Patrick Jarvis

Professor Little

Mont. Identifying Patterns

18 November 2009

Lab on Correlation, Regression, and Data Analysis Report

A). When analyzing the data for the various populations many conclusions can be drawn. By calculating the correlation coefficient and also the residuals for the regression it becomes apparent that there is not a linear relationship between x and y. This is a result of the residual graphs which have a definite pattern. If the residual plots didn’t have a then it would be considered a linear relationship. The lack of a linear relationship can also be surmised by observing the graph which appears to resemble the graph of y= (1/x) with a domain of x>0. In this question however the correlation coefficient wouldn’t be helpful overall because it has a resulting value close to negative one, in this case, the correlation only reflects the general uniformity of the data points. The correlation coefficient would however be helpful in determining the general pattern of the data, which is downward sloping. Similar conclusions can be drawn when analyzing the data for the relationship between ln (y) and x. By observing the plot of the residuals you can see a definite pattern meaning there is no linear relationship. Much like the relationship between x and y, the lny and x relationship has a negative correlation close to one but once again that is only a reflection of the general uniformity of the data points.

The relationship of ln x and ln y however varies greatly from the other two relationships. By observing the graph of the residuals it becomes apparent that there is a relationship between lny and lnx. There is a relationship because there is no pattern in the residual plot signifying a linear relationship. This relationship also has the strongest correlation coefficient at approximately -.985.

If the first eight populations were removed the resulting relationships would become more linear. This is due to how unrepresentative the first eight points are, they would be deemed outliers. This is because they throw the data of so much, especially the first three which are more than 3 times the average. This skew that is created is best illustrated by the new average and standard deviation that excludes the first eight points, the average decreases by 700,000. The standard deviation decreases by over 160,000 which is more than a 50% decrease. The resulting residuals plots also reflect the increase in the linear relationship, because they become more disorganized and lack a definite pattern compared to their respective counterparts.

The residual plots appear to have a linear relationship but through calculations it becomes apparent that a functional relation would best represent the data. When calculated separately using the regression values the approximate relationship would best be expressed as y= x^ -1.33 + e^19.5. This would result in a curved line that mimics the original data plot.

If you were to look at the populations levels for the other countries you would expect to see similar results. In every country you will find major cities that house much of the countries populations. These major cities would best be represented by the outliers removed in question four. The majority of the cities however have similar populations that slowly decrease after the outliers. The results given in the lab are based on true values and would be expected to hold true for other areas.

The conclusions drawn from this experiment appear to be accurate. You wouldn’t expect the data follow a perfectly linear model explaining why the best model to use would be a functional model. The outliers which drastically alter the data are also extremely plausible and are comparable to major urban cities like New York and Boston. The decrease is also expected because they are ranked highest to lowest. This means that no matter what the data will be downward sloping because a lower population should result in a lower rank and vice versa.

B). In the second data set there are two different x variables, the Oxide Thickness and the Deposition time. In order to accurately determine the uniformity you need to do calculate the standard deviation of the wafer thickness across all thirteen locations. The standard deviation is then used as your y value because it represents the spread of the data and the variance of the wafer thickness across all thirteen locations. This is what you would use to determine the uniformity of the polysilicon thickness. To determine if a value would be an outlier you would again have to calculate the standard deviation for all locations excluding the value in question. Within this data location thirteen is considered an outlier. To determine this you compare the standard deviation value for all thirteen to the standard deviation value for the first twelve. When compared the twelve has a lower standard deviation on for every different wafer. This lower standard deviation proves that location thirteen is an outlier because it skews the data and makes it less uniform on a whole.

To account for the two x values when calculating the regression line you would have to use a multiple regression data analysis tool. The resulting regression line includes both of the x values while producing the normal regression data and residual plot. By doing the regression sequence used in data set one, i.e. calculating the regression for the pairs (x,y); (x,lny); and (lnx,lny), it is possible to see whether a linear relationship occurs with the data. In general each residual plot lacks a strong definite pattern resulting in a good fit for the data. The similarity of the three residual graphs in terms of a good fit can be seen by the fact that the correlation coefficient for each different point is very similar. For example whenever you calculate the correlation for X1 you get a negative correlation around .8, while when calculating X2 you get a correlation of around .48. The combination of the similar correlation and residual plots that share no distinct pattern reflects the good fit associated with the data set.

The signs of the coefficients reflect the role that each x variable plays in the resulting thickness of the polysilicon wafer. For example for all three relationships have a X1 coefficient as a small negative number. This means that the X1 variable which is oxide thickness has a small effect on the wafer thickness. The effect it does has though is a negative one, the oxide thickness will play some role in the shrinking of the wafer at different locations. The X2 variable has a positive coefficient reflecting that deposition time will play some role in the wafers thickening.