Question 1

Problem of Interest

Is there evidence to justify the educational attainment has an impact on the level of income a person receives? Data was collected 2,584 Americans from a population of adults that were selected for the National Longitudinal Study of Youth in 1979 and who had income generating jobs in the year 2005. The data categorized their highest level of educational attainment and the income received in the year 2005.

Is there evidence that there education level has an impact on income? An Analysis of Variance Test (ANOVA) will be used to determine if there is evidence that at least one mean income category is different from the rest.

Assumptions of Test

In order to conduct an ANOVA test, several assumptions must be met. The underlying assumptions of ANOVA are that the data comes from a normal distribution, the population standard deviations are equal, and that the observations are independent of each other (both within and between samples).

The data was investigated graphically in order to determine if the data meets the assumptions required for ANOVA. First, the data was displayed using boxplots in order to understand the distribution of the data. A boxplot of the data is shown in **Figure 1**.

Figure 1

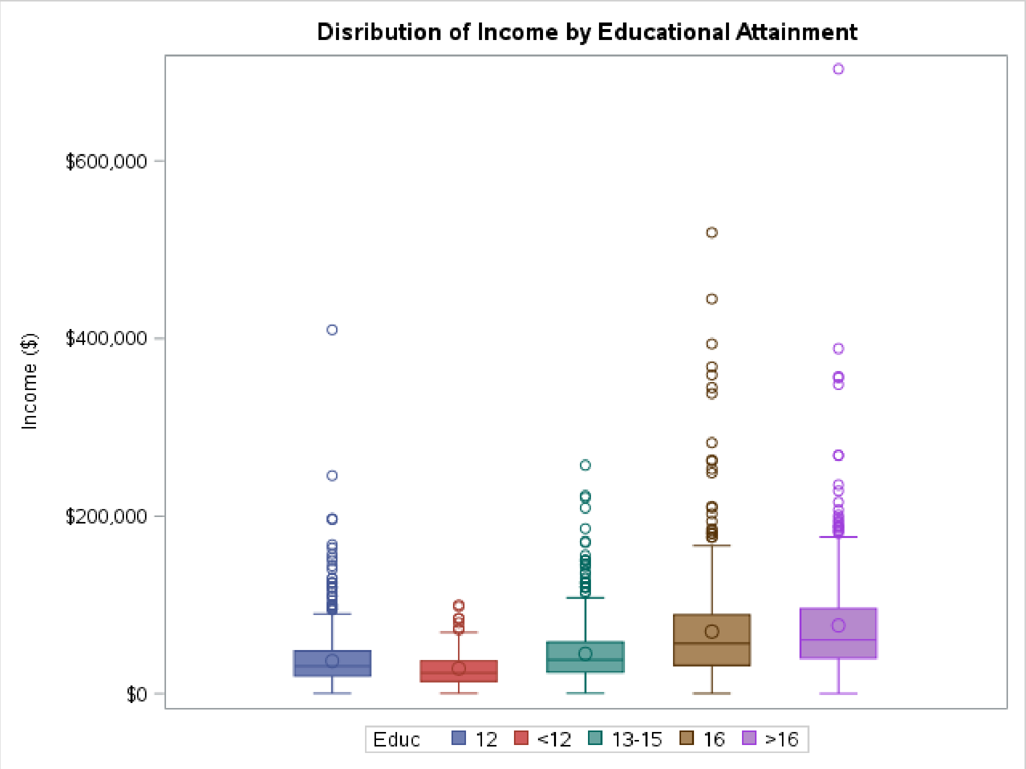


Figure 1: Distribution of Income by Educational Attainment. Where Educ is the education level achieved. For example "<12" indicates less than 12 years of education.

This boxplot quickly shows that the datasets are different from each other and that there is a significant amount of outliers for each category which is skewing the data. This leads me to believe that the population standard deviations for each group are different from each other.

Next, QQ-plots were used to determine if the groups are normally distributed. The QQ-plots, shown in **Figure 2, Figure 3, and Figure 4** (found in the **Appendix** of this report) reveal that the data is not normally distributed. Based on the boxplots and QQ-plots it seems reasonable to assume that the data would benefit from log transformation.

Figure 5, shows the distribution of the data after the data has been log transformed. The data now appears to be far less skewed. QQ-plots of the log transformed data reveal that the data is much more normally distributed after the log transformation (These QQ-plots can be found in the Appendix of this Report). Based on this analysis, there is not enough visual evidence to suggest that there are differences in variance between the groups once the data has been log-transformed.

