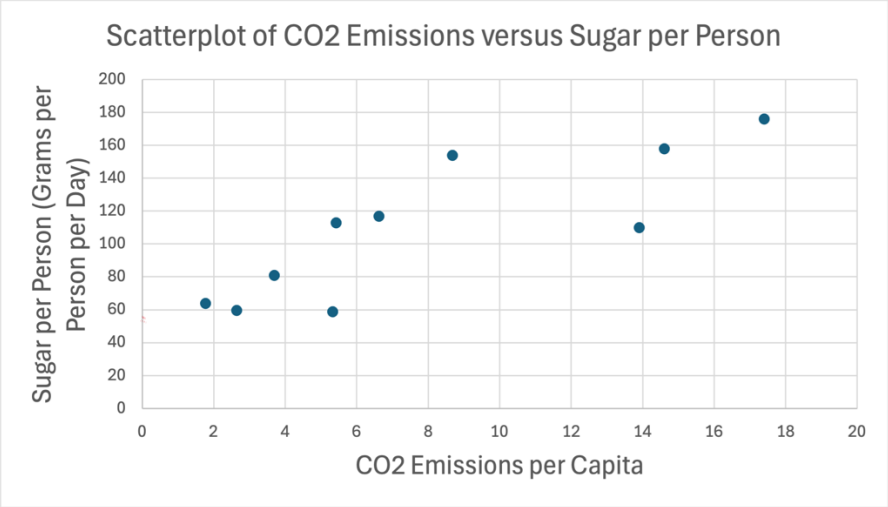
**ACTIVITY 3: Worksheet**

When the relationship is linear, we can model the relationship with a least squares regression line. The least squares regression line is a linear equation that best fits the data. Excel can be used to find an estimate of the slope and y-intercept of a linear equation that best fits a data set.

The relationship below shows CO2 emissions per capita versus sugar per person (grams per person per day).




The slope found using Excel was m = 6.6 and the y-intercept was 56.5. Answer the questions below based on the information given.

**1.** Use the slope and y-intercept above to create an equation for the least squares regression line.

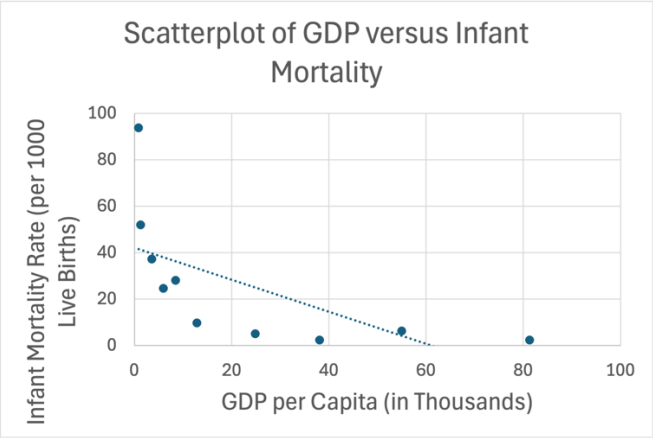
**2.** Interpret the y-intercept in terms of the data above.

**3.** Interpret the slope in terms of the data above.

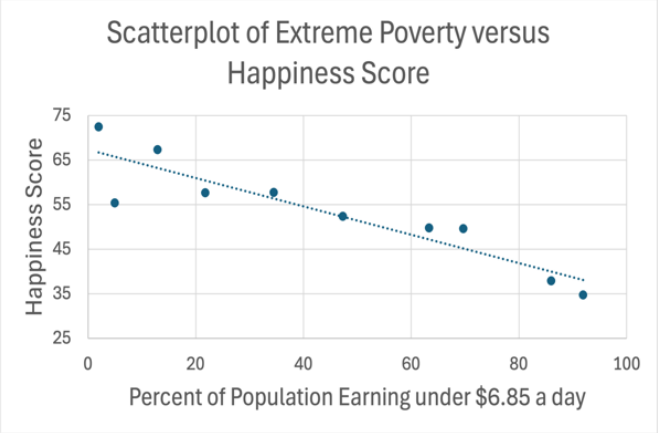
**4.** Roughly draw the least squares regression line on the scatterplot above. Hint: Use the equation to find two points and draw the line.

**5.** For this example, there is a correlation, but can you argue for causation? Explain!

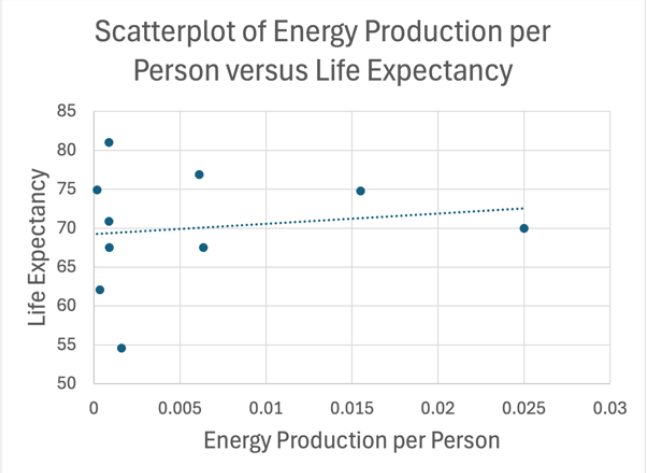
It is not always appropriate to find a least squares regression line for data. Answer the questions below based on the graphics given.




