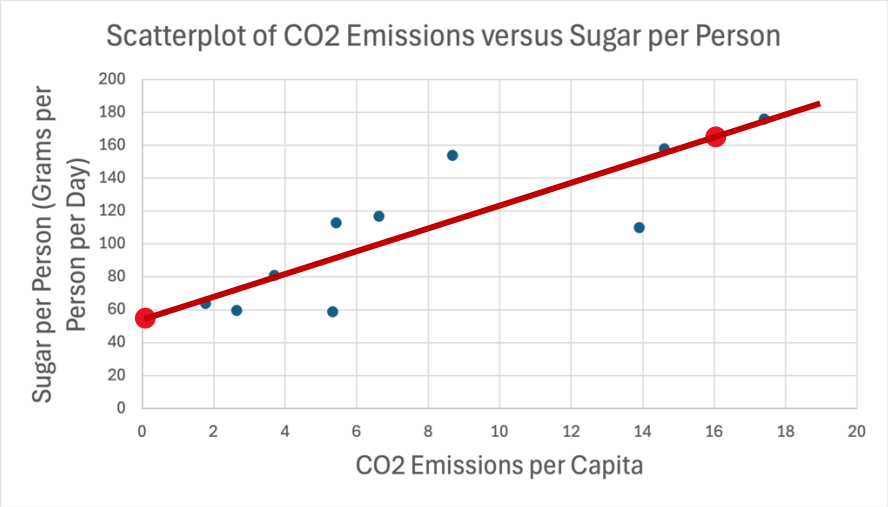
ACTIVITY 3: Worksheet Answers

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.
$\hat{y}$ = 6.6x + 56.5

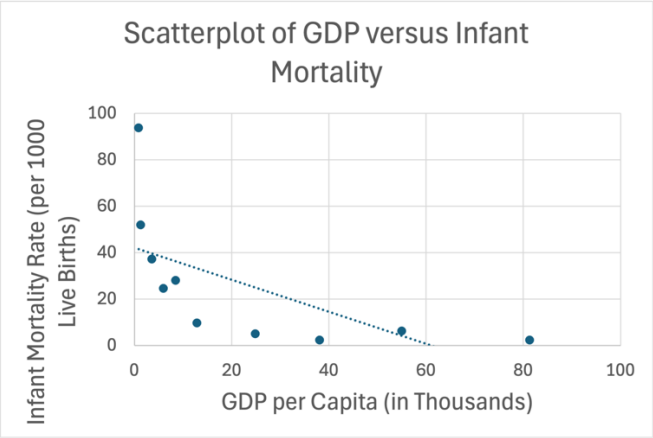
2. Interpret the y-intercept in terms of the data above.
When the CO2 emissions per capita for a country are close to zero, the number of grams of sugar per person per day is close to 56.5.

3. Interpret the slope in terms of the data above.
For every increase of 1 in CO2 emissions per capita for a country, the number of grams of sugar per person per day will increase by 6.6 grams.

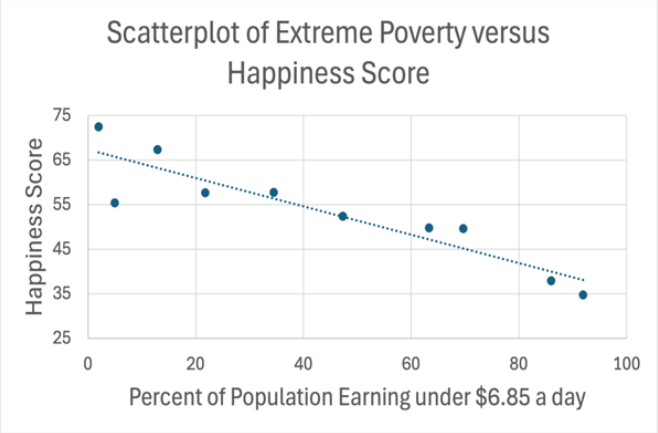
4. Roughly draw the least squares regression line on the scatterplot above. Hint: Use the equation to find two points and draw the line.
When x = 0, y = 6.6(0) + 56.5 = 56.5 and x = 16, y = 6.6(16) + 56.5 = 162.1

5. For this example, there is a correlation, but can you argue for causation? Explain!
It would be very difficult to argue that a higher CO2 emission per capita for a country would cause sugar consumption to go up. Most likely, an outside variable such as the wealth of the country is causing the correlation. Wealthier countries produce more CO2 and have the money to consume more sugar on average.