Figure 5

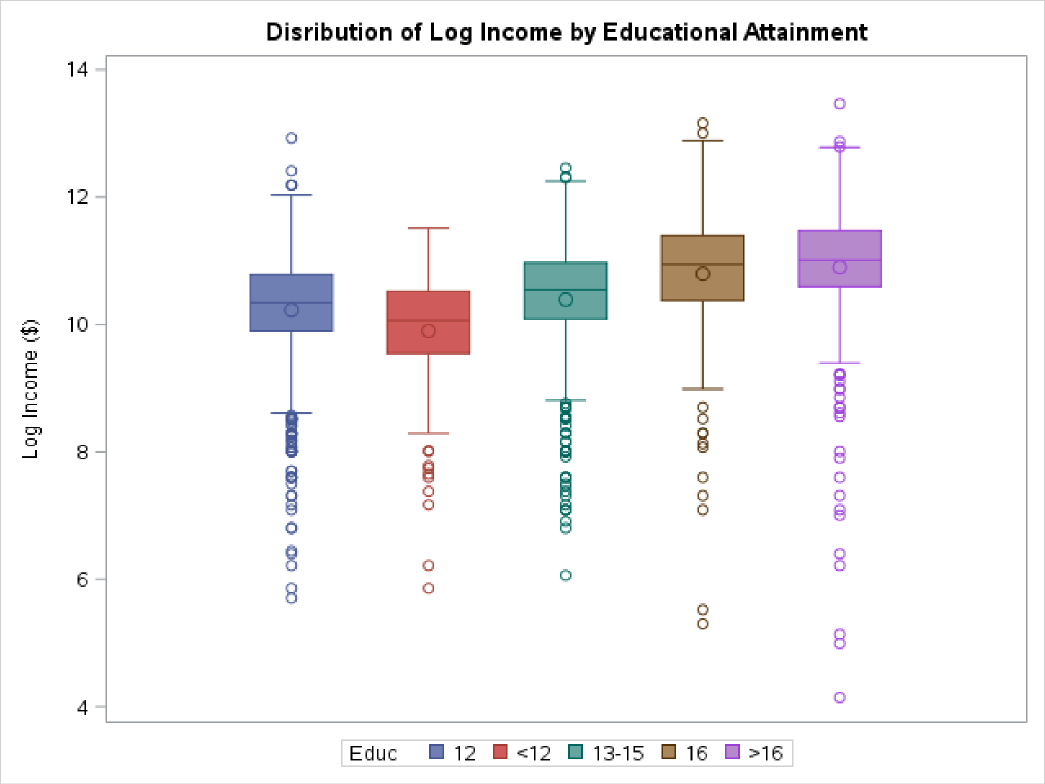


Figure 5: Distribution of Log Income by Educational Attainment. Where “Educ” is the education level achieved. For example “<12” indicates less than 12 years of education.

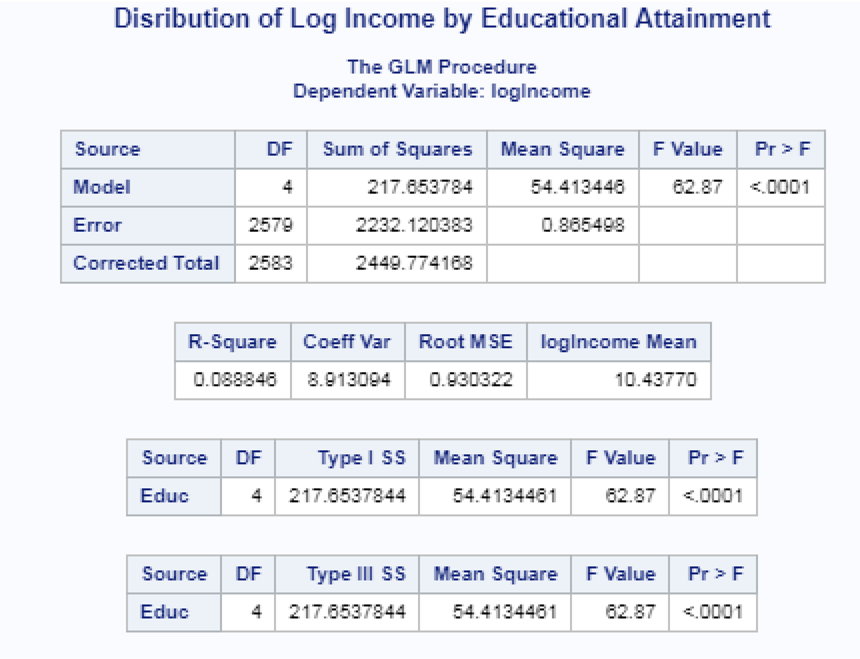
Hypothesis:

- H0: The mean of the log transformed data are all the same
- HA: The mean of at least one group is different from the other groups.

Analysis

There is strong evidence that the mean income earned by based on education attainment is different for at least one group (F-Statistics of 62.87 with an associated p-value of <0.0001). However, the ANOVA test did not explain much of the variation. The R^2 of associated with this test was 0.089 which indicates that on 8.9% of the variation was explained via ANOVA. The mean square error for this test was 0.865 with 2579 degrees of freedom. This analysis was conducted in SAS. The code used to conduct this ANOVA analysis, along with the code used to conduct the analysis is shown below. The entire code used throughout this analysis can be found in the appendix.

```
proc glm data=work.logedu;
      class educ;
      model logIncome = educ;
      lsmeans educ / stderr pdiff cl;
run;
```



Disribution of Log Income by Educational Attainment

The GLM Procedure
Dependent Variable: logIncome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

R-Square	Coeff Var	Root MSE	logIncome Mean
0.088846	8.913094	0.930322	10.43770

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Educ	4	217.6537844	54.4134461	62.87	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Educ	4	217.6537844	54.4134461	62.87	<.0001

Conclusion

There is strong evidence to reject the null hypothesis. This means that the mean income of one group is different from the other groups.

