Lab 4: Correlation Coefficient and Linear Regression

|  |  |
| --- | --- |
| Due: | Monday, June 10th by the end of class |
| Assignment Type: | Group (Up to 6) |
| Assignment Title: | Spreadsheets |
| Style: | One problem at a time on Discord – WORK TOGETHER |
| External Sources: | You are allowed notes, books, and searches |
| Description: | We are going to use Excel or Google Sheets to calculate correlation coefficient and linear regression. If you are using Google sheets and are unsure how to calculate these – here is a website that shows you: <https://www.wired.com/2011/01/linear-regression-in-google-docs/>  If you use Google sheets, make sure to export the file as Excel to turn into Blackboard. |
| Points | 50 |
| Starting Files | None needed – data included here |

These problems are a little bit more time consuming each. Please complete one at a time and have them checked before moving on to the next question. There are only 3 questions for this lab.

This lab was introduced by Jeremy Dixon

# Calculating Correlation Coefficient and Linear Regression

1. The file for this lab should be named lab4.xlsx and include tabs for each of the parts.

Your final answer for this question should have two tabs, “1\_correlation\_coefficient” and “1\_linear\_regression”.

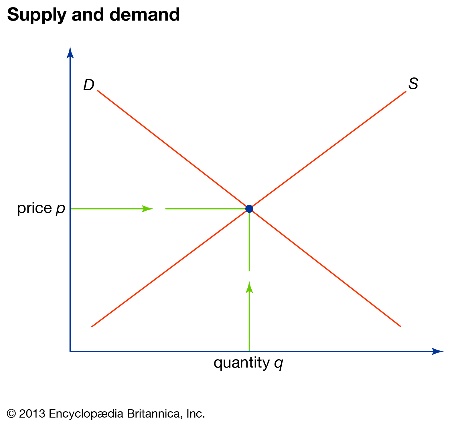


* In a tab named, “1\_correlation\_coefficient”, calculate the correlation coefficient **by hand** in Excel or Google Sheets for the data below. Show each step as in the video. Double check your calculation using the build in calculation.
* In the tab named, “1\_correlation\_coefficient” describe if there is there an expected correlation with this data? What type of relationship?
* In a tab named, “1\_linear\_regression”, calculate the slope of the regression line for the data.
* In the tab named, “1\_linear\_regression”, create a scatter plot of the data and insert the trendline.
* In the tab named, “1\_linear\_regression”, use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

# Calculating Correlation Coefficient and Linear Regression

1. Add new tabs to the lab4.xlsx named “2\_correlation\_coefficient” and “2\_linear\_regression” for this question.

Supply and demand, in economics, is the relationship between the quantity of a commodity that producers wish to sell at various prices and the quantity that consumers wish to buy. It is the main model of price determination used in economic theory. The price of a commodity is determined by the interaction of supply and demand in a market. The resulting price is referred to as the equilibrium price and represents an agreement between producers and consumers of the good. In equilibrium the quantity of a good supplied by producers equals the quantity demanded by consumers. The graph below shows the relationship between supply and demand.



This problem includes two tables: price-supply data and price-demand data for soybeans.



* Enter the data into the “2\_correlation\_coefficient” tab.
* Create the scatter plots for the price-supply, where x is the supply (in billions of bushels) and y is the price (in dollars). Do the same for price-demand.
* Adjust the minimum and maximum of the axes of each plot to slightly below and slightly above the data values.
* Compute the regression equations for supply and for demand using linear regression on each of the plots. The trendline will be y = ax + b for some values of a and b. Round a and b to 3 decimal places.
* Use the trendlines to find the equilibrium price for soybeans. (Hint: The supply model will be an increasing linear function. The price model will be a decreasing 3 linear function. Set the two equations equal to each other and solve for the equilibrium value x, and then find the corresponding value for y equilibrium price.)

1. Add new tabs to the lab4.xlsx named “3\_correlation\_coefficient” and “3\_linear\_regression”

For this problem, we are looking at the number of internet hosts by year. Based on this, we can extrapolate what we think should be occurring in 2019. Then we can check the actual number to see how close our estimate was.



* + Enter this data into the 3\_correlation\_coefficient and calculate the correlation coefficient by hand (use built-in functions to check)
  + Calculate the linear regression by hand
  + Extrapolate for the year 2018 and check the actual data here: <https://www.statista.com/statistics/264473/number-of-internet-hosts-in-the-domain-name-system/>
  + Did the data follow the expected pattern? Why or why not?
  + If you add the actual data from 2018, does the slope change? Is that important? What happens if we were projecting out to 2021 or 2024? How does that slope change effect the results?