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 at 12:45pm.
4. I will record in the cloud.

Unit 3: Basic regression

1. Introduction to linear model

Stat 140 - 02

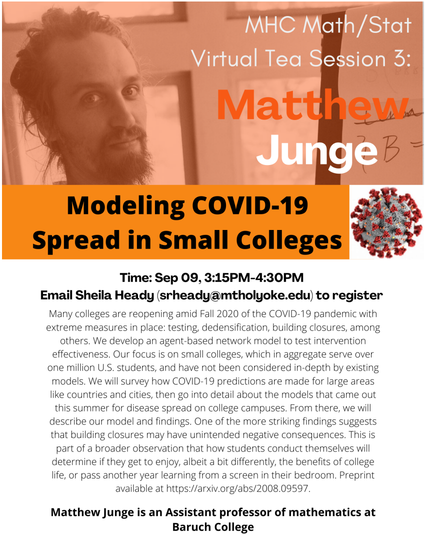
Mount Holyoke College

Dr. Shan Shan

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

1. Announcements
2. This week: Linear models
3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict
4. 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

- EA 01:
  - We (TAs and I) are still grading
  - Each person grades one problem to minimize bias
  - Grades will be posted later this week
- EA 02:
  - Due Friday 09/11 6PM EST
  - Peer Collaboration Evaluation: https://forms.gle/HPXHtPru981vAWE16
- EA 03: again a group project.
  - Please post your team information on Piazza by Friday.

In support of the strike, I want to

- Acknowledge the importance of work that we all need to do to be anti-racist in society, at MHC, and in our department. Specifically, as individuals and a department we are discussing the anti-racism plan that President Stephens laid out for the college and are engaged in ongoing work.
- Reaffirm my commitment to being open to listening to BIPOC students to understand your experiences and better be able to support you.

Learn more...

I encourage you to follow the following links if you want to educate yourself

- try the common read for MHC this year, the 1619 project
- read about the action plan of the American Mathematical Society in support of the Black community
- learn about how algorithms can reinforce and exacerbate biases in this article in Nature: `https://www.nature.com/articles/d41586-018-05469-3`

Outline

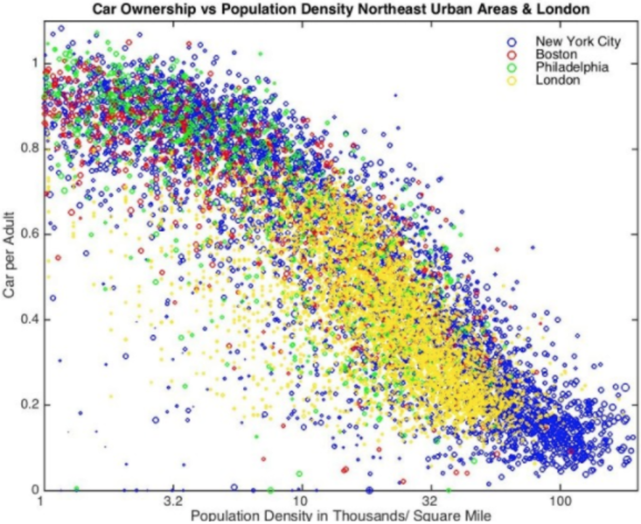
1. Announcements

2. This week: Linear models

3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict

4. Summary

Jamie's finding

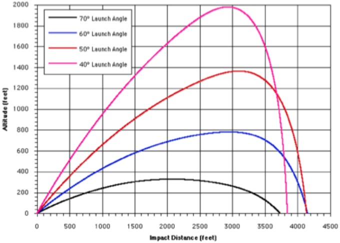


Lines

The algebraic equation for a line is

$$Y = b_0 + b_1 X$$

The use of coordinate axes to show functional relationships was invented by Rene Descartes (1596-1650). He was an artillery officer, and probably got the idea from pictures that showed the trajectories of cannonballs.



