**Classwork 6**

MAT 135

**Correlation and Linear Regression**

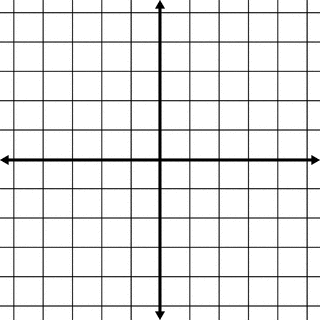
Instructions: Take out your TI calculator. For the following classwork, enter the following datasets into the L1 and L2 lists

*L1 = { 1930, 1940, 1950, 1960, 1970, 1980, 1990, 2000, 2010 }*

*L2 = { 92, 103, 300, 543, 1076, 2863, 5980, 10285, 14964 }*

L1 represents the year and L2 represents the US Gross Domestic Product in billions of dollars. You will be using this *bivariate* dataset in the following classwork.

Part 1: Scatter Plot

**Problem**

1. Create a scatter plot of the datasets, using the year as the independent variable and the US GDP as the dependent variable. If you need a refresher on how to create a scatter plot on your calculator, open up *Classwork #3* and review the procedure for generating scatter plots. Sketch the output on the provided graph. Label and scale your axes appropriately:

a. Does there appear to be a negative or positive correlation?

Part 2: Sample Correlation

Tutorial #1: Calculating the Sample Correlation

While visual inspection is a good indicator whether or not correlation exists, we need numerical evidence to conclude the correlation is actually statistically significant. Therefore, we will calculate the sample correlation using the 2-variable statistics provided to us by our calculators and then compare it against a *critical value* to decide whether or not the correlation is significant.

First things first, calculate the 2-variable statistics for this sample of data using the **2-Var Stats** program. Navigate to the STATS menu and then the CALC submenu

*STAT > CALC > 2: 2-Var Stats*

This will bring up the program menu. For the *Xlist:,* select the year. For the *Ylist:* select the US GDP. Ignore the *FreqList:* (*this would be used to specify the frequency of each observation; the frequencies would be contained in a separate list, say, L3.)*

Execute this program to view the 2-variable statistics for this sample of data.

**Problems**

2. Fill out the following table with the **2-Var Stats** output (only fill in the indicated statistics; we do not need some of the results to calculate the sample correlation.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| **n** |  |

3. Recall the short cut formula for sample correlation is given by

Using the 2-variable summary statistics calculated in #2, find the sample correlation.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Open up D2L and navigate to *Course Materials > Content > Resource > Statistical Tables* and open the Pearson Correlation Coefficient Table. Using a significance level of 5%, find the critical value for significant correlation from a sample of *n* data points.

Critical Value : \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Compare the calculated correlation t to the critical value for correlation. Is there evidence to suggest the correlation observed in the scatter plot and calculated from sample statistics is statistically significant?

Part 3: Linear Regression

Once the correlation in a bivariate dataset has been established as significant, we can use the correlation to extrapolate and make predictions about future values of the dependent variable. We fit a line to the data, known as the *line of best fit*. The line of best fit is found by minimizing the prediction error, called the *mean square error* or MSE for short

Where *y* is the *observed* (or actual) value of the dependent variable and is the *predicted value* of the dependent variable, according to the line of best fit. The *linear regression model,* or the equation of the *line of best fit*, is given by

(Linear Regression Model)

The TI calculator will fit this model to the datasets entered into L1 and L2 if we use the **LinReg** program, which can be found in the same location as the **2-Var Stats**

STAT > CALC > 4 : LinReg (ax + b)

The program screen for **LinReg** looks very similar to the **2-Var Stats** program screen. In fact, we essentially do the same thing: Enter the year into the *Xlist:* and the US GDP into the *Ylist:.* You can ignore all of the other options. Execute the program and you should see the equation of the line of best fit.

You can superimpose the equation of the line of best fit over the scatter plot itself by navigating to the PLOT menu, typing in the equation and then switching to the GRAPH screen. Just make sure to use an appropriate viewing window!

**Problems**

6. Write out the linear model with the values of and calculated from the **LinReg** program:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7. Sketch the line of best fit (*linear regression model*) on top of the scatter plot you created in #1. Use your calculator's GRAPH function to make sure the sketch is as accurate as possible.

8. If the trend between the year and US GDP continues into the future, predict the US GDP in the following years:

a. 2020

b. 2030

c. 2100