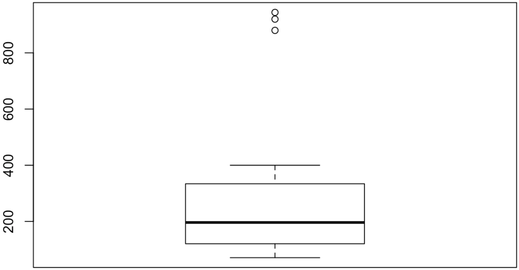
Step 1: Begin with an example

Let's remove the three outliers in the following boxplot.



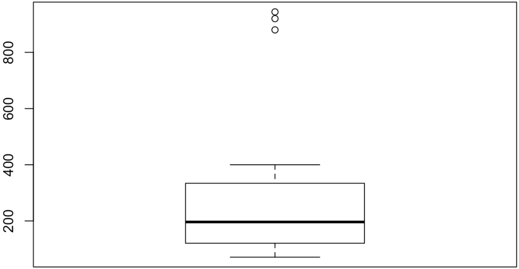
Step 2: Write down what you did.

1. Find the three points that are bigger than 800
2. Remove those three points



Step 2: Write down what you did, being as specific as possible.

1. Sort the data in descending order
2. Remove the first three rows



# How do you remove an outlier?

Step 3: Translate this to code

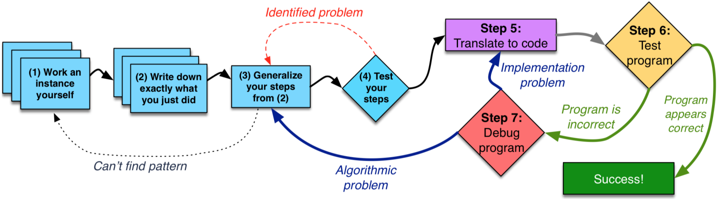
1. Sort the data in descending order
2. Remove the first three rows

```
new data <-   data %>%
                arrange(desc(var)) %>%
                slice(4:n())
```

# The Seven Steps

A Technique for Translation from Problem to Code

Andrew Hilton, Genevieve Lipp, and Susan Rodger
Duke University, ECE and CS Departments



# Outline

1. Announcement

2. Today: How useful is your linear model
    1. Finding the least square line in R
    2. How useful is the model?

3. Summary

# Outline

1. Announcement

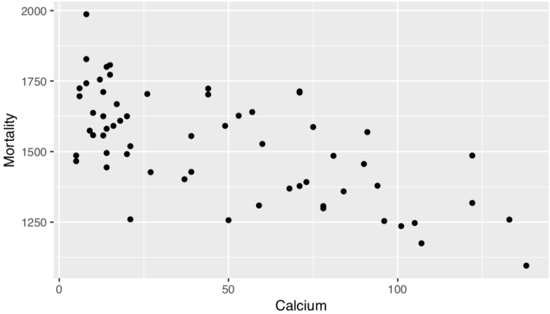
2. Today: How useful is your linear model
    1. Finding the least square line in R
    2. How useful is the model?

3. Summary

# Mortality and hard water

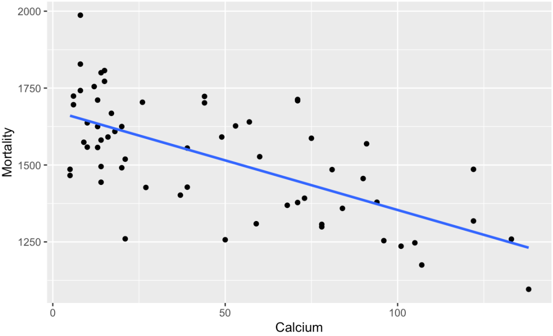
Scientists believe that hard water (water with high concentrations of calcium and magnesium) is beneficial for health.

We have recordings of the mortality rate (deaths per 100,000 population) and concentration of calcium in drinking water (parts per million) in 61 large towns in England and Wales.



# Add a line

```
ggplot(data = mortality_water, mapping = aes(x = Calcium, y = Mortality)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```



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

# A new function

General form of ‘lm‘ command:

```
lm(y_variable ∼ x_variable, data = data_frame)
```

Use this to estimate the intercept and the slope of line in the Mortality/Health data.

