Benjamin Goodwin

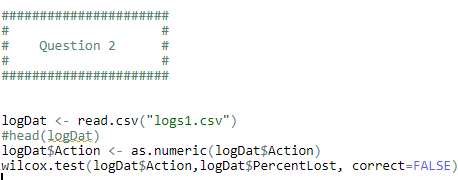
MSDS6371

HW4

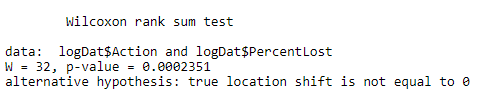
1. Read Chapter 4 from Statistical Sleuth and answer the conceptual problems at the end of the chapter. Note: You do not need to type these up and turn them in. The answers are at the very end of the chapter.
2. A. Perform a complete analysis using a rank sum test in SAS. (Logging data

B. Verify the p-value and confidence interval by running the rank sum test in R

R Input



R Output



3) A. State the problem, address the assumptions. Be sure to support with your knowledge of theory (CLT) as well as with histograms, box plots, q-q plots, etc

A) A study was conducted concerning income levels of people in two groups, the first was non-college education people, and the second was college educated people. The purpose of the study was to determine if there was a difference in income levels between the groups. We would like to statistically determine if there is any difference in income levels between the two groups of people. In further detail the researchers conducted a study of subjects between ages 41 and 49 years old in the year 2006. They are hoping to answer their initial question using this population, which was, “is there a difference in mean income level between subjects in the group who went to college vs those who did not.”

B) Show all 6 steps, including a thoughtful, thorough, yet non-technical conclusion. Include a confidence interval.

Step 1: H\_0:

Step 2: H\_A:

Step 3:

Step 4

Step 5

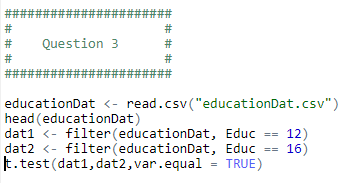
Step 6

Conclusion and confidence interval:

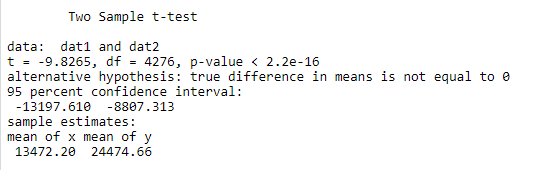
C) Scope of inference: The data is a subset of National Longitudinal Survey of Youth (NLSY79). The question did not indicate random sampling. However, the sample size was quite large, all subjects were between 41 and 49. I believe this data can be inferred om the population of subjects sampled.

D) Verify the Welch’s t statistic and p-value with R (using R function t.test).

R input:



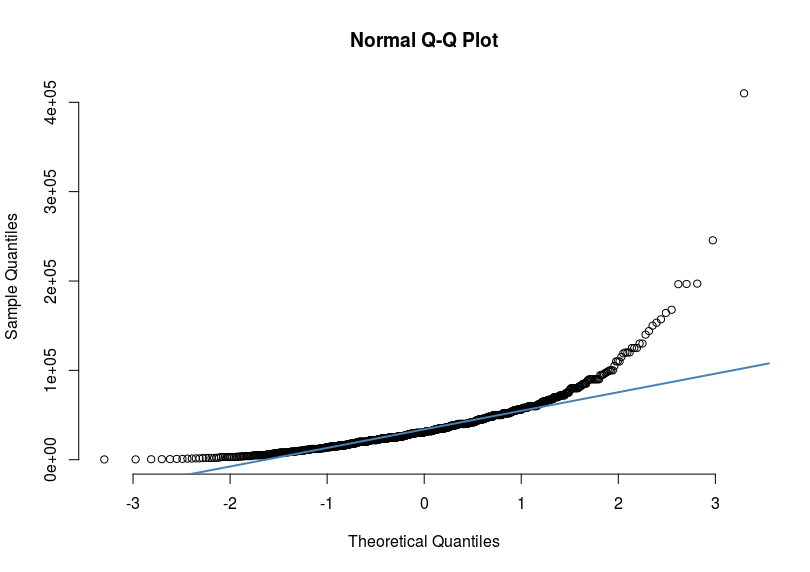
R output:



