Anthony Cunningham

STAT 3200

Due 2/8/2017

**Homework 2**

1. a. > library(car)

> ?Leinhardt

> boxplot(Leinhardt$income ~ Leinhardt$region, main="original income")



\* The IQRs of the four regions do not line up very well, as the group Europe’s IQR is far removed from the other three groups. The Europe group is also much more spread out than the other three groups, while group Africa’s spread is tiny. It’s obvious that the constant variance assumption will not hold with the data set in its current form. There are some abnormally high outliers for the Africa, Americas, and Asia groups, while the Europe group has no outliers.

b. > temp = Leinhardt[Leinhardt$region == "Africa",]

> temp[rev(order(temp$income)),]

income

Libya 3010 \*Libya, South Africa, and Tunisia are the outliers in group Africa

South.Africa 1000

Tunisia 434

> temp = Leinhardt[Leinhardt$region == "Americas",]

> temp[rev(order(temp$income)),]

income

United.States 5523 \*United States and Canada are the outliers in group Americas

Canada 4751

> temp = Leinhardt[Leinhardt$region == "Asia",]

> temp[rev(order(temp$income)),]

income

New.Zealand 3723 \*New Zealand, Australia, Israel, Saudi Arabia, Iran, and Singapore

Australia 3426 are the outliers in group Asia (which includes Asian and Oceanic

Israel 2526 countries)

Saudi.Arabia 1530

Iran 1280

Singapore 1268

c. > logincome = log(Leinhardt$income)

> boxplot(logincome ~ Leinhardt$region, main="log-transformed income")



\* The spread of the groups are a lot more similar than the original data, although not too similar. Asia’s IQR and range appear to be larger than the other three groups.

> sqrtincome = sqrt(Leinhardt$income)

> boxplot(sqrtincome ~ Leinhardt$region, main="sqrt-transformed income")



\* The spreads are more similar than the original data set. Europe’s IQR and range appear to still be larger than the other three groups, but it is not as blatant as in the original data set.

d. While both transformations work in making the spreads more similar than the original data set, I would prefer the log transformation over the square root transformation because the spreads appear to be more similar under a log transformation than a square root transformation. For instance, group Africa’s IQR is much more similar to the other groups’ IQRs.

2. a. > mydata<- read.csv(file="http://www.stat.uiowa.edu/~ernli/ALRdata/HW2scatterplot.csv", head=TRUE, sep=",")

>attach(mydata)

> plot(Y ~ x)

> fit = lm(Y ~ x, data=mydata)

> abline(fit)



> cor(x,Y)

[1] 0.9473655

\* Despite the regression line being a strong fit to the data, as told by the correlation coefficient being near 1, I don’t think that it is appropriate to use simple linear regression on this data in its current form. I also believe that the linearity assumption doesn’t hold for this data. It looks like a quadratic regression would be a much cleaner fit for this data, since the scatter plot resembles one half of an upward-facing parabola.

b. Decreasing the ladder of powers for the variable Y would allow the data to conform to linearity. This is due to Musteller and Tukey’s bulging rule, which provides a diagram in selecting linear transformations based on the general shape of the current data. This data’s “bulge” appears to be pointing down and to the right. By the bulging rule, we could either either transform the variable x up the ladder of powers or transform the variable Y down the ladder of powers. Since the question asks how one would transform Y to allow the data to conform to linearity, then one would just suggest transforming Y by decreasing the ladder of powers.

c. I think that the square root transformation on Y works the best in conforming the data to linearity. The Y^2 transformation obviously doesn’t work since we need to decrease, not increase, the ladder of powers. Both log transformations over-adjusted the data, resulting in a bulge pointing up and to the left. The square root transformation, however, was a happy medium, as the data fell in a pretty straight-looking line.

> sqrtY = sqrt(Y)

> plot(sqrtY~x)

> fit = lm(sqrtY ~ x, data=mydata)

> abline(fit)



> cor(x,sqrtY) \*The correlation coefficient for the square root-transformed Y data

[1] 0.9837989 is closer to 1 than the correlation coefficient for the original data.

3. > library(car)

> ?Freedman

> prop = Freedman$nonwhite / 100

a. > par(mfrow=c(1,2))

> hist(prop, n = 20)

> boxplot(prop)



\* No, the values of the proportions do not cover values on the negative real line because negative values do not make sense in this context. It is impossible to have negative non-white population; the lowest this data can go is 0.0.

b. The transformation commonly used on proportions is the logit transformation, which essentially takes the log of the odds (probability of success divided by probability of failure; in this case, the probability of randomly selecting a “non-white” individual divided by the probability of randomly selecting a “white” individual). The formula of the logit transformation = ln(P/(1-P)), where P is a given proportion in one’s data set.

c. > logitprop = logit(prop)

> par(mfrow=c(1,2))

> hist(logitprop)

> boxplot(logitprop)



\* Yes, the transformed proportions do cover both the positive and negative sides of the real line, although a vast majority of the transformed proportions fall on the negative side, which is decidedly not symmetric about 0. This is due to the skewness of the original data set, as most of the proportions fall near 0, while to the distribution is skewed to the right.