Score of Interest

Since this is an observation study with random selection, it is valid to make generalization about the entire population.

Same Analysis but using R instead of SAS

The same analysis was conducted in R. The code and output of R is shown below.

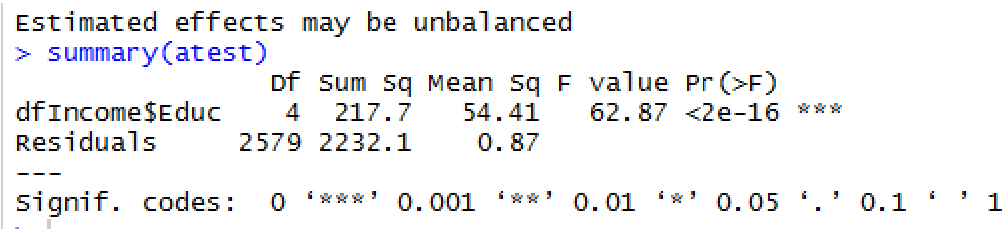
Code

```
#Load the data
dfIncome = read.csv("ex0525.csv")

#log transform the data
dfIncome["logIncome"] = log(dfIncome$Income2005)

#Run the test
atest = aov(dfIncome$logIncome~dfIncome$Educ)
summary(atest)
```

Results



```
Estimated effects may be unbalanced
> summary(atest)
                  Df Sum Sq Mean Sq F value Pr(>F)    
dfIncome$Educ      4  217.7   54.41   62.87 <2e-16 ***
Residuals       2579 2232.1    0.87                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 2

1. **Hypothesis**

- H0: The mean income of the group with 16 years of education is the same as the mean from the group with more than 16 years of education.
- Ha: The mean income of the group with 16 years of education is **not** the same as the mean from the group with more than 16 years of education.

2. **Find Critical value**

   The critical value associated with an alpha of 0.05 and 2579 degrees of freedom (from ANOVA in questions 1) is 1.961. The code used to find this critical value is shown below.

```
data critVAl;
        crit = quantile("T", 0.975, 2579);
run;
```

3,4. Analysis

The t-Value associated with this extra sum of squares test is -1.51 with an associated p-value of 0.1307. There were 2579 degrees of freedom used in this test.

5. **Fail to Reject the Null Hypothesis**
6. **Conclusion**

   There is not sufficient evidence to suggest that there is a difference in mean income for education attainment groups with 16 years of education and more than 16 years of education.

Code used to Answer Question 2

```
proc sort data=work.logedu out=logSort; by educ; run;

data critVAl;
     crit = quantile("T", 0.975, 2579);
run;

proc ttest data=work.logsort;
     class educ;
     var logIncome;
run;

proc glm data=work.logsort order=data;
     class educ;
     model logIncome = educ;
     estimate 'Estimate of BYOA' educ 0 0 1 0 -1;
run;
```

Question 3

Problem:

“How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others?”

Assumptions

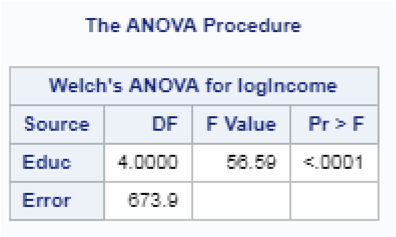
I am assuming that the standard deviation is not the same across populations. Please see question 1 for the graphs used during exploratory data analysis of this data.

Hypothesis

- H0: The means of the log transformed data are all the same
- HA: The means of at least one group is different from the other groups.

Analysis

In order to test data that does not meet the assumptions of equal variance/standard deviations I preformed the Welch’s ANOVA. The results of Welch’s ANOVA are as follows: F-value of 56.59, 4 degrees of freedom, and an associated p-value of <0.0001. The output of the test, and the code used to create it, are shown below.



The ANOVA Procedure

Welch's ANOVA for logIncome			
Source	DF	F Value	Pr > F
Educ	4.0000	56.59	<.0001
Error	673.9		

