Kyle Walraven

Mont106N

Lab on Correlation, Regression, and Data Analysis

A) After plotting the original data for SMAs in SP 1 (Pop vs. Rank), a clear view was given of the poor relationship between the populations of the given SMAs and the rank they have compared to other SMAs. A plot of the regression line also shows that fit was not very good at all. With a regression coefficient or .786 and clear patterns developing in the regression data, the relationship between population and rank is shown to be not linear. After taking the natural logarithm of the populations and graphing it with the rank in SP 2 (LN vs. Rank), it became obvious that this is closer to a true representation of the relationship. Unfortunately, strong patterns persisted in the regression data. A final graph, however, of the natural logarithm of the population plotted against the natural logarithm of the rank proved to be the closest representation yet. Regression data showed only weak patterns, and the regression coefficient was .985.

 Some might consider the first eight data points as outliers, or poor representations of the data as a whole. Therefore, it may be useful to exclude them in the data compilation. After plotting the previous three graphs again, this time without the first eight data points (SP 1 (-First8), SP 2 (-First8), SP 3 (-First8)), it became clear that this point of view held some validity. The regression coefficient increased in every case. Additionally, patterns in the regression data were less profound. Therefore, the data leads one to believe that the most appropriate representation of the relation between the population of the SMA and its rank is the natural logarithm of both data sets. This leads to the conclusion that a power law governs this relationship. That is one where the population is a function of a constant multiplied by the rank which had been affected exponentially.

B) A statistic that can be used to measure the uniformity of the polysilicon thickness across all the sites on one of the wafers is the standard deviation. The higher the standard deviation, the more widely spread out the data points are from the average. This means that wafers with high standard deviations had surfaces that were not very uniform. The closer the standard deviation is to zero, the closer the wafer’s surface is to being flat. After reviewing the data and making a few quick computations of average, it may be seen that the data from Location 13 on all of the chips is significantly lower than the previous heights. For each case, it persists that the height in Location 13 is lower than the average of Locations 1-13, and much lower than the average of Locations 1-12. In this way, excluding the data from Location 13 would help to streamline the data and provide more accurate conclusions.

 Using a multiple regression, we can compile the relationship between the oxide thickness, deposition time, and the standard deviation of the heights of the wafers and multiple locations. This will give us a formula in the form of y=m(1)x(1) +m(2)x(2)+b where y is the standard deviation, x(1) is oxide thickness, and x(2) is the deposition time. It can be seen from the data readout that m(1) is -0.071, m(2) is 1.225, and b is 115.336. The signs of the m’s indicate that oxide thickness tends to decrease when the thickness of polysilicon thickness increases and deposition time increases when polysilicon thickness increases. The regression coefficient of .932 indicates that this relationship is a good fit for the data. Furthermore, the residual data has no clear, developed patterns, reinforcing the relationship of the data.