Assignment 2 CS 4070

Nirmal Roy, 4724429

December 4, 2017

Question 1

We use the number, Hinge Ratio $U = \frac{UpperHinge - Median}{Median - LowerHinge}$ to check the symmetry of a distribution. A number close to one indicates the distribution is symmetric. Whereas $U > 1$ indicates a positive skew and $U < 1$ indicates a negative skew. We check the value of U and make transformation to make the data symmetric.

Population

For population, $U = 1.19$ and hence the data is positively skewed as can be seen from the first plot of figure 1. Hence, we need to climb the ladder of powers to find a transformation that makes U close to 1. We ran the code [1] in Appendix and found that for an exponent of 0.76, $U = 1.003$. Hence, we show a boxplot comparison of the original variable and a transformed variable with an exponent of 0.76

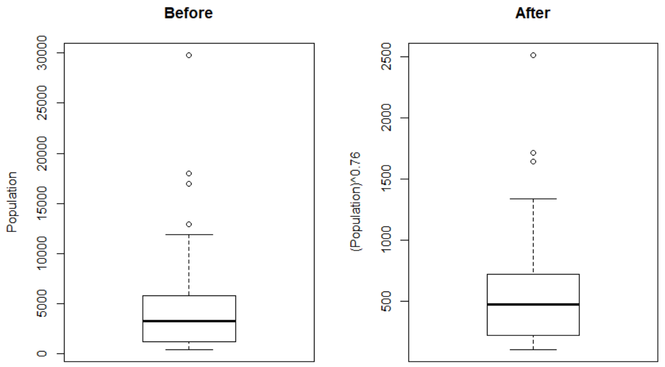


Figure 1: Boxplot comparison for the original Population variable and the transformed variable

Verbal SAT score of students

For SATV, $U = 1.537$ indicating a high positive skew. In this data the ratio of maximum value to minimum value is 1.287(close to 1). Hence we add a negative start of 397 which is the minimum value for this data. We run a code similar to [1] in Appendix but with the S(ATV-397) variable and we find that for an exponent of 0.25 value of U becomes close to 1 and we have symmetry as seen in fig 2. We plot $(SATV - 397)^{0.25}$ after transformation.

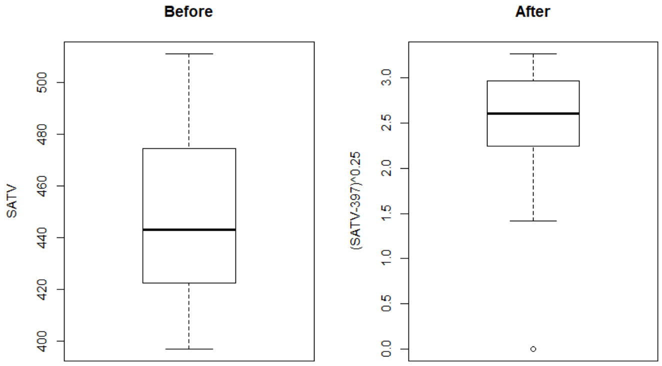


Figure 2: Boxplot comparison for the original SATV variable and the transformed variable

Maths SAT score of students

For SATM, similar to SATV $U = 1.625$ indicating a high positive skew of the data, having a max to min ratio close to 1. With a negative start of 437and running the same code we find that for an exponent of 0.02 value of U becomes 1.02. Since the exponent is close to zero we change the offset to 436 and do a log10 transformation to make the data symmetric as shown in fig 3

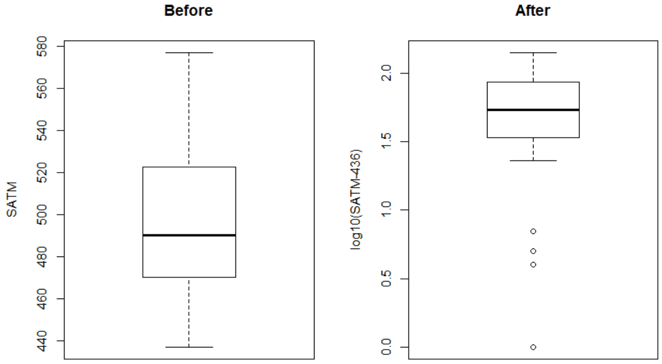


Figure 3: Boxplot comparison for the original SATM variable and the transformed variable

Percentage of graduating high school students

The percent variable describes the percentage of graduating high-school students in the state who took the SAT exam. It's U value is 2.41. Being positively skewed and a percentage, we applied the logit transformation, to which the value of U became 1.49. It has obviously made the distribution more symmetric but it can be made more symmetric and so we use the log10 transformation. With log10 transformation the value of U becomes .107 and hence we do the boxplot comparison with log10 transformed variable.