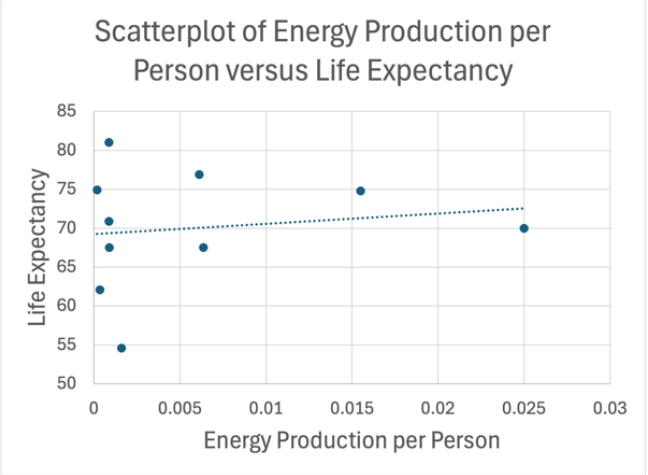
It is not always appropriate to find a least squares regression line for data. Answer the questions below based on the graphics given.



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?
The data has a distinct curve. It would be better to find an equation that follows the curve as a model for this relationship.

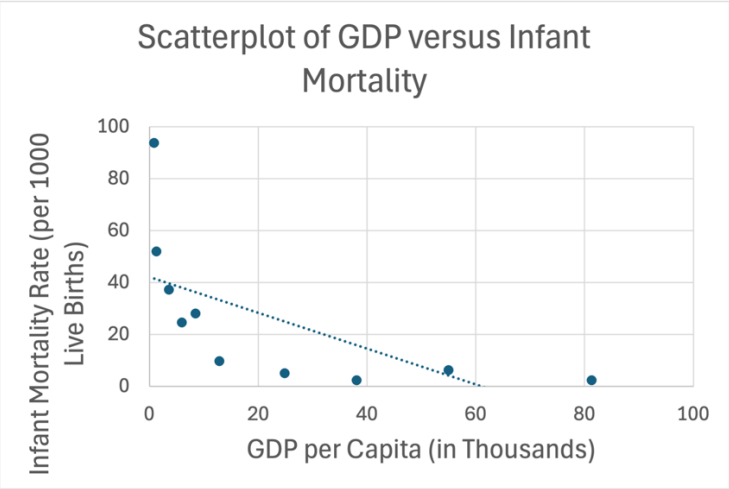


7. For the scatterplot on the left, would it be reasonable to find the least squares regression line to model the relationship? Explain!
Yes, the relationship between these two variables is approximately linear, so it would be reasonable to use the least squares regression line to model the relationship. More specifically, the happiness score decreases at a near-constant rate of change as the percentage of the population earning under $6.85 a day increases.

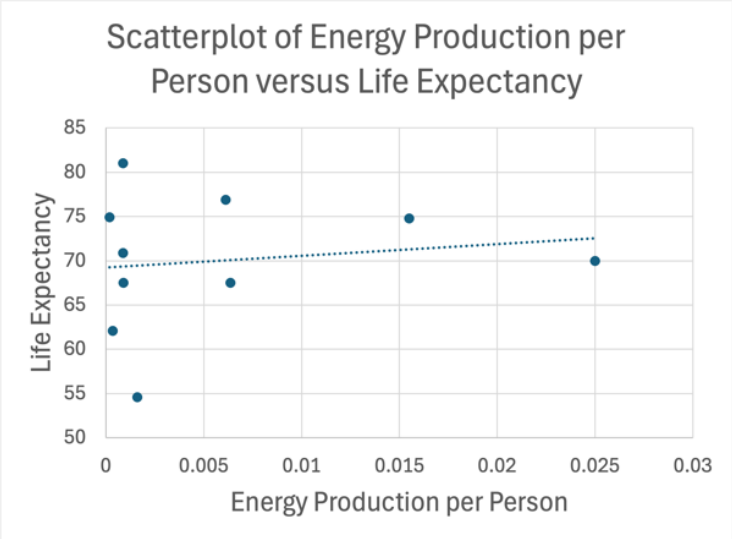


8. For the scatterplot on the left, would it be reasonable to find the least squares regression line to model the relationship? Explain!
No, the relationship does not appear to be linear.

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.



Based on the scatterplot to the left, the relationship between GDP and infant mortality rate (per 1000 live births) does not appear to be linear. Consequently, it would be better to find an equation that fits the curve to use for prediction.

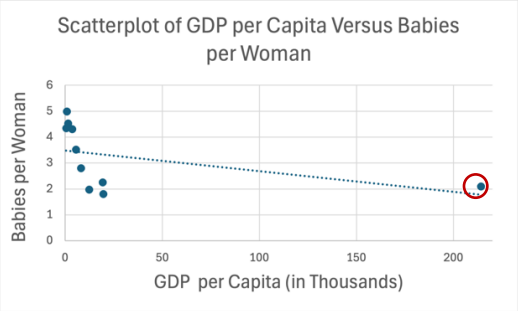


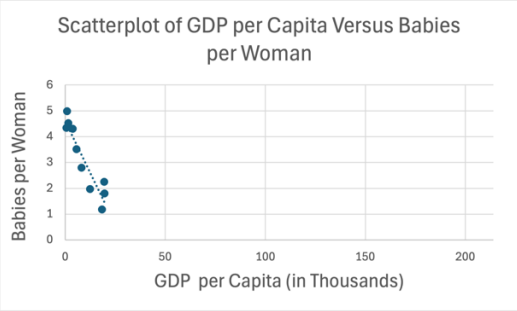
Based on the scatterplot to the left, no relationship was found between energy production per person and life expectancy. Averaging the y-values would be the best prediction in this case.