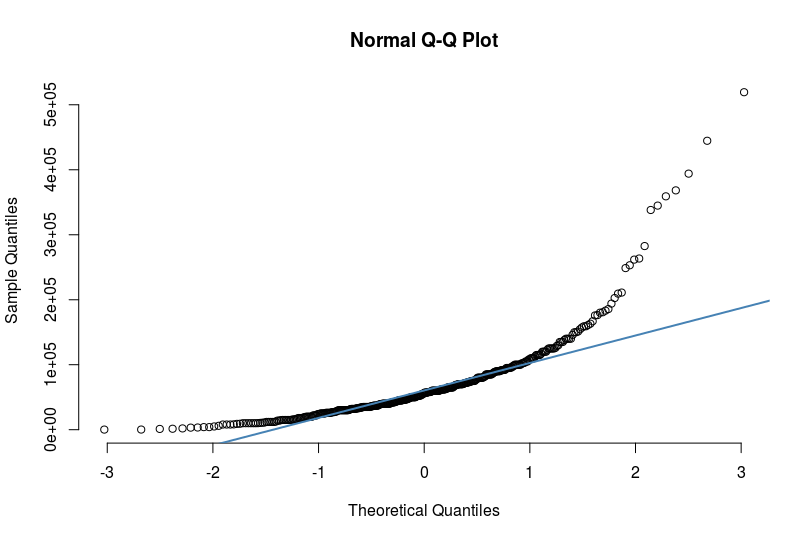
Would you prefer to run the log transformed analysis you ran in HW3, or do you feel this analysis is more appropriate?

I think that the log transformed analysis is more appropriate, this is because I feel the log transformed model better conforms to model assumptions of normality of data, equal variances, and independent data. I have attached the plots below of the transformed data and non-transformed data.

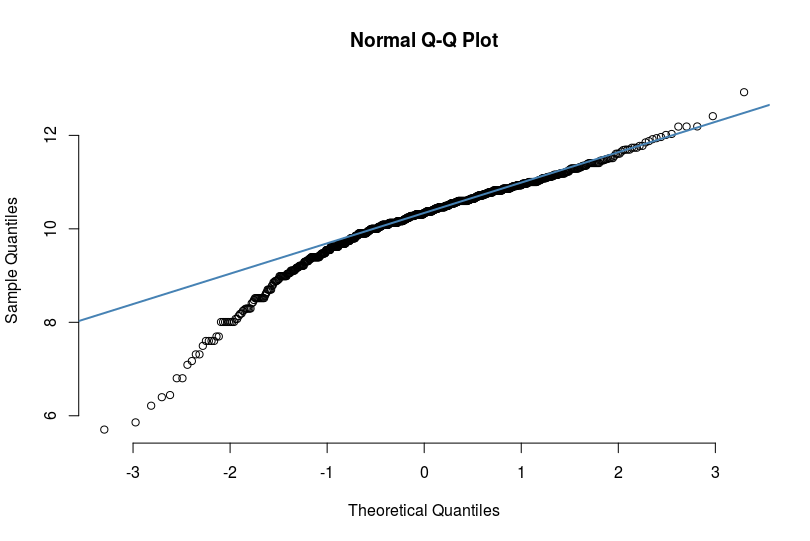
Plot 1: (QQ-plot for non-transformed data (12-year education data))