**6.** For the scatterplot on the left, why does it not make sense to find the least squares regression line to use as a model for this relationship?

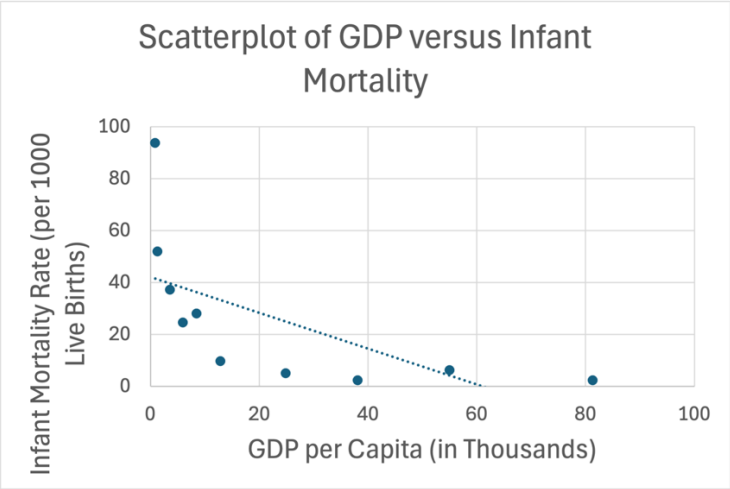


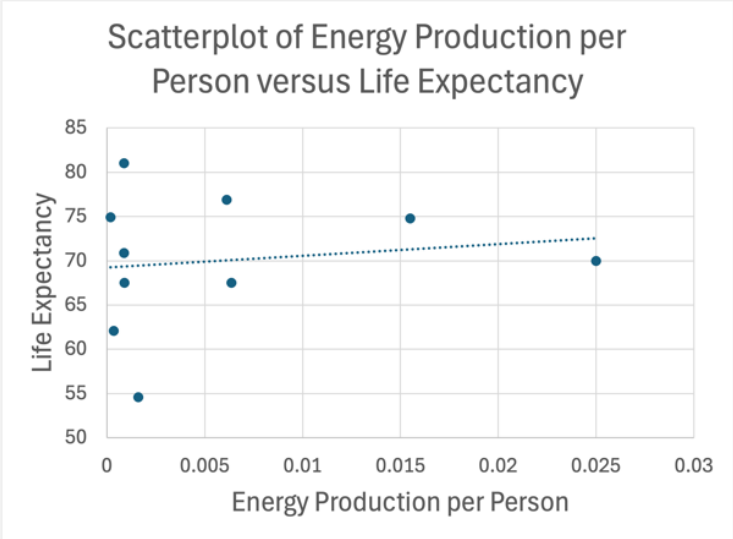

**7.** For the scatterplot on the left, would it be reasonable to find the least squares regression line to model the relationship? Explain!




**8.** For the scatterplot on the left, would it be reasonable to find the least squares regression line to model the relationship? Explain!

**9.** For each of the graphs below, explain why using the least squares regression line to predict the dependent variable makes no sense. Also, explain what a better prediction would be.

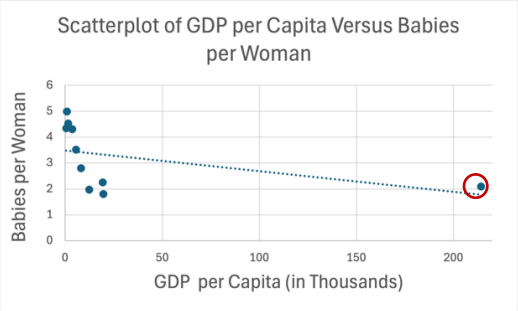


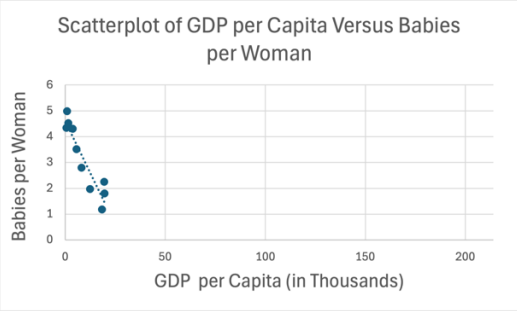



**How do we deal with outliers?**

Remember that outliers in the **x-direction** are also considered influential points. Answer the questions below based on the scatterplots of GDP per capita versus babies per woman.

**Note:** The scatterplot of the left includes an extreme outlier in the x-direction. Monaco has an extremely high GDP because of the low personal tax rate which draws people with considerable wealth. The graph on the right is the same data with Monaco removed.




**10.** Does the least squares regression line on the scatterplot on the left follow most of the points? Why do you think this is?