```
proc anova data=work.logedu;
            class educ;
            model logincome = educ;
            means educ / welch;
run;
```

Conclusion

There is strong evidence to reject the null hypothesis. This means that the mean income of one group is different from the other groups.

Scope of Interest

Since this is an observation study with random selection, it is valid to make generalization about the entire population.

Question 4

You could conduct the above tests on the data even if it were not log transformed. The t-tests and ANOVA tests are fairly robust to skewed data such as the data seen here (although the differences in sample size could still be concerning). If you are not log transforming the data then you won’t have to do any back transformation on the results; nor would you have to think about the multiplicative impact on the untransformed data either. Additionally, the results would be dealing only with means, you would not have to think about median values at all. In order to conduct these analysis you would still have to assume that variance is equal across the groups.

Appendix

QUESTION 1 SUPPLEMENTAL DATA

QQ-Plots of Original Data

Figure 2: 12 Years and 13-15 Years Education

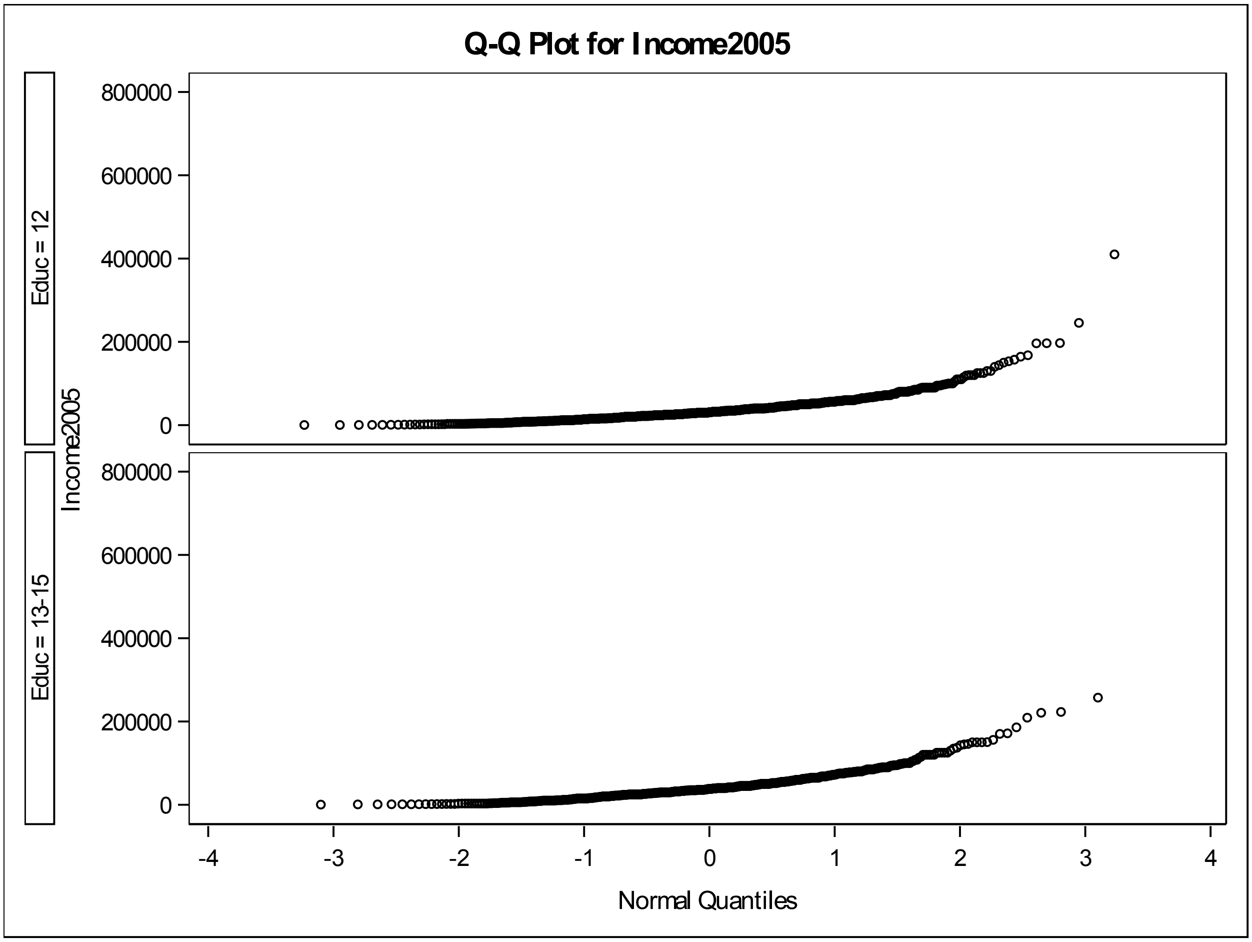


Figure 3: 16 Years and Less than 12 Years Education

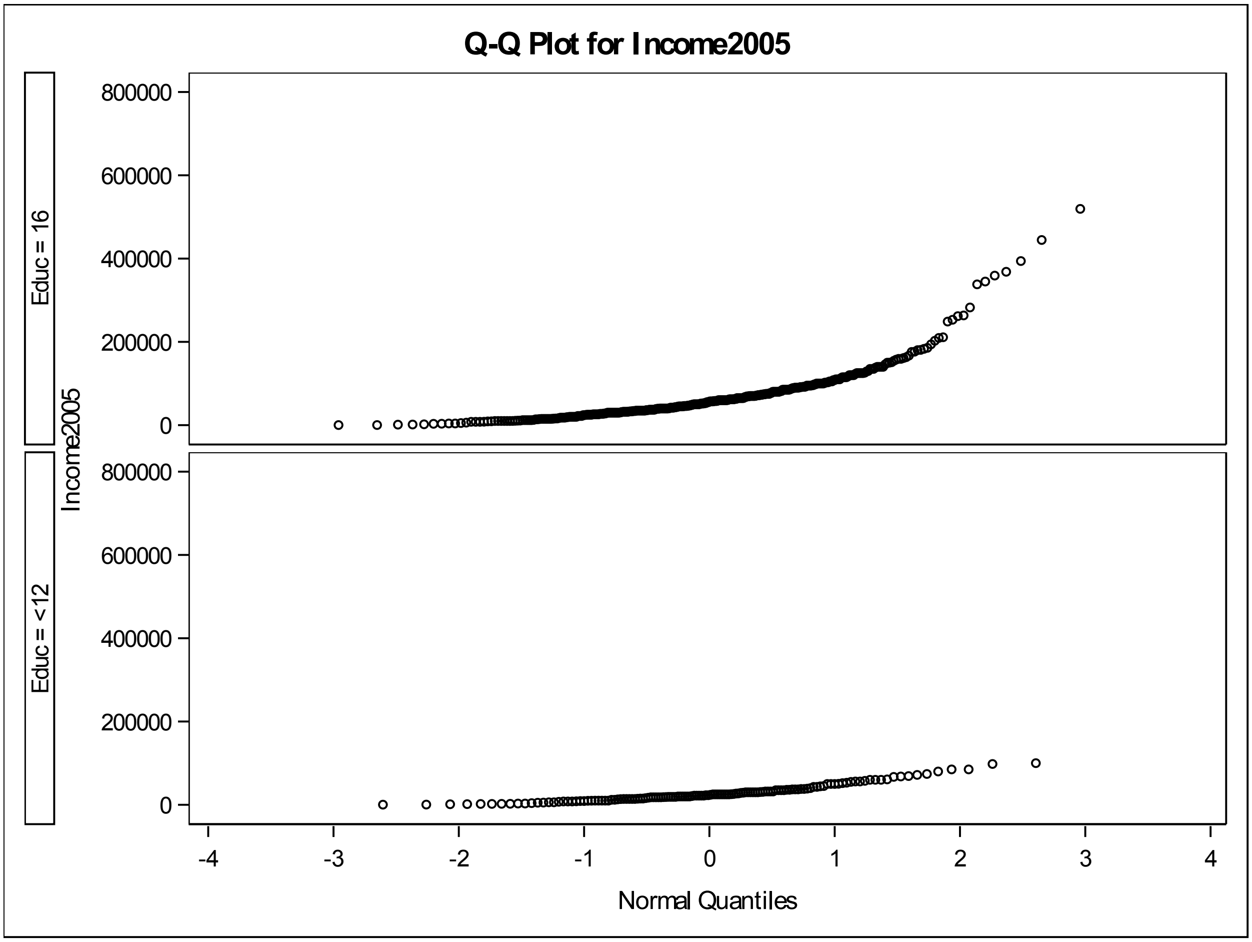