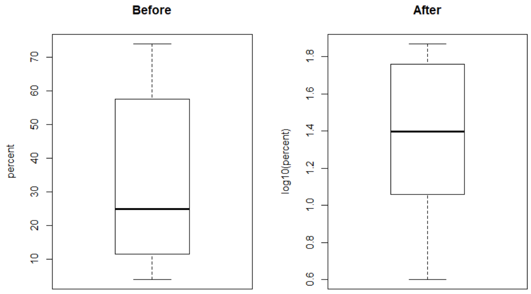


Figure 4: Boxplot comparison for the original percent variable and the transformed variable

State Spending in Education

The U value for dollars variable is 0.932 which indicates a bit of negative skewness in the data. And hence unlike previous variables, we have to climb up the ladder of power. Running code [] in Appendix , we see that at an exponent of 1.38, U becomes 1.02. The data was already fairly symmetric, hence we might not need a transformation on it. Nonetheless we show a boxplot comparison of the original variable with the transformed variable raised to the power of 1.38

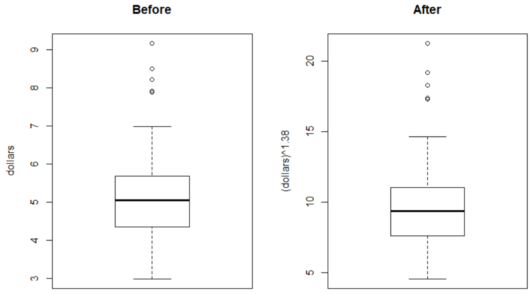


Figure 5: Boxplot comparison for the original dollars variable and the transformed variable

Average teacher's salary in the state

The U value for the pay variable is 1.4 indicating a positive skew again and we follow previous steps to check that at an exponent of -2.13, the value becomes 1.0196. In fig we show the boxplot comparison of the original variable with the transformed variable making it symmetric. The negative sign at front preserves the direction of the data when we do prower transform with negative exponent.

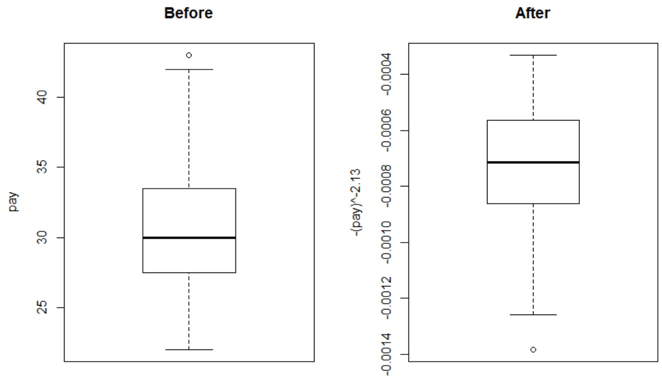


Figure 6: Boxplot comparison for the original pay variable and the transformed variable

Question 2

Relationship between Per Capita Income and Infant Mortality Rate

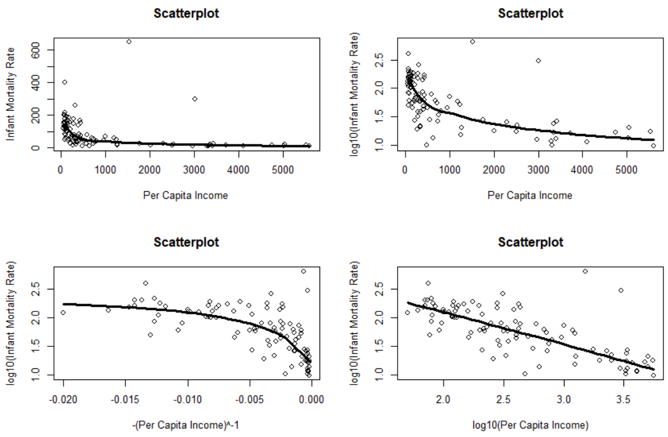


Figure 7: Boxplot comparison for the original pay variable and the transformed variable

For this part, we plot four figures(fig 7) to realize the best possible transformation to make the relationship reliable. Going clockwise from top left, the first figure is the original plot between the two variables. The lowess fit indicates a curve towards the bottom left hence intuitively we need to climb down the ladder of powers for both the explanatory and response variable. We start by log transforming only the response variable and the resulting plot is the second figure. We still have some curve towards the bottom right and hence we need to transform the explanatory variable too. Hence in the third plot we transform the explanatory variable by raising it to a power of -1 but we see it's a bit too much. Hence after several rounds of trial and error iteration of transforming each and both of them we see that when we log10 transform both we get a linear relationship between the two variables.