1. Announcements

2. This week: Linear models

3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict

4. Summary

1. Announcements

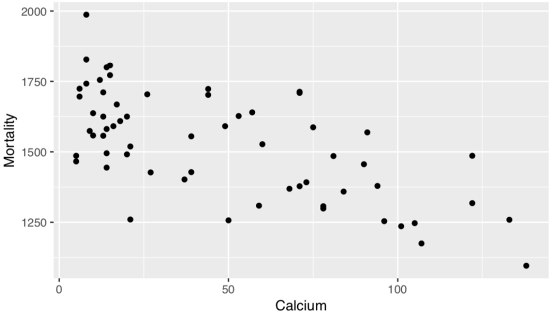
2. This week: Linear models

3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict

4. Summary

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

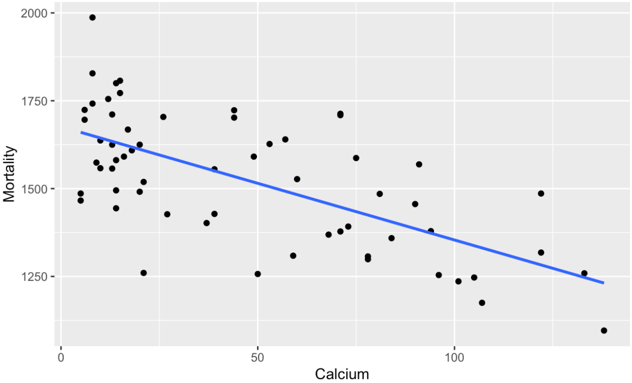
We have recordings of the mortality rate (deaths per 100,000 population) and concentration of calcium in drinking water (parts per million) in 61 large towns in England and Wales.



Review of terms

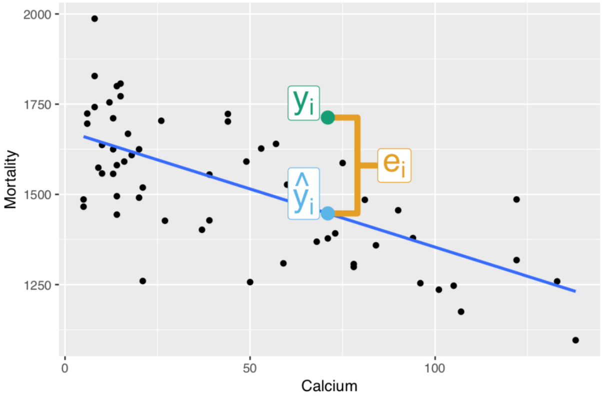
- **Response variable**: variable whose behavior or variation you are trying to understand, on the y-axis (dependent variable)
- **Explanatory variables**: other variables that you want to use to explain the variation in the response, on the x-axis (independent variables); these are also referred to as predictors or features.

Add a line

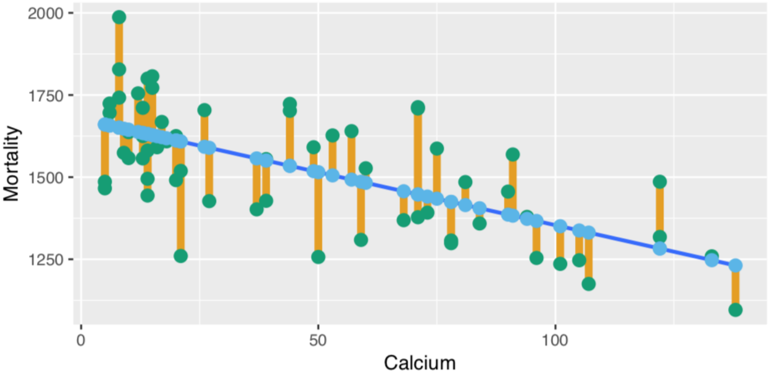


$$\text{Residual} = \text{Observed - Predicted}$$

1. **Predicted value** $\hat{y}$: estimate made from a model
2. **Observed value** y: value in the dataset



The line of best fit is the line for which the sum of the squared residuals is smallest, the **least squares line**.



Outline

1. Announcements

2. This week: Linear models

3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict

4. Summary

Notation

- x: the explanatory variable (calcium concentration)
- y: the response variable (mortality rate)
- x_i: value of x-variable for observational unit number i
- y_i: value of y-variable for observational unit number i
- $\bar{x}, \bar{y}$: sample mean of x and y
- s_x, s_y: sample standard deviation of x and y
- R: correlation between x and y

Finding the least square line by hand

The least square line has

- slope:

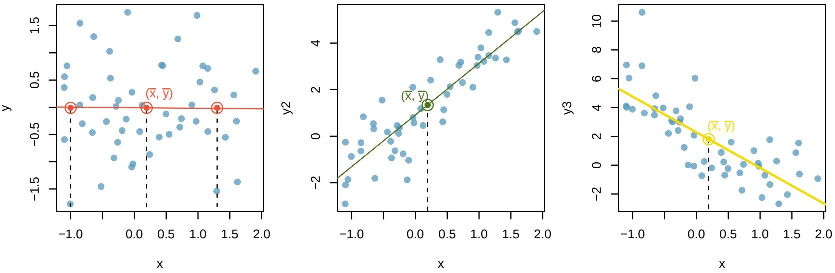
$$b_1 = \frac{s_y}{s_x} R$$

- intercept (the value at $x = 0$):

$$b_0 = \bar{y} - b_1 \bar{x}$$

The calculation of the intercept uses the fact the a least square line always passes through $(\bar{x}, \bar{y})$.