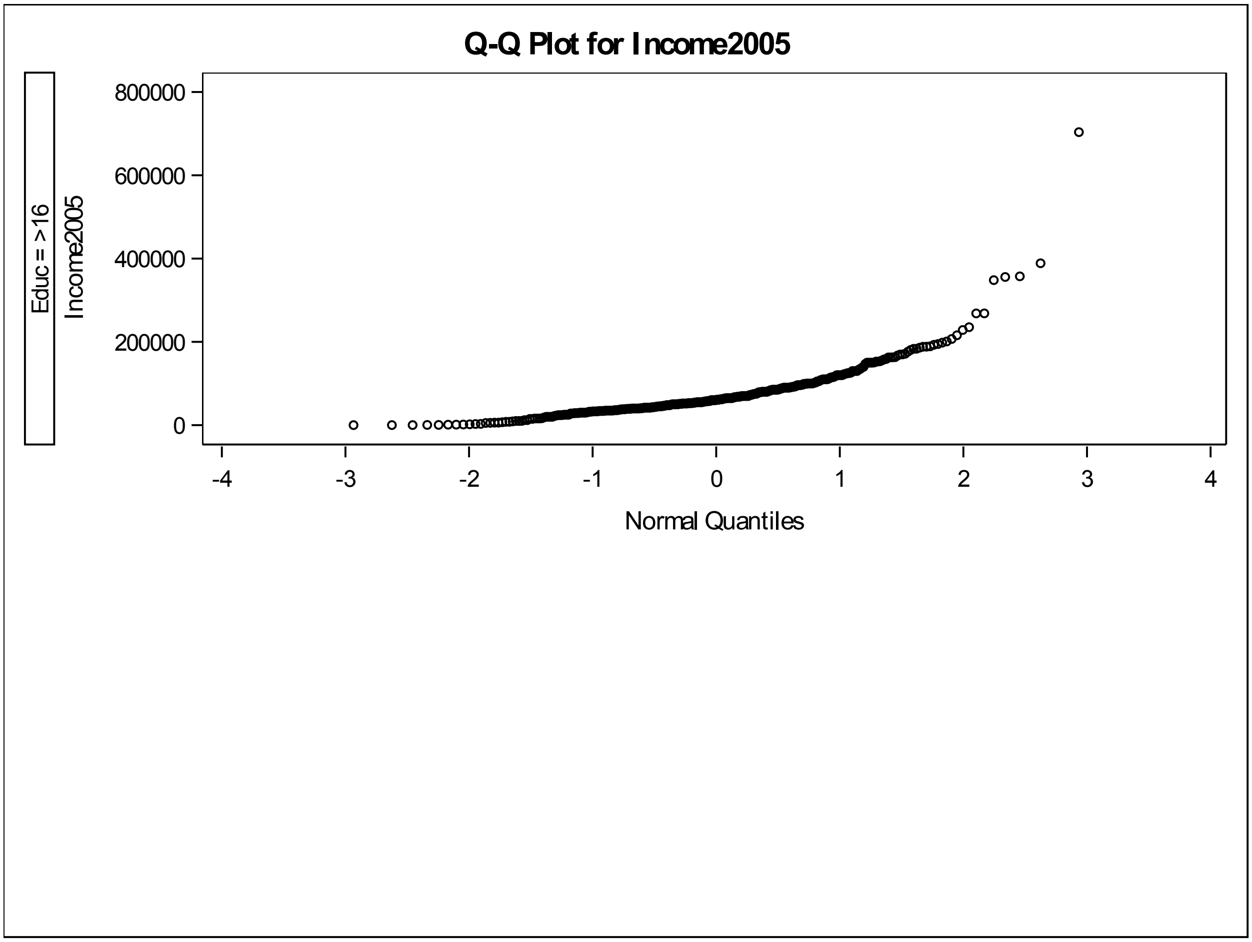
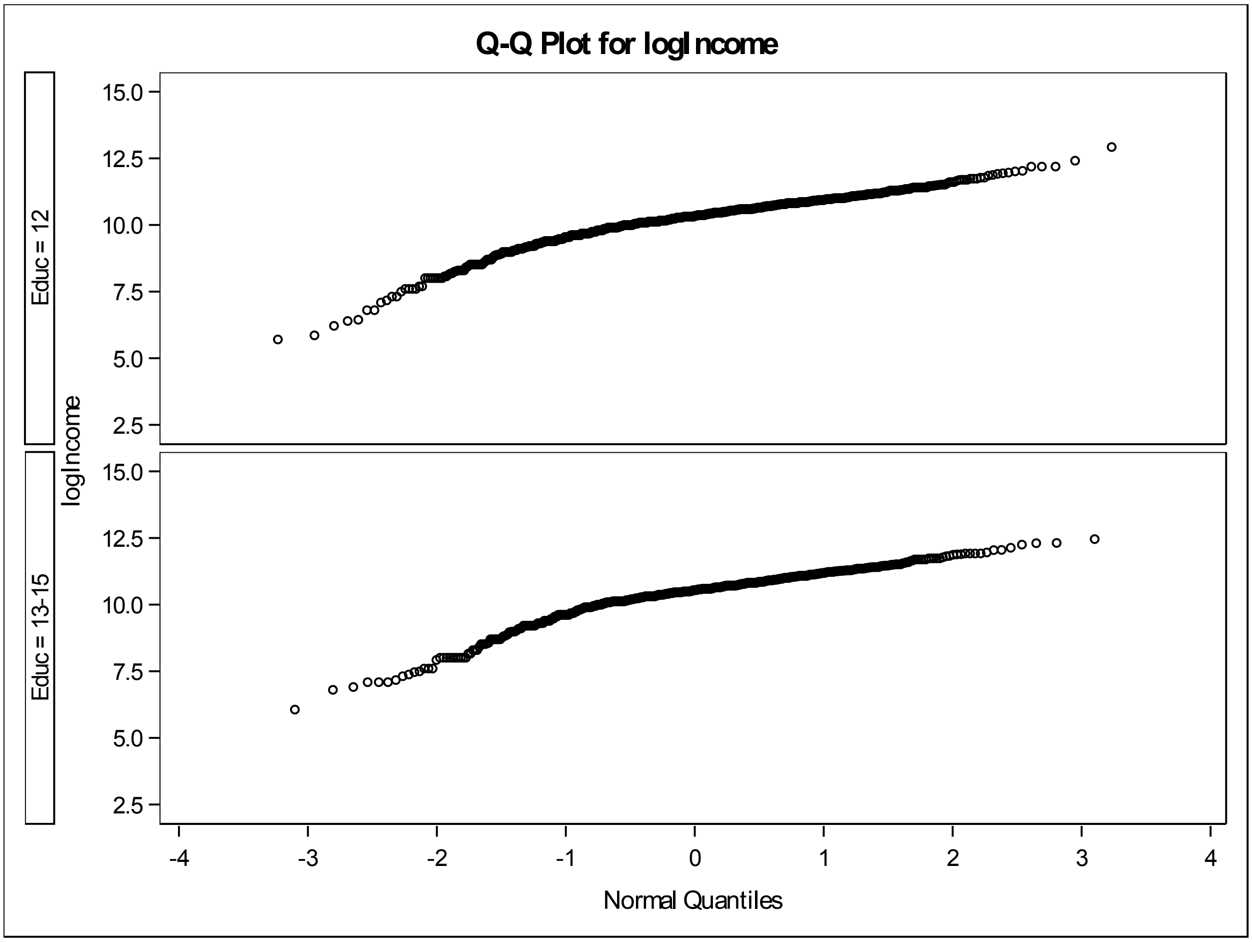
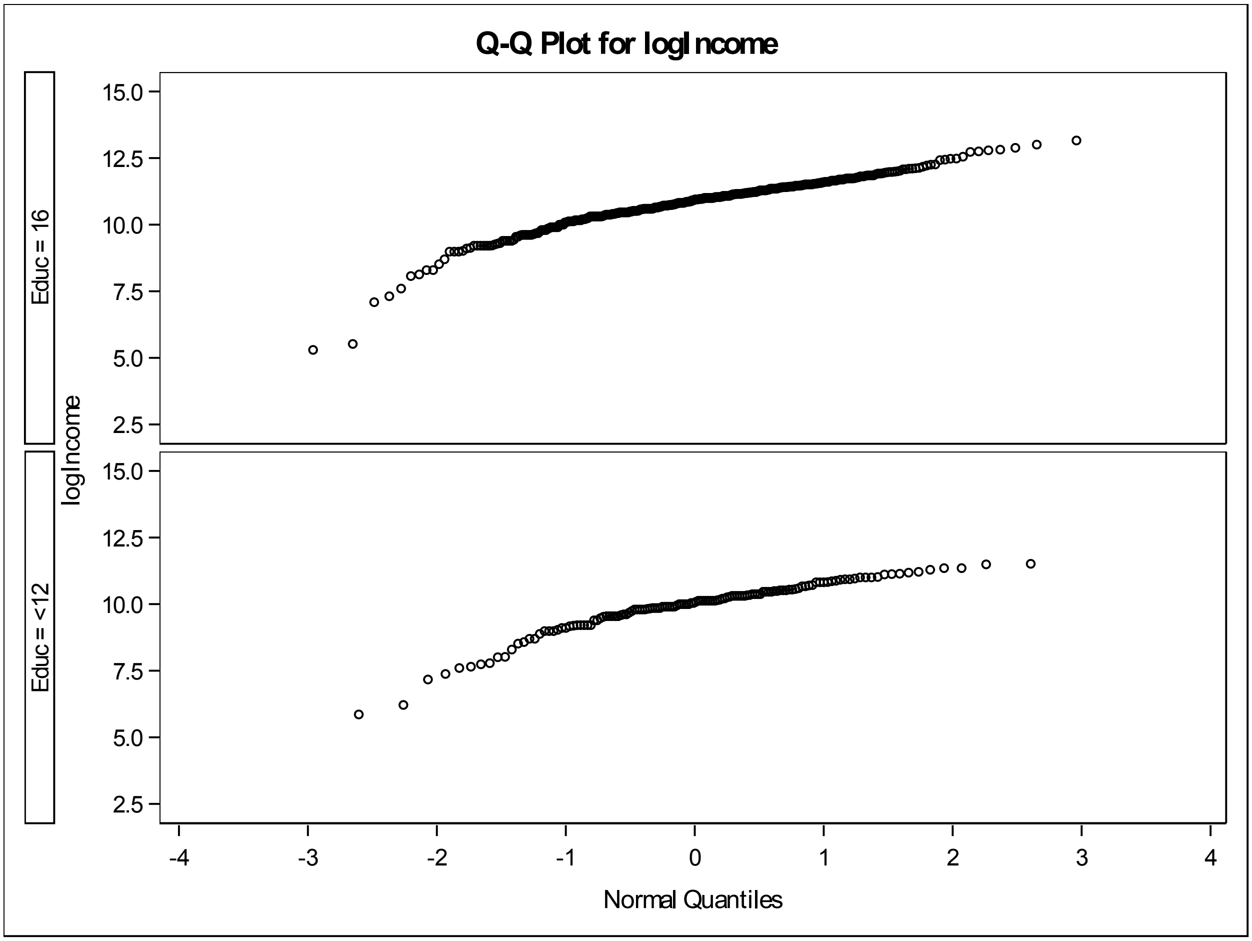


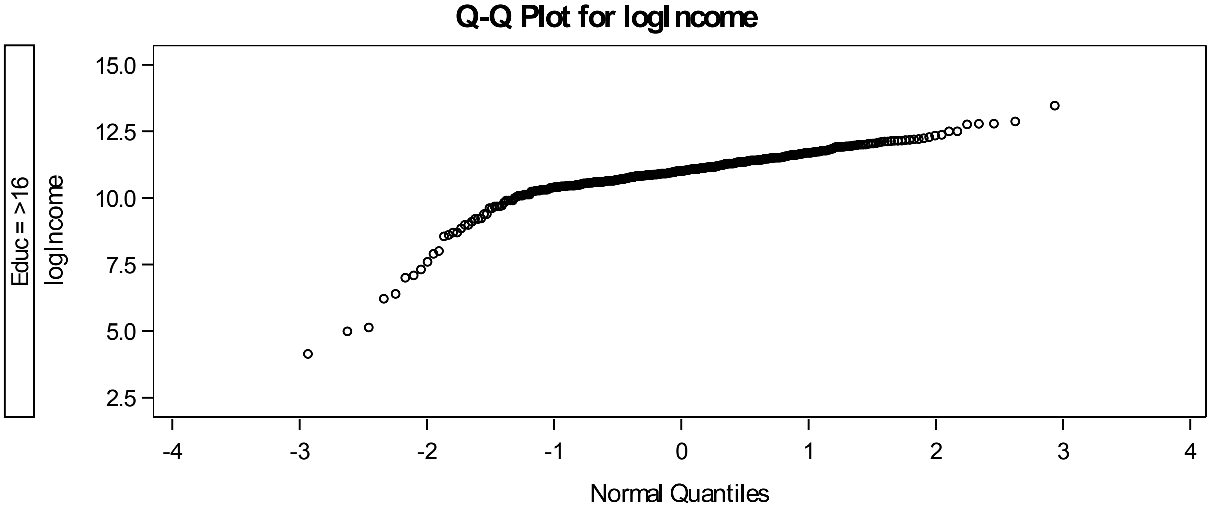
Figure 4: More than 16 Years Educaiton



QQ-Plots of Log Transformed Data




Q-Q Plot for logIncome
Educ = 16
Educ = <12
logIncome
15.0
12.5
10.0
7.5
5.0
2.5
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles



Code used to conduct the entire SAS analysis for Question 1, 2, and 3

