Andrew Coury

Professor Little

Montserrat: Identifying Patters

November 20, 2009

Lab on Correlation, Regression, Data Analysis

A: Population

Analysis of the data in part A lends itself to many possible statistical inferences concerning numerous populations. There is no linear relationship evident between x and y, rank and population, because there is a distinct pattern present on the residual plot. This conclusion is also evident when looking at the overall shape of the graph, which resembles the first quadrant of the function Y= 1/X. The proximity of the correlation coefficient to negative one demonstrates the general uniformity of the data pertaining to the population and rank. This strong negative correlation indicates that there is a downward sloping trend in the data. There is also no linear relationship between ln(y) and x, because the residual has a clear pattern. As in the aforementioned relationship between x and y, ln(y) and x show a negative correlation that is close in magnitude to one, demonstrating the interrelatedness of the data values.

Unlike the relationships between x and y and ln(y) and x, the relationship between ln(x) and ln(y) has no definitive pattern to the residual plot. Therefore, there would be an indication of a linear relationship, as depicted in the residuals graph. The correlation coefficient, r= -.985, is closer in magnitude to one, and is the strongest correlation coefficient.

If populations 1 through 8 were taken out of the data set, there would be a more linear relationship; these first 8 populations are considered outliers since they are the most extreme cities. The other cities that would remain after this change to the data are relatively close to each other. To quantify the disparity between these points and the others, the first three points are triple the average. With the removal of the first eight points, from the data set there is a drastic change in the standard deviation and average. The Standard Deviation is decreased by a quantity of more than 160,000 while the average decreases by a significant 700,000. These result in values of approximately half their original magnitude. This effectively results in a stronger linear relationship due to the lack of a pattern in the residual plot of the new data.

While it seems at first as though the residual plots have a linear relationship, the ensuing calculations indicate that a function relation rather than a linear one would illustrate the data correctly. This expression is namely y=x^ -1.33 + e^19.5.

Data set containing information for different populations would most likely yield conclusions. The outliers in the form of densely populated cities are present in every country, similar to the first eight populations from question four. The population dynamic of the remaining cities, however, would be quite similar resulting in a significant decrease after the removal of these extreme population outliers. These lab results are representative of real-life trends, and are applicable in similar circumstances.

For the most part, the conclusions made in this experiment are indeed accurate ones. While it would be unlikely for the data to be exactly linear it seems much more practical that with the functional model that is used. It is makes sense to see such extreme outliers representing large populations in cities. For this reason it also understandable to witness a decrease since they are placed in order of their population rank in descending order from highest to lowest population. As a result there is a negative correlation and a downward slope since the greater populations have higher ranks while the smaller populations have lower ranks.

B: Oxide thickness

Data set number 2 contains two x variables, namely thickness of oxide and deposition time. The standard deviation is necessary in order to gauge the uniformity of the polysilicon thickness across the sites of the wafers. These values of thirteen wafer thickness are shown in the attached excel sheet. The standard deviation is the y-value, since it demonstrates the variability of wafer thickness at different locations. You can find out whether or not a value is one of the outliers, you calculate the standard deviations for all of the locations other than the one found in the question. For our data, the outlier is location number thirteen. This is found by comparing the standard deviation for the all of the thirteen locations, and then for just the first twelve locations. The standard deviation for comparison between the first twelve points and all thirteen points reveals that the thirteen have a higher standard deviation on each of the wafers. As a result of the skew that occurs, location thirteen is the outlier in the data set.

 In order to calculate the regression line with respect to two x values, it is necessary to use the multiple regression data analysis tool on excel. This regression line is still able to yield the normal residual plot and regression data. If you want to determine whether or not a linear relationship exists in the data of the residual plot you must use the regression methods for calculating the pairs (x,y),(x,ln(y)), and (ln(x), (ln(y)). The similar correlation coefficients as well as the residual plots without a pattern demonstrate a good fit. Overall, the residual plots all lack decisive patterns making them a good fit for the data. The correlation coefficients of the points are very close showing a good fit. For instance, X1 has a strong negative correlation of approximately .8 while the correlation of X2 is about .48.

 The coefficient signs demonstrate the overall trend of the data. The X coefficient which is negative demonstrates that the x variable increases as y decreases. The magnitude of the number represents the amount of the impact made, with a small number not having as big of an effect as larger ones. The X1 coefficient, a small negative number does not greatly influence the thickness of the wafer. The X2 coefficient with a positive coefficient plays the larger role in wafer thickening.