Brandon Nunn

Lab on Correlation, Regression, Data Analysis

**USMA’s File Questions**:

1. There seems to be a somewhat strong negative relation between rank(x) and population(y) as displayed by the correlation coefficient of -0.703. However, this relation does not appear to be linear. This is because there is a strong pattern in the residual plot indicating lack of fit for a linear relation. They are not randomly spread out which is good evidence for no real linear relation.
2. There seems to be a strong negative relation between rank and the natural log of the populations. The correlation coefficient is -0.954 which indicates the strong negative correlation. This correlation does not seem to be a good linear fit because the residuals follow some pattern and are not randomly plotted.
3. It seems like there is a linear relation between the natural log of the ranks and the natural log of the populations. There is a very strong negative relation (r=-0.985) and there is no real strong pattern in the residual plot. Since the residuals are fairly randomly plotted, it seems as though this is a good fit for a linear relation between these two variables.
4. If you repeat these steps on just the 9th through the 75th SMA’s, all three of the correlation coefficients grow closer to -1, indicating stronger negative and more linear relations between the variables. The correlation for the rank vs. population grows to -0.890 but because there is a pattern in the residual plot, it still does not seem to be a good fit for a linear relation. The correlation for rank and the natural log of the populations grows to -.979 and although the residuals are more spread out then before, there is still a small pattern which indicates no true linear fit for the relation between the two variables. The correlation for natural log of rank and the natural log of the populations grows to -.995. There is no pattern in the residual plot which indicates that it is a good fit for linear relation.
5. There is a functional relation between rank and population and it is a regression line in the form y=mx +b. This is the line that minimizes RMS error. It is not a very good fit for a linear relation because there are patterns in the residuals. The functional form for rank and natural log of the populations is in the form y=ce^kx. This is derived from exponential laws, and c and k are constants. You take the natural log of both sides and get ln(y)=kx + ln (c). K is the slope and ln(c) is the y intercept. This also does not seem to prove a good linear fit because there is still some pattern in the residual plot. The functional form for natural log of rank and natural log of the populations is in the form y=cx^a. This is derived from power laws and c and a are constants. If you take the natural log of both sides, you get ln(y)=a ln(x) + ln(c). The slope of the line would be a and the intercept is ln(c). Out of all of our models, this is the best linear fit for the variables, because there is no real pattern in the residual plot. They are fairly randomly scattered, indicating linear relation between natural log of the ranks and natural log of the populations.

**Semiconductor File Questions:**

1. The statistic used to measure uniformity of the polysilicon thickness across all the sites on one of the wafers is the standard deviation. This measures how far each of the values of polysilicon thickness (in angstroms) are from the mean value of the 13 sites. Thus, it is a good measure of uniformity.
2. The data from site 13 on each wafer seems to be an outlier. The data should be excluded because the polysilicon thickness in angstroms from this site is the lowest for every single wafer besides B2. In many of these cases the numbers for site 13 are drastically lower than the numbers for the other sites, and this discrepancy is why the data from site 13 should be excluded.
3. Y(SD) M1(oxide thickness)=-0.06 M2(deposition time)=3.05 Intercept(b) = 225

Y=m1x1+m2x2+ b

1. The relation stated above seems to be a good fit for the data. The data is roughly football shaped. Both of the residual plots for the deposition time and oxide thickness have no patterns, so both variables have a linear relation with the standard deviation. This relation is a good fit for the data.
2. The signs on the coefficients m1 and m2 indicate that the slope of the regression line is either positive or negative. The signs will prove whether oxide thickness and deposition time have positive or negative relationship with the standard deviation. These signs indicate which way the oxide thickness and deposition time affected the measure of polysilicon thickness uniformity. There is a negative sign for oxide thickness, which means it is negatively related to polysilicon thickness uniformity. Since, its coefficient is much smaller than the one for deposition time, deposition time affects polysilicon thickness uniformity more than oxide thickness. The positive sign for deposition time means that it is positively related to polysilicon thickness uniformity.