```
/*STATS HW5*/
/*I used the data from the Sleuth3 R library and exported the data*/
filename ref "/folders/myshortcuts/SMU/MSDS6371 - Stats/Unit 
5/ex0525.csv";

proc import file=ref dbms=csv out=work.hw5;
      getnames=yes;
run;

/*set up data, both numerical and categorical labels*/
data work.education;
      set work.hw5;
      if educ = "<<12" then eduYears = "less than twelve";
      if educ = "<12" then eduYears = "tweleve";
      if educ = "13-15" then  eduYears = "13 to 15";
      if educ = "16" then eduYears = "sixteen";
      if educ = ">16" then eduYears = "more than 16";
      if educ = "<<12" then eduNum = 1;
      if educ = "<12" then eduNum = 2;
```

```
     if educ = "13-15" then  eduNum = 3;
     if educ = "16" then eduNum = 4;
     if educ = ">16" then eduNum = 5;
run;

/*log transform the data*/
data work.logEdu;
     set work.education;
     logIncome = log(Income2005);
run;

/*Exploritory data analysis on original data*/

/*Notes:
Data is very unnormal and heavily skewed at times*/
proc univariate data=work.education;
     class educ;
     var  Income2005;
     histogram Income2005;
     QQPLOT Income2005;
run;

proc sort data=work.education; by eduNum Income2005; run;

/* proc boxplot data=work.education; */
/*    by eduNum; */
/*    plot Income2005*eduNum; */
/* run; */

proc sgplot data=work.education;
     title "Disribution of Income by Educational Attainment";
     vbox Income2005 / group=Educ;
     yaxis label="Income ($)";
     format Income2005 dollar6.0;
run;




/*Exploritory data analysis on transformed data data*/
/*Notes:
     The transformation helps the normality assumption quite a
deal*/
proc univariate data=work.logEdu;
     class educ;
     var  logIncome;
     histogram logIncome;
     QQPLOT logIncome;
run;

proc sort data=work.logEdu; by eduNum logIncome; run;
proc boxplot data=work.logEdu;
     by eduNum;
     plot logIncome*eduNum;
run;

proc sgplot data=work.logedu;
     title "Disribution of Log Income by Educational Attainment";
     vbox logIncome / group=Educ;
     yaxis label="Log Income ($)";
     *format logIncome dollar6.0;
run;


/*Set up and run the ANOVA and GLM */

proc anova data=work.logEdu;
     class educ;
     model logIncome = educ;
run;


proc glm data=work.logedu;
     class educ;
     model logIncome = educ;
     lsmeans educ / stderr pdiff cl;
run;
/*CONCLUSION: At least one of the means is different based on the
logged income data */

/* ~~~~~~~~~~~~~ BYOA -- Build Your Own Anova -- HW Question 2
~~~~~~~~~~~~~~~~~~~~~~*/
/* data work.byoa; */
/*   set work.logedu; */
/*   if educ = "16" or educ = ">16"; */
/* run; */
/*  */
/* data work.byoa2; */
/*   set work.logEdu; */
/*   if(educ ne "16") & (educ ne ">16") then delete; */
/* run; */

proc sort data=work.logedu out=logSort; by educ; run;

data critVAl;
     crit = quantile("T", 0.975, 2579);
run;

proc ttest data=work.logsort;
     class educ;
     var logIncome;
```

```
run;

proc glm data=work.logsort order=data;
      class educ;
      model logIncome = educ;
      estimate 'Estimate of BYOA' educ 0 0 1 0 -1;
run;




/*~~~~~Question 3 - Kruskal Wallace test - non parametric ANOVA
~~~~~~~~~~~~*/
/*Is non-equal standard deviation messing things up? Lets find out.*/

/* first lets do a regular anova and see what happens */
proc glm data=work.logedu;
      class educ;
      model logincome = educ;
      means educ / hovtest=bf;
run;

/* Now run a non parametric test and see what happens */
proc npar1way data=work.logedu wilcoxon;
      class educ;
      var logincome;
run;

proc anova data=work.logedu;
      class educ;
      model logincome = educ;
      means educ / welch;
run;



/* Conclusion: */
/* There is no difference in the p-value between the normal ANOVA and
the non-parametric test.  */
/* The conclusion remains the same. */
/* This is rather unsurprising as the data passed earlier when the
assumptions were met with the */
/* transformed data */
```