Relationship between Region and Infant Mortality Rate

For the last two relationship we use Tukey's graph of log hinge-spread against log median which is provided by the function '*spreadLevelPlot*' in R. It outputs a 'suggested' power transformation of 0.226545 . The suggestion comes from the equation $log(\text{spread}) = a + blog(\text{level})$, where transformation is given by $(1 - b)$. Since this value is close to 0 we use the standard log10 transformation to do the boxplot comparison as seen fig

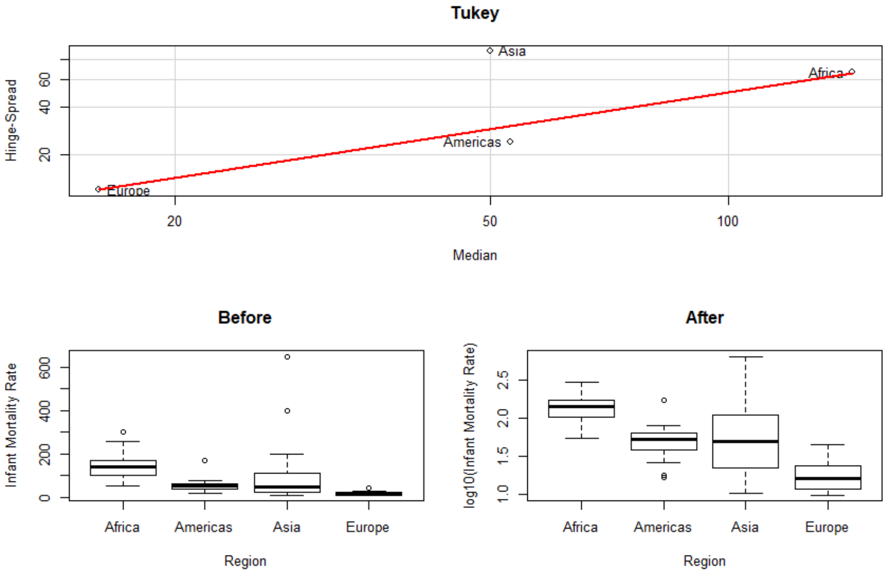


Figure 8: Tukey graph and boxplot comparison for the original variable and the transformed variable indicating relationship between Region and infant mortality rate per 1000 live births with boxplots

Relationship between whether a country exports oil and Infant Mortality Rate

From the Tukey graph(the spreadLevelPlot function), we have the suggested -power transform of -0.288 and again this value being close to 0 we use the standard log10 transform of the mortality rate variable to show the boxplot comparison

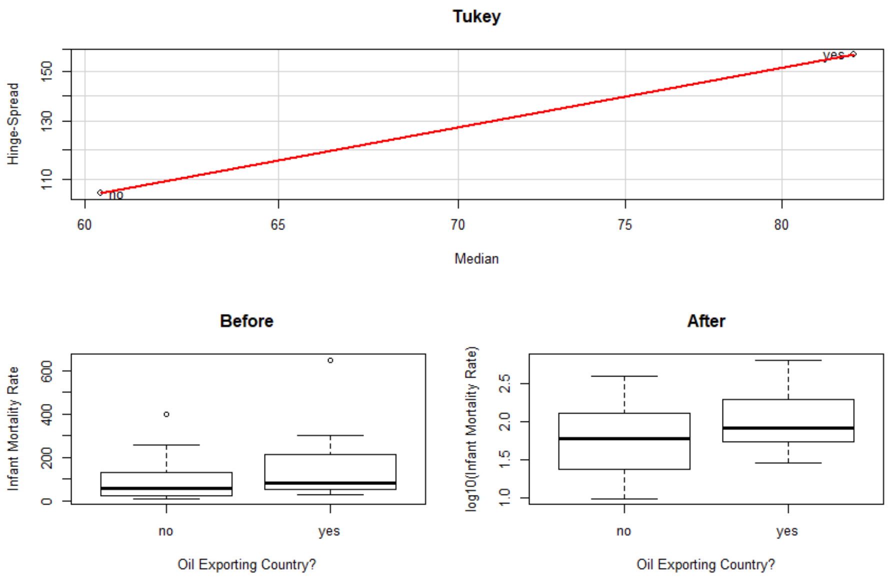


Figure 9: Tukey graph and boxplot comparison for the original variable and the transformed variable indicating relationship between oil export and infant mortality rate per 1000 live births with boxplots