```
linear_fit ← lm(Mortality ∼ Calcium, data = mortality_water)
linear_fit
```

```
Coefficients:
(Intercept)      Calcium
   1676.356       -3.226
```

From the coefficients, we can write the regression line in the Mortality/Health data as

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

Abstractly,

$$\hat{y} = 1676 - 3x$$

# Predict

One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

*By hand:*

$$\widehat{\text{Mortality}} = 1676 - 3 \times 71 = 1463$$

The predicted value for the mortality rate in that town is 1463 deaths per 100,000 population.

# Predict

*By R:*

The general form of ‘predict‘ command:

predict(linear_model, newdata = data_frame)

*Code:*

```
linear_fit <- lm(Mortality ~ Calcium, data = mortality_water)
predict_data <- data.frame( Calcium = 71)
predict(linear_fit, newdata = predict_data)
```

*Output:*

```
       1
1447.303
```

# Predict

*By R:*

The general form of ‘predict‘ command:

predict(linear_model, newdata = data_frame)

*Code:*

```
linear_fit <- lm(Mortality ~ Calcium, data = mortality_water)
predict_data <- data.frame( Calcium = 71)
predict(linear_fit, newdata = predict_data)
```

*Output:*

```
       1
1447.303
```

The outputs by hand and by R are different because of rounding errors.

1. Announcement

2. Today: How useful is your linear model
    1. Finding the least square line in R
    2. How useful is the model?

3. Summary

# Sales/Ads Example

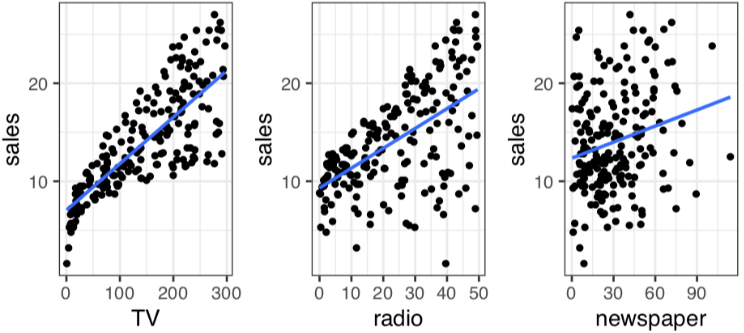
We have a data set of 200 markets, and we are interested in the relationship between sales and advertising budget. We look at the following variables.

- **sales** is a measure of sales volume in thousands of units
- **TV** is TV advertising budget
- **radio** is radio advertising budget
- **newspaper** is newspaper advertising budget

Poll question

Which of these models would you prefer to use for predicting sales?

- a TV
- b radio
- c newspaper



## Tutorial exercise: 15 minutes

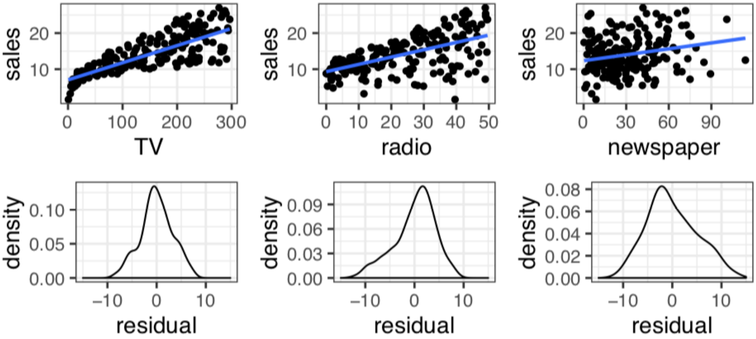
Being as specific and concrete as possible, write down a rule for selecting your preferred model

1. based only on **visual characteristics** of the plot. (That is, your rule should not involve any calculations of numeric quantities).
2. based only on **a quantitative summary** of the data. You can describe how you would calculate your numeric summary of the data in a general sense; if you'd like you can write down a formula.

Each group will nominate a spokesperson for a class-wide discussion.

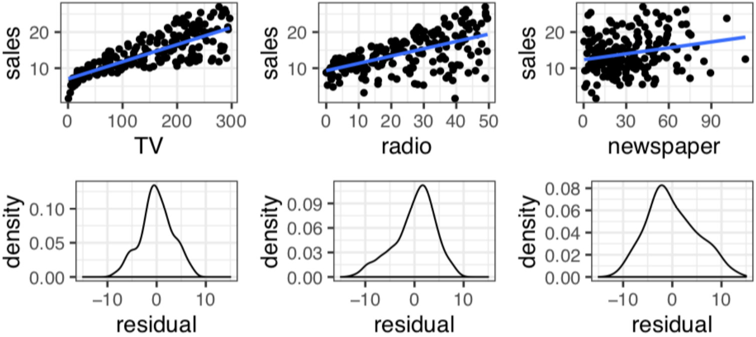
Residuals:

- $e_i = y_i - \hat{y}_i$ (vertical distance between point and line)
- Smaller residuals mean the predictions were better.