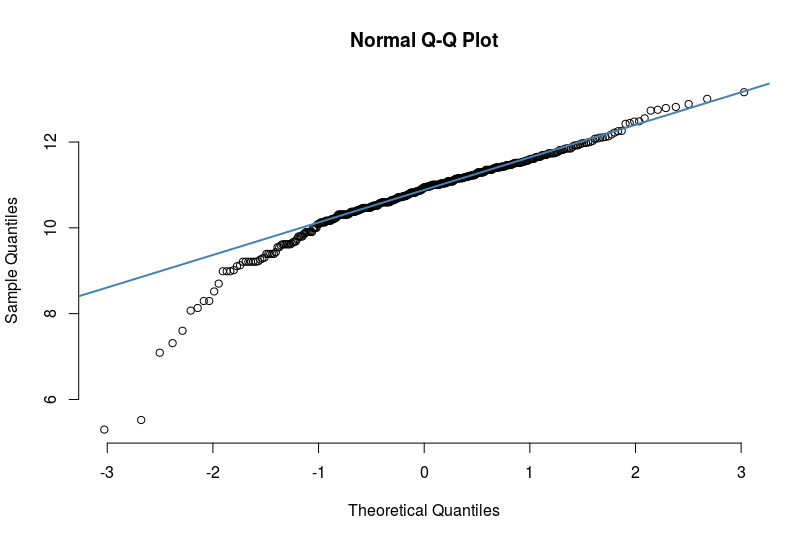


Plot 2:(QQ-plot for non-transformed data (16-year education data))



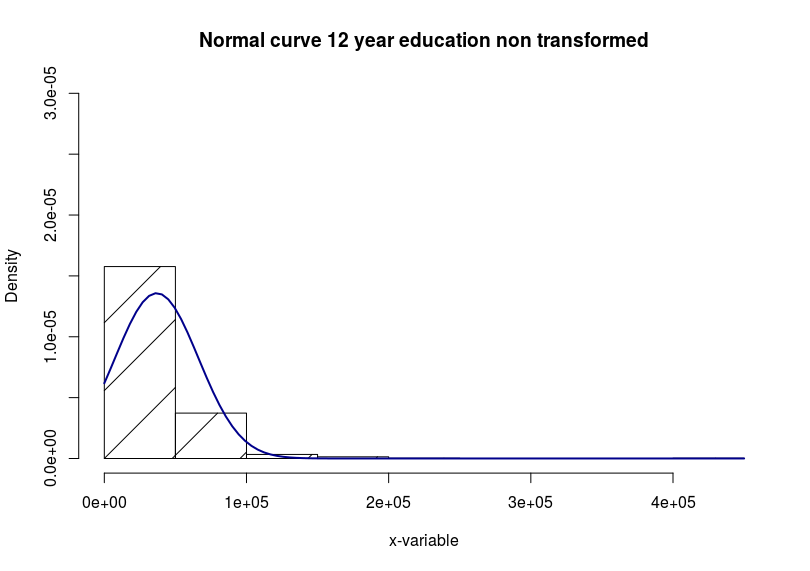
Plot 3:(QQ-plot for transformed data (12-year education data))



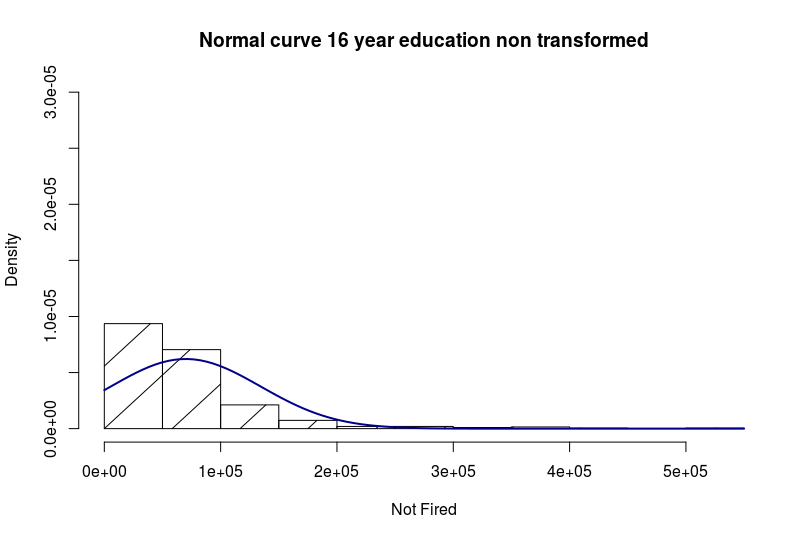
Plot 4: (QQ-plot for non-transformed data (12-year education data))