Question 3

1. In fig10, we see the scatter plot of the response variable Acres/Gardener on Consumers/Gardener. We use the R function ‘*summary*’ and ‘*lm*’ to get the statistics of the simple linear regression model fitted on this data. The black line with equation $Y = 1.3756 + 0.5163X$ is the regression line on this data. We see that, $S_E = 0.4543$ and $r^2 = 0.1411$. The intercept value $A = 1.3756$ does not have much significance whereas $B = 0.5163$ indicates that the society is somewhere between a society characterized by ‘primitive communism’ and a society in which redistribution is purely through the market.

2. When we remove the fourth household, which was more of an outlier in this data, and use the above mentioned functions to get the regression statistics, we get a reduced error(according to our expectation) $S_E = 0.3681$ and $r^2 = 0.3264$. We plot the red line which is the fitted simple regression curve for the data without the fourth household. The line follows the

equation $Y = 1 + 0.7216X$. A higher slope $B = 0.7216$ shows that without the fourth household, the society is more of the second kind, where each household have to work in proportion to it's consumption needs. That is higher the number of consumers in a household, they should have more acres.

3. The second regression line perhaps does a better job but according to me none of them really does a good job in characterizing the entire data. Because if we take the mean of the response variable, it's 2.162 and when we compare the values of error to it, they are not insignificant or low. Hence, both produce relatively high error and maybe simple linear regression is not the best fit for this data.

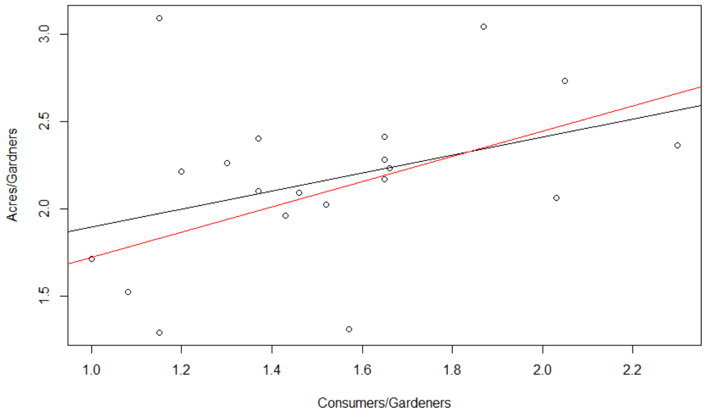
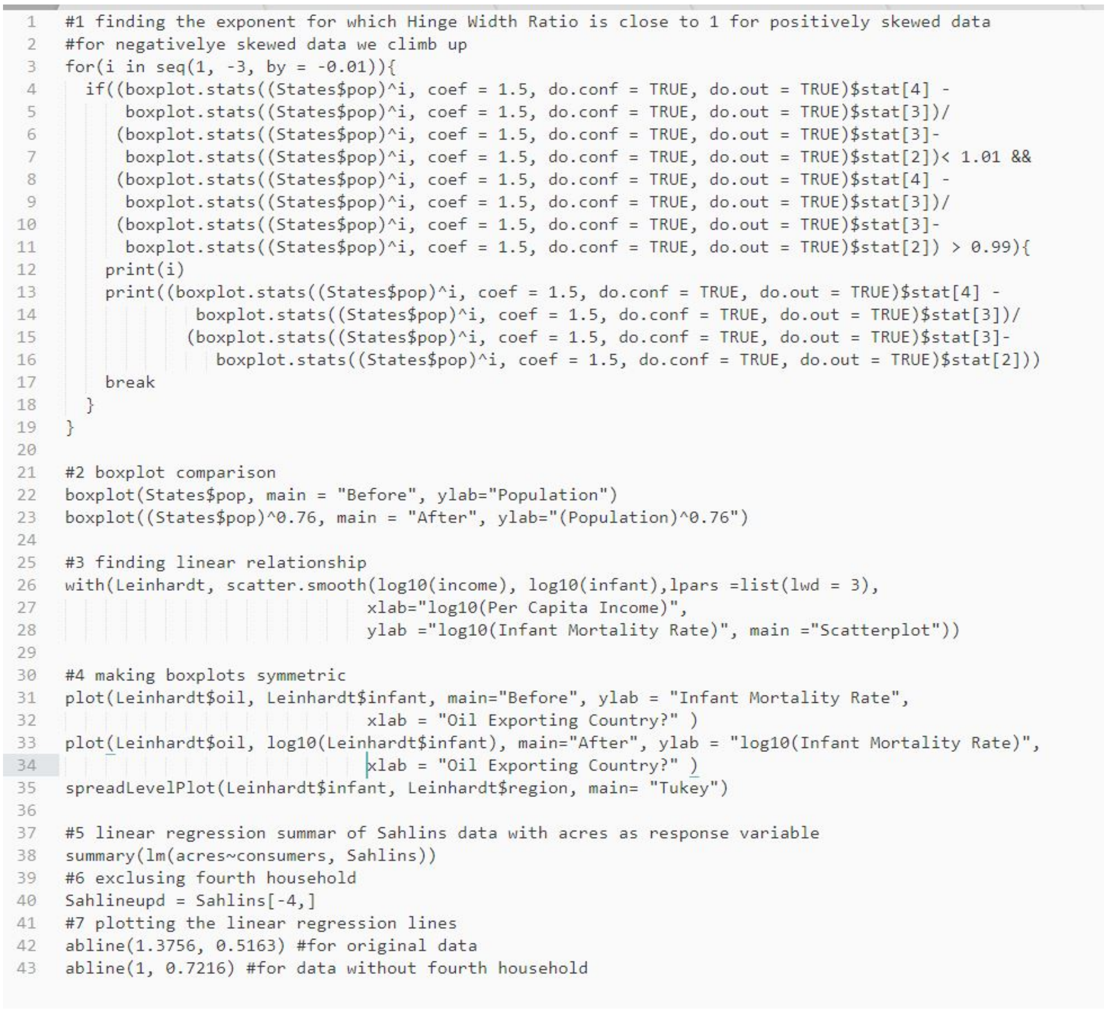