**11.** Do you think you would get accurate predictions if you were to use the regression line shown on the scatterplot to the left? Explain!

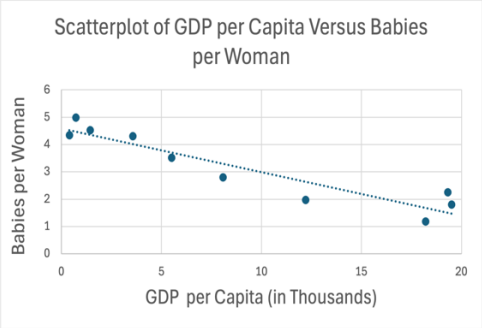
**12.** If you removed the outlier, would you get more accurate predictions? Look at the regression line in the scatterplot on the right. Explain!

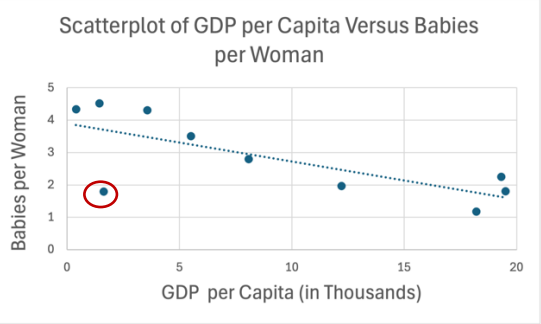
**13.** Do you think that you could use the regression line found on the right to predict the babies per woman for a country with a GDP of 100? Explain!

**14.** From the above, what is the best way to deal with outliers in the x-direction that don't follow the general trend of the data?

Outliers in the y-direction are typically not influential unless they are extreme or there is very little data. Answer the questions below based on the scatterplots of GDP per capita and babies per woman.

**Note:** The scatterplot on the right includes the same data as the scatterplot on the left with one outlier in the y-direction circled in red, added.




**15.** Do you see a drastic change in the least squares regression line? Explain!

**16.** Do you think you will still get good predictions for the regression line calculated with the outlier? Explain!

From the previous questions, when is the only time that a least squares regression line can be used for prediction?

Even when you can find the least squares regression line and use it for prediction, it is important to follow another rule. To discover the third rule, work through the following exercise.

**There is an Excel spreadsheet with the data you will be using for this exercise.**
For the data on Sheet 1, find the scatterplot and the least squares regression line and answer the questions below. Instructions are at the end of this activity and you may start with Step 2 as Step 1 is inputting the data.

**1.** Describe the relationship between average daily income and life expectancy for countries.

**2.** Would it be appropriate to find the least squares regression line? Explain!

**3.** Find the slope and y-intercept for the least squares regression line and write the equation below.

**4.** Paraguay has an average daily income of $17.30, predict the life expectancy for Paraguay using your least squares regression line equation.

**5.** The actual life expectancy for Paraguay is 73.1 years. How accurate was your prediction?

**6.** Is $17.30 within the range of the independent variable?

**7.** Singapore has a daily income of $127, predict the life expectancy for Singapore using your least squares regression line equation.

**8.** The actual life expectancy in Singapore is 83.8 years. How accurate was the prediction?

**9.** Is $127 within the range of the independent variable?

**10.** Go to Sheet 2 which has values above $19.20 and create a scatterplot of the data. Describe the relationship between average daily income and life expectancy for countries.

**The prediction for Singapore required what is called "Extrapolation".**

In linear regression, **extrapolation** involves estimating a value of the dependent variable from a value of the independent variable that is outside the range of the independent variable used to calculate the least squares regression line.

**11.** Do you think from the previous example that extrapolation is advisable? Why or why not?

**12.** Based on all the activities you have completed, what are some rules you should adhere to when computing the least squares regression line and using it to predict the dependent variable from the independent variable?

You are now ready to continue working on your PowerPoint presentation by creating a new slide. On Slide 4, state what would be the best predictor for the relationship of your data and explain why. If your best predictor is the least squares regression line or the average of your dependent variable, calculate it. Also, state the range of x-values that can be used for prediction.

**Excel Instructions:**

1. Type your data into two columns with the independent variable in the left column and the dependent variable in the right column.
2. Highlight both columns of data, click “Insert”, select the graph that looks like a scatterplot, and choose the upper left option.
3. **Note:** Typically, you want the points to be spread out across all values of your variables. You may need to adjust the bounds of one (or both) of your axes. To do so, double-click on one of the numbers on the axis you want to adjust. Under the “Format Axis” option, click on the bar graph icon. You’ll then notice that you can adjust the bounds under “Axis Options.”
4. If your data is linearly related, you can put in a trend line by selecting “Add Chart Elements”; followed by “Trend Line”; and then followed by “Linear”.
5. To calculate the slope and y-intercept for the least squares regression line, click on an empty square in Excel, and type in
   “=LINEST(“. Then highlight the second column of data, type a “,” highlight the first column of data, and type a “)”. You can then hit enter. The first number is the slope, and the second number is the y-intercept.
6. **Note:** Your final equation will look like =LINEST(B2:B26,A2:A26).