Ben Goodwin

DS6371

HW5

1. Plotting raw data:

Chart, box and whisker chart

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Data after log transform:

Chart, box and whisker chart

Description automatically generatedChart, histogram

Description automatically generated

There are 3 main assumptions for ANOVA:

1. The responses for each factor level have a normalpopulationdistribution.
2. These distributions have the samevariance.
3. The data are independent.
4. After log transforming the data, it went from being heavily right skewed to more left skewed. And based on the qq-plots, the data does not appear to be normally distributed.
5. The plot suggests we have non-constant variance, and the null hypothesis of H\_0: sigma\_1=sigma\_2=Sigma\_n is strongly rejected. F<0.0001
6. Independence is a strange one, as it should have been part of the deisngof the study. Since we have worked with this dataset on a few hw’s and I have spent some time with it. The assumption of independence is met.

Below are the diagnostics plots for the log-transformed data:

Diagram, engineering drawing

Description automatically generated

And below are the plots for the non-transformed data:

Graphical user interface, diagram

Description automatically generated

I do believe the assumptions are better met by the log-transformed data.

* 1. State the problem: 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.” Since we are performing an analysis of variance experiment we are merely interested in seeing if any of the groups are different.
  2. Address the assumptions

There are 3 main assumptions for ANOVA:

1 The responses for each factor level have a normalpopulationdistribution.

* + - 1. These distributions have the samevariance.
      2. The data are independent.

1. After log transforming the data, it went from being heavily right skewed to more left skewed. And based on the qq-plots, the data does not appear to be normally distributed.
2. The plot suggests we have non-constant variance, and the null hypothesis of H\_0: sigma\_1=sigma\_2=Sigma\_n is strongly rejected. F<0.0001
3. Independence is a strange one, as it should have been part of the design of the study. Since we have worked with this dataset on a few hw’s and I have spent some time with it. The assumption of independence is met.

Below are the diagnostics plots for the log-transformed data:

Diagram, engineering drawing

Description automatically generated

And below are the boxplots of the data by which I visually looked at variance:

Chart, box and whisker chart

Description automatically generated

3) Conduct the test: F-Value is 45.22, yielding a p-value <0.001

4) There is strong evidence to suggest that the at least one of the groups has a mean income that is different than the others (p-value < .0001 from an ANOVA).   
5) 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

Six step test:

1. H\_0: Mu\_0=Mu\_1=Mu\_2\_Mu\_3\_Mu\_16
2. H\_A: At least one Mu differs
3. F-Statistic: 45.22
4. P-Value: <0.0001
5. Reject H\_0!
6. We rejected the null and concluded that there is at least one of the group means is different from the others
7. 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

Additional things!

R^2: 0.108080

MSE: 1900991197.1

R Input:

Graphical user interface, text, application, email

Description automatically generated

R output:A picture containing text

Description automatically generated

Question 2:

Text

Description automatically generated

Output:

A picture containing text

Description automatically generated

Output:

Text, letter

Description automatically generated

F-test

Step 1: sigma\_1=sigma\_2

Step 2: sigma\_2 != sigma\_2

Step 3: F-statistic: 1.2393

Step 4: P-value: 0.03444

Step 5: Decision, Reject H\_0

Step 6: Conclusion, we reject H\_0 and conclude there is a significant difference between the two groups in terms of income.

And based on the evidence in the full and reduced models, and based on Tukey HSD, we will fail to reject the null hypothesis and conclude that there isn’t a difference between the groups.

3)

Step 1: Null hypothesis: Mu0=Mu1=mu2=mu3=mu16

Step 2: At least one of the groups differ

Step 3: F-statistic: 45.22

Step 4: P<0.001

Step 5: Reject H\_0

Step 6: We reject H\_0 and conclude that there is a difference between at least one of the groups. We will use Tukey HSD to determine which groups are actually different.

Step 7: 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.

R input:

* A picture containing text

  Description automatically generated

R output:

Chart, scatter chart, box and whisker chart

Description automatically generated

From the chart above, we can now visually see that there are differences between the groups, and we can see between which groups the differences are more pronounced. We can see all of our pairwise comparisons. The extended lines show 95% confidence intervals. Between 16 and 3 we can see that the line crosses zero, and is not considered statistically significant. It appears there is fairly strong evidence supporting differences in income between the educational groups besides the two highest educational groups.