Based on the qq-plots, it seems that log transformed data better adheres to the assumption of normally distributed data

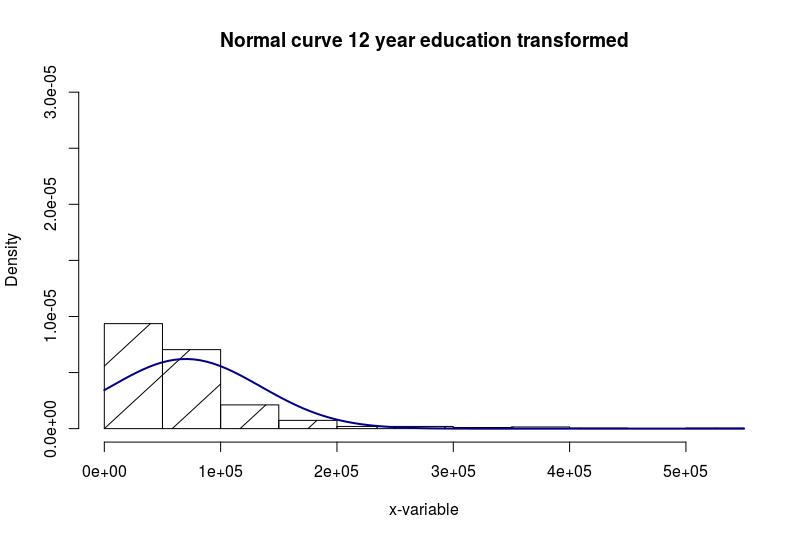
Histogram 1: (Plot of normal curve overlaid by non-transformed data (12-year education data))



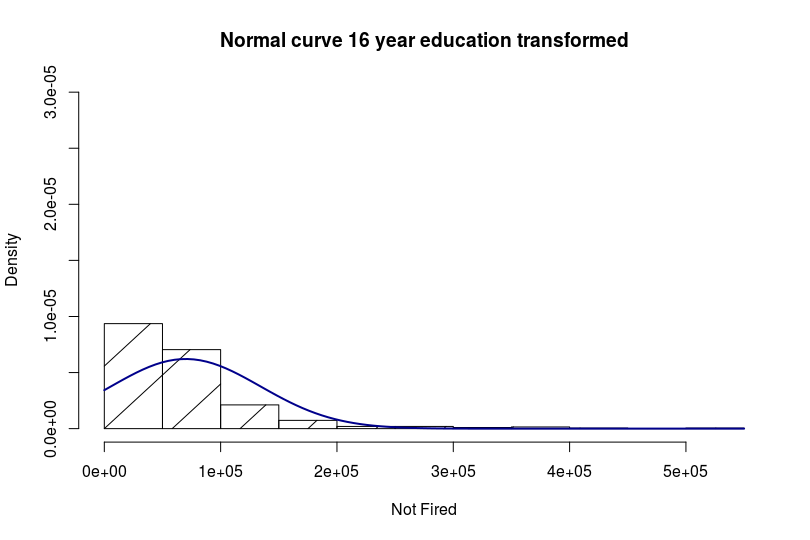
Histogram 2: (Plot of normal curve overlaid by non-transformed data (16-year education data))



Histogram 3: (Plot of normal curve overlaid by transformed data (12-year education data))



Histogram 4: (Plot of normal curve overlaid by non-transformed data (16-year education data))



Based on all the model assumptions (plots above) and general subject knowledge, we can see that the transformed model more correctly follows our assumptions for a t-test. Transforming the data has the interesting effect of changing the conclusion, instead of rejecting the null, we will now fail to reject the null on the transformed data.