Figure 10: Scatterplot of response variable acres on explanatory variable consumers. The black line is a simple linear regression fit on the original data. The red line is a simple linear regression fit on the data without the fourth household

Appendix



```
#1 finding the exponent for which Hinge Width Ratio is close to 1 for positively skewed data
#for negativelye skewed data we climb up
for(i in seq(1, -3, by = -0.01)){
  if((boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[4] -
      boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[3])/
     (boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[3]-
      boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[2])< 1.01 &&
     (boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[4] -
      boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[3])/
     (boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[3]-
      boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[2]) > 0.99){
    print(i)
    print((boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[4] -
             boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[3])/
            (boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[3]-
               boxplot.stats((States$pop)^i, coef = 1.5, do.conf = TRUE, do.out = TRUE)$stat[2]))
    break
  }
}

#2 boxplot comparison
boxplot(States$pop, main = "Before", ylab="Population")
boxplot((States$pop)^0.76, main = "After", ylab="(Population)^0.76")

#3 finding linear relationship
with(Leinhardt, scatter.smooth(log10(income), log10(infant),lpars =list(lwd = 3),
                              xlab="log10(Per Capita Income)",
                              ylab ="log10(Infant Mortality Rate)", main ="Scatterplot"))

#4 making boxplots symmetric
plot(Leinhardt$oil, Leinhardt$infant, main="Before", ylab = "Infant Mortality Rate",
                              xlab = "Oil Exporting Country?" )
plot(Leinhardt$oil, log10(Leinhardt$infant), main="After", ylab = "log10(Infant Mortality Rate)",
                              xlab = "Oil Exporting Country?" )
spreadLevelPlot(Leinhardt$infant, Leinhardt$region, main= "Tukey")

#5 linear regression summar of Sahlins data with acres as response variable
summary(lm(acres~consumers, Sahlins))
#6 exclusing fourth household
Sahlineupd = Sahlins[-4,]
#7 plotting the linear regression lines
abline(1.3756, 0.5163) #for original data
abline(1, 0.7216) #for data without fourth household
```

Figure 11: Relevant Codes