- The key is to measure the spread of residuals.



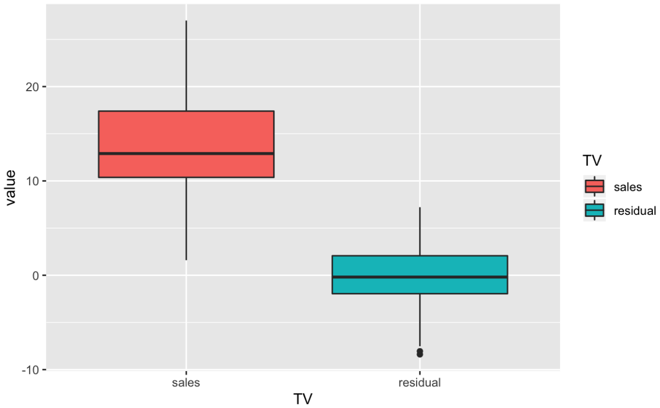
Measure spread of residuals with the standard deviation. We call this the **residual standard error**, $s_{\text{RES}}$.

- TV: 3.26
- radio: 4.28
- newspaper: 5.09

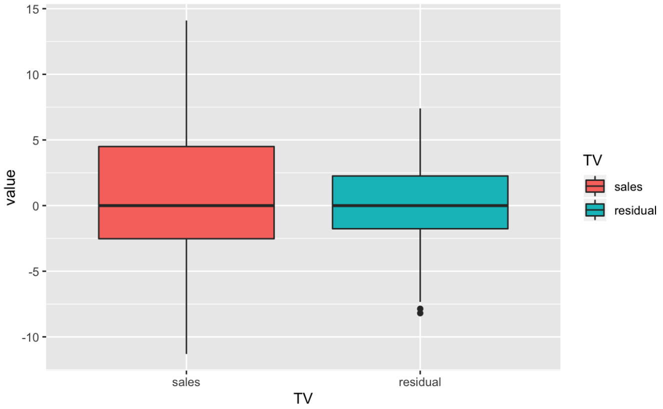


# Variation accounted by the model

The variability in the residuals describes how much variation remains after using the model



The variability in the residuals describes how much variation remains after using the model



# Variation accounted by the model

Let's compute the reduction in variation.

$$\frac{s^2_{\textsf{sales}} - s^2_{\textsf{RES}}}{s^2_{\textsf{sales}}} = 0.61$$

This number describes the amount of variation in the $y$-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in sales can be accounted for by the TV advertisement budget.

# Variation accounted by the model

Let's compute the reduction in variation.

$$\frac{s^2_{\text{sales}} - s^2_{\text{RES}}}{s^2_{\text{sales}}} = 0.61$$

This number describes the amount of variation in the $y$-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in sales can be accounted for by the TV advertisement budget.

How do we get 61%?

Let's compute the reduction in variation.

$$\frac{s^2_{\text{sales}} - s^2_{\text{RES}}}{s^2_{\text{sales}}} = 0.61$$

This number describes the amount of variation in the $y$-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in sales can be accounted for by the TV advertisement budget.

How do we get 61%?

Statisticians found this value is $R^2$, the **square of Correlation**.

# $R^2$

Square of the correlation coefficient $R$: between 0 and 1, closer to 1 is better.

$R^2$ describes the amount of variation in the $y$-variable that is explained by the least squares line.

- TV: 0.61
- radio: 0.33
- newspaper: 0.05

meaning, 61% of the variation in sales can be accounted for by the TV advertisement budget; 33% of the variation in sales can be accounted for by the radio advertisement budget; 5% of the variation in sales can be accounted for by the newspaper advertisement budget.

# Outline

1. Announcement

2. Today: How useful is your linear model
    1. Finding the least square line in R
    2. How useful is the model?

3. Summary

# Summary of main ideas

1. Finding the least square line in R
2. How useful is the model?

Interpretation is the most important thing in this module.

- slope
- intercept
- residual
- $R^2$

# Work in groups

### Tutorial exercise: For the rest of class

Finish Exercise 09

Goals:

1. Practice finding the least square line in R
2. Practice interpretation of a linear model

Submit on Moodle. Each group only needs to submit once.