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?
The regression line does not follow the points. It looks like the outlier in the x-direction has a strong pull on the regression line.

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!
No, the predictions would not be accurate because the regression line does not follow the general trend of most of the data.

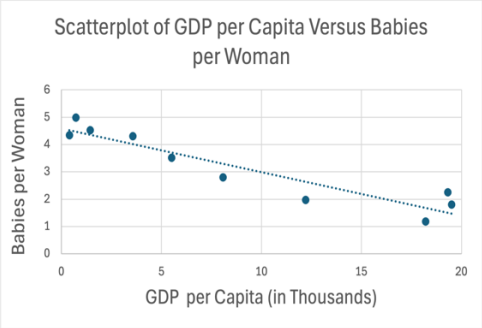
12. If you removed the outlier, would you get more accurate predictions? Look at the regression line in the scatterplot on the right. Explain!
Yes, the regression line now follows the trend of the data so the predictions will be more accurate.

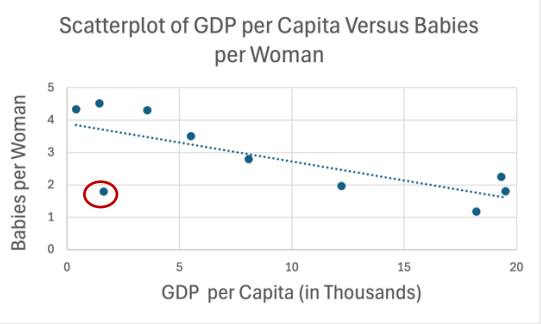
13. Do you think that you could use the regression line found on the right to predict the babies per woman for a country with a GDP of 100? Explain!
No, because looking at the graph, you would get a negative number of babies per woman which would not be possible.

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?
For outliers in the x-direction, it is better to remove them as they can drastically change the equation of the least squares regression line.

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!
No, while the outlier pulls the regression line towards it, the points above the outlier counter the effect that the outlier has thus minimizing it.

16. Do you think you will still get good predictions for the regression line calculated with the outlier? Explain!
Yes, it would not be necessary to remove the outlier before calculating the regression line. The regression line still follows the trend of the data with the outlier included.

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

The least squares regression line can be used for prediction

- when the relationship between the two variables is linear.
- when there are no extreme outliers in the x-direction that don't follow the general trend of the data.

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.
The relationship between average daily income and life expectancy for countries is linear, positive, and moderate with one outlier in the y-direction.

2. Would it be appropriate to find the least squares regression line? Explain!
Yes, since the relationship is linear, it would be appropriate to find the least squares regression line.

3. Find the slope and y-intercept for the least squares regression line and write the equation below.
$\hat{y} = 0.408x + 65.66$

4. Paraguay has an average daily income of $17.30, predict the life expectancy for Paraguay using your least squares regression line equation.
$\hat{y} = 0.408(17.30) + 65.66 \approx 72.72$ years

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

6. Is $17.30 within the range of the independent variable?
Yes, there are x-values below and above it.

7. Singapore has a daily income of $127, predict the life expectancy for Singapore using your least squares regression line equation.
$\hat{y} = 0.408(127) + 65.66 \approx 117.48$ years

8. The actual life expectancy in Singapore is 83.8 years. How accurate was the prediction?
The prediction was very inaccurate. It would be impossible for a country to have a life expectancy of 117.48 years. Very few people live to that age.

9. Is $127 within the range of the independent variable?
No, the highest x-value is $19.20 and $127 is significantly above this value.

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 relationship between average daily income and life expectancy for countries starts linear but then levels off. At some point, the average daily income of a country can no longer increase the longevity of its citizens.

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?
It is not a good idea to extrapolate as it is unlikely that your estimate will be very accurate. Even though the relationship may appear linear for the range of x-values used, it is unknown what the relationship looks like outside of those values.

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?
1 – Always look at a scatterplot first to determine whether the relationship is linearly related.
2 – Remove outliers in the x-direction that are influential.
3 – Avoid extrapolation as it is often very inaccurate.

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.

For the example data collected in Activity 1, the following would be on Slide 4.

Slide 4:
The best predictor would be the regression line $\hat{y} = -0.03x + 4.34$ because the relationship between contraceptive use and babies per woman is linear.

The regression line can be used for prediction for countries that have between 29.3% and 84.5% contraceptive use.

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).