Lab 13: Exploring Linear Regression

In this class, we've looked at several scenarios where we were able to take a formula, use it to create a table of data, represent those as points on a graph, and connect them with a perfectly straight line. Then we can look at things like the slope of the line and relate them back to our data and to the numbers we see in the formula. (Think about our "total cost of owning a car" examples from the Formulas Class Notes.)

The purpose of this lab, however, is to kind of "go the other way" …. what if you ***have*** the data already, and want to find its equation? And what if that data is ***kinda*** lying on a line, but not ***exactly***?

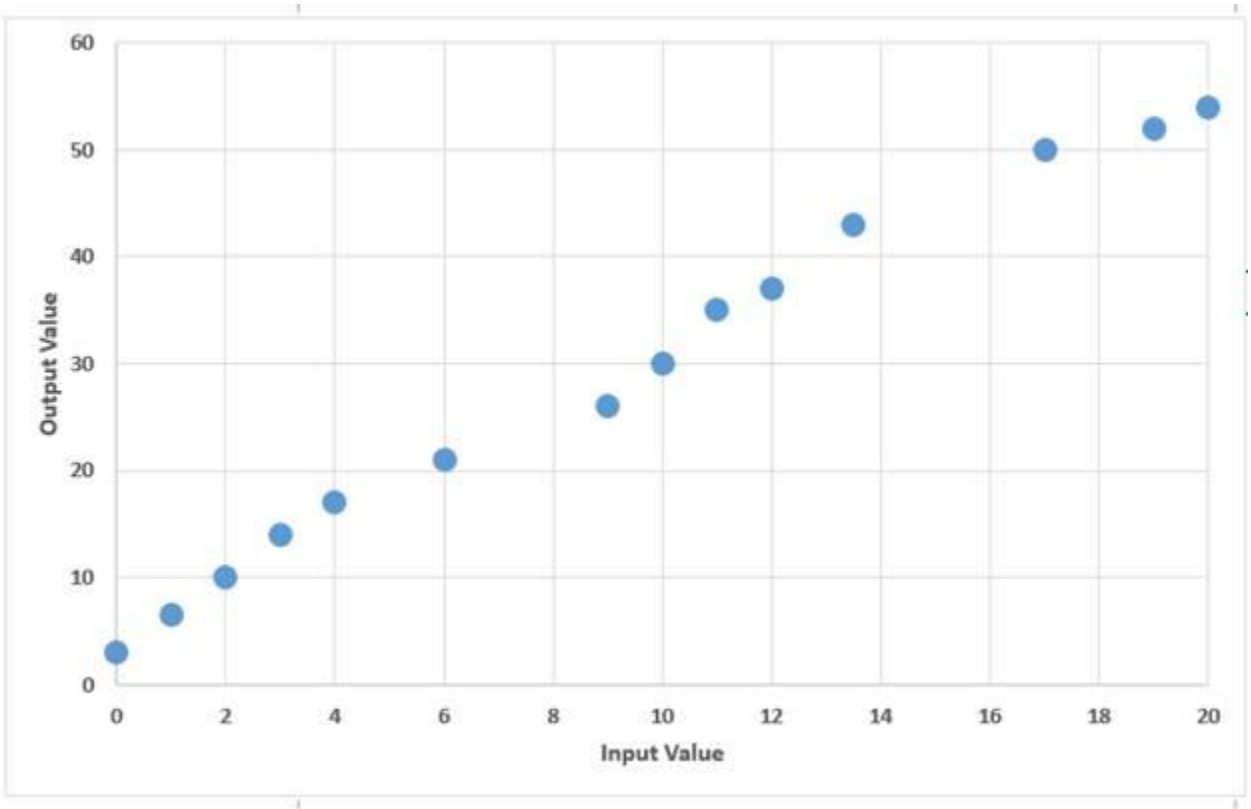
In this lab, we'll do both! You see, in reality, data that you encounter isn't always neat and clean around the edges (you've already seen some of that in this class!). But, for starters (and to get some technoliogy basics down), we'll look at some completely "perfectly" linear data first. Open the spreadsheet for today: Lab 13 Exploring Linear Regression Data, You'll need to click the "Make a copy" button, and then make sure you are on the **"Perfect" Linear Data** tab.

Part 1: "Perfect" Linear Data

Watch this video to learn (or, maybe relearn) a few things! Make sure to work along with me so you can see the commands I use!

Part 2: "Not So Perfect" Linear Data

Now I'll have you take a look (not in Sheets just yet) at some "not so perfect" linear data.



1. **(3 points)** When I say "not so perfect" linear data, what do you think I mean? Write a couple of sentences, specifically: why "not so perfect"? Why "linear"?

2. **(2 points)** Draw what you think the best-fit line would be for that data! Here's a video where I show you how to do this using the Snipping Tool and Microsoft Paint. Please include that screenshot with your best fit line as your answers to this question!

3. **(2 points)** See how we're missing inputs of 5, 7, and 8? Pick one of them and approximate its output using your best-fit line!

4. **(5 points)** (**w**) Technically speaking, we're also "missing" an input value of 100. What would the approximate output be for an input of 100? Also explain how you figured it![1]

Now let's use ***Google Sheets*** to figure the "best-fit line" for this data. Click over to the tab called **"Not So Perfect" Linear Data**! You might need to rewatch the video from Part 1!

5. **(1 point)** What's the best-fit line's slope?

6. **(1 point)** How close was your slope from #2?

7. **(2 points)** Using the best-fit line from Sheets, what's the best output for the input you used back in #3?

Part 3: MTH 105 Data

When you open the next tab (MTH 105 Data), you'll find a data set that lists some MTH 105 student grades. In particular, you'll see each student's average project grade (out of 25 points), average quiz grade (out of 10 points), and overall grade in the course (out of 100 points).

8. **(2 points)** Plot "Project Average" vs. "Course Grade" on one scatterplot, and "Quiz Average" vs. "Course Grade" in a second scatterplot (make sure "Course Grade" is on the vertical axis in each graph). [**Note:** To select two columns that aren't right next to each other, hold down the "**Ctrl**" button while selecting one data set and then the other.] Which of the two data sets is *more* linearly related to "Course Grade"? by this, I mean which of the sets of points more tightly cluster around a linear shape?

9. **(4 points)** Give one reason why you think this might be. Hint: look at the paragraph above!

On your "Project Average" vs. "Course Grade" chart, add a trendline. Make sure to display its equation!

10. **(2 points) (w)** According to this equation, what course grade goes along with a project average of 22?

11. **(1 point)** Find the only student with a project average of exactly 22. Did they do better or worse in the course than the trendline would suggest?

[1] **Word of statistical caution**: when you're building mathematical models on data, it's dangerous to go "too far away" from the data you used to build your model. As an example, consider the 2008 housing bubble that burst spectacularly—if you built a housing price model based on 2000 through 2007 data, it wouldn't have done a great job in 2009! It's generally a good idea to stay "close" to your dataset. Now, if you're wondering, "what's ***close*** mean?"…hang tight. We'll get there. ☺