Why does the least square line **always** pass through $(\bar{x}, \bar{y})$?



Mortality/Health Example

Suppose we have the following information.

	mortality rate	calcium concentration
	(y)	(x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

Mortality/Health Example

Suppose we have the following information.

	mortality rate	calcium concentration
	(y)	(x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

1. Calculate the slope.

$$b_1 = R \times \frac{s_y}{s_x} = -0.65 \times \frac{188}{38} = -3.21$$

Mortality/Health Example

Suppose we have the following information.

	mortality rate	calcium concentration
	(y)	(x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

1. Calculate the slope.
$$b_1 = R \times \frac{s_y}{s_x} = -0.65 \times \frac{188}{38} = -3.21$$
2. Calculate the intercept.
$$b_0 = \bar{y} - b_1 \times \bar{x} = 1524 + 3.21 \times 47 = 1674.87$$

Suppose we have the following information.

	mortality rate (y)	calcium concentration (x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

1. Calculate the slope.

$$b_1 = R \times \frac{s_y}{s_x} = -0.65 \times \frac{188}{38} = -3.21$$

2. Calculate the intercept.

$$b_0 = \bar{y} - b_1 \times \bar{x} = 1524 + 3.21 \times 47 = 1674.87$$

3. Write out the linear model.

$$\widehat{\text{Mortality}} = 1675 - 3 \text{ Calcium}$$

Slope and intercept

In general, the regression line is

$$\hat{y} = b_0 - b_1 x$$

1. **Slope** b_1: Slopes are always expressed in y-units per x-unit. They tell how the y-variable changes (in its units) for a one-unit change in the x-variable.
2. **Intercept** b_0: the value the line takes when x is zero

How to interpret intercept and slope in the context of data?

Outline

1. Announcements

2. This week: Linear models

3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict

4. Summary

What is the interpretation of slope?

Units:

- x-variable: calcium concentration (parts per million)
- y-variable: mortality rate (deaths per 100,000 population)

The slope, -**3**, says that for 1 unit increase in x-variable, we can expect, on average, to have 3 units less in y-variable.

Units:

- x-variable: calcium concentration (parts per million)
- y-variable: mortality rate (deaths per 100,000 population)

The slope, -**3**, says that for 1 unit increase in x-variable, we can expect, on average, to have 3 units less in y-variable.

This means, for 1 part per million increase in calcium concentration, we can expect, on average, to have 3 deaths per 100,000 population less in mortality rate.

Units:

- x-variable: calcium concentration (parts per million)
- y-variable: mortality rate (deaths per 100,000 population)

The slope, -**3**, says that for 1 unit increase in x-variable, we can expect, on average, to have 3 units less in y-variable.

This means, for 1 part per million increase in calcium concentration, we can expect, on average, to have 3 deaths per 100,000 population less in mortality rate.

Less formally, for each additional parts per million increase in calcium concentration, the predicted number of mortality rate decreases by 3 deaths per 100,000 population.

What is the interpretation of intercept?

Algebraically, that's the value the line takes when x is zero.

Here, our model predicts that when the water does not have any calcium, on average, the mortality rate is 1676 deaths per 100,000 population.

Note that the intercept serves only as a starting value for our predictions, and we don't interpret it as a meaningful predicted value.

Outline

1. Announcements

2. This week: Linear models

3. Main ideas
    1. Add a line to your plot
    2. Find the least square line
    3. Interpret intercept and slope
    4. Predict

4. Summary

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:

$$\widehat{murder} = 1675 - 3 \times 71 = 1462$$

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

Rounding errors

The calculations in this unit are particularly sensitive to the rounding. If your answers differ from what I provided here, they may still be correct in the sense that you calculated them by a correct method, buy you may have used values that were rounded differently than those used to find these answers.

In general, we use the original data and do no intermediate rounding.

Don't be concerned about minor differences — what is important is your interpretations of the results.

1. Announcements

2. This week: Linear models

3. Main ideas
   1. Add a line to your plot
   2. Find the least square line
   3. Interpret intercept and slope
   4. Predict

4. Summary

Summary of main ideas

1. Add a line to your plot
2. Find the least square line
3. Interpret intercept and slope
4. Predict

Tutorial exercise: For the rest of class

Begin working on exercises sent via zoom chat

Goal:
(1) Practice finding the least square line by hand
(2) Interpretation of the least square line
(3) Practice finding prediction and residuals

Participation:
Write your answers on Piazza. Submission optional.