Homework #5 Solutions

You earned 5 points just for turning in the assignment!

## Question 1 (35 points total)

Simply Answer Question 25 on pg. 147 from the Statistical Sleuth (read it!):

Plot the raw data, and also plot the data after a log transform. After a log transform, do the data satisfy the assumptions better? The data is in ex0525.csv or ex0525.xlsx. Perform this analysis in SAS. [Depending on where you find the data set, you may see the value **<<12**. Note that **<<12 = 12**.]

Also in SessionData.xlsx

Regardless of whether the assumptions of the original data or log transformed data are met, please include a **complete analysis** on the **log transformed** data.

1. State the Problem
2. Address the assumptions. Comment on each assumption (Use the visual test, as the Brown-Forsythe test will be overpowered due to the large sample size. This simply means that it is able to detect very small effect sizes-here, differences in standard deviations-which may not be big enough to practically affect the test). Comment on your thoughts of the assumptions, but, in the end, assume there is not enough visual evidence to suggest the standard deviations of the log transformed data are different.
3. Conduct the Test (an example is in the Unit 5 PowerPoint).
4. Write a conclusion (an example is in the Unit 5 PowerPoint).
5. State the Scope. (Can we generalize to the entire population or just the sample that was taken? Is there a causal relationship present?)

*Looking to the future! This is not an additional problem. Just FYI: The next step will be to look at these pairwise if we reject the to discover WHICH pairs have evidence of different means/medians.*

ADDITIONAL THINGS TO INCLUDE (for the logged data):

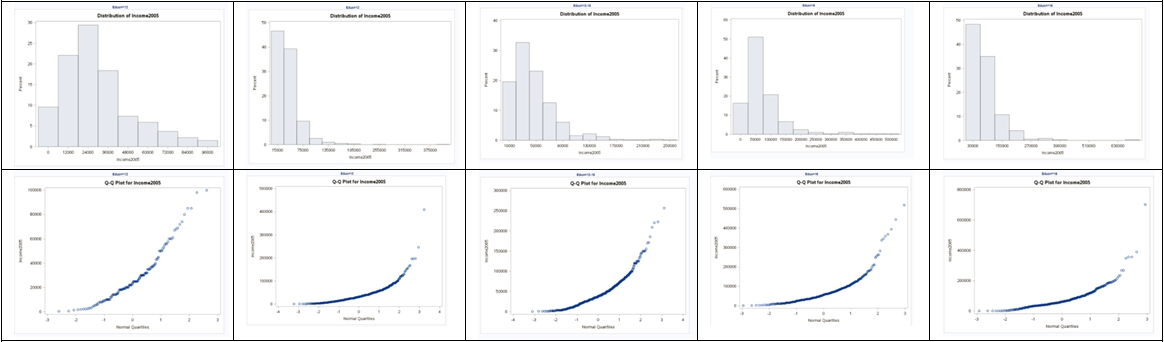
1. Please also identify .
2. Also specify the mean square error and how many degrees of freedom were used to estimate it.
3. Provide the code to perform the ANOVA in R and a screen shot of the output.

**Problem (1 point): How strong is the evidence that at least one of the five population distributions of education level has a different mean income than any of the others?**

**Assumptions: The Assumptions of the ANOVA are: the incomes in each educational group come from a normal distribution, the variances of these normal distributions are equal, the data are independent within each group, and the data are independent between each group.**

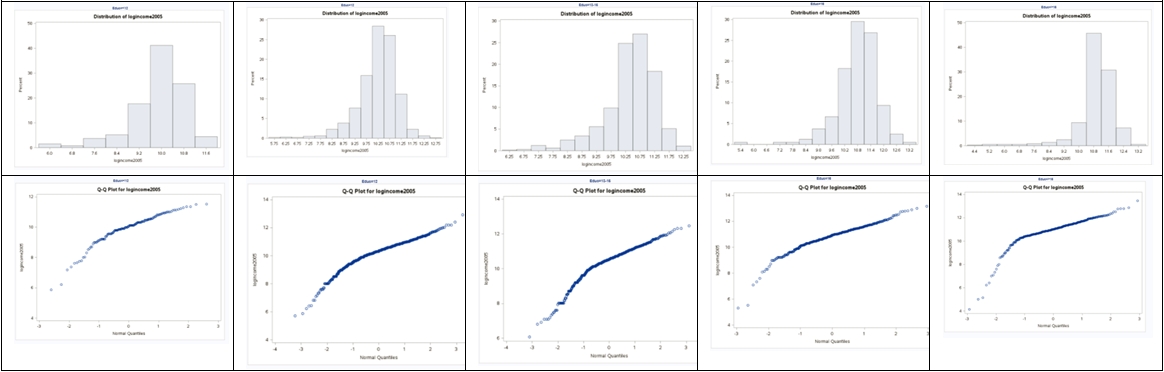
**Normality (3 points): The histograms and QQ plots below (of original data) appear to each show strong evidence of right skew, and thus provide evidence against coming from a normal distribution. This is not unexpected, as income data is often right skewed. However, each group has a sample size greater than 130, thus allowing the CLT to enable the ANOVA to be robust to this assumption. The log transformed data appears to be slightly less skewed (in the other direction), but only slightly.**

\*To address ANOVA assumptions on original data with histograms and QQ plots;  
proc univariate data = incomedata;  
by educ;  
histogram income2005;  
qqplot income2005;  
run;



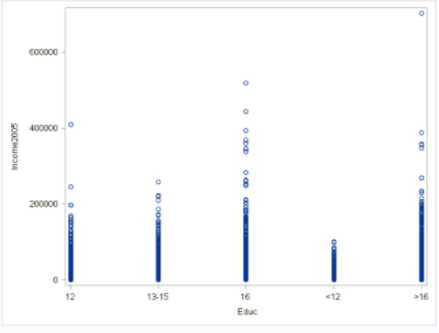
\*Perform a log transform;  
data incomedata;  
set incomedata;  
logincome2005 = log(income2005);  
run;

\*To address ANOVA assumptions on log transformed data with histograms and QQ plots;  
proc univariate data = incomedata;  
by educ;  
histogram logincome2005;  
qqplot logincome2005;  
run;

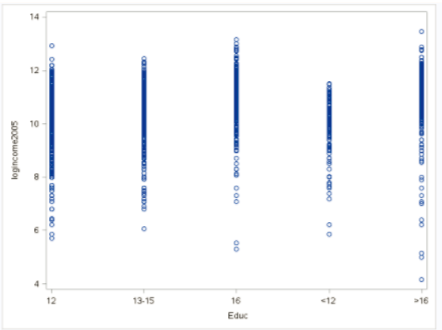


**Equal Standard Deviations (3 points): It appears that the original data shows evidence against equal standard deviations in the scatter plot. We were given in the problem that we are able to assume that the standard deviations between the groups are equal (homoscedasticity) for log transformed data. This is a safe assumption visually. Small deviations in sd will have less effect on the test than larger deviations. Remember, all models are wrong, but some are useful. [George Box]**

\*To address ANOVA assumptions on original data with scatter plots;  
proc sgplot data = incomedata;  
scatter x= educ y = income2005;  
run;



\*To address ANOVA assumptions on log transformed data with scatter plots;  
proc sgplot data = incomedata;  
scatter x= educ y = logincome2005;  
run;



**Independence (3 points): We will assume the data are independent, both between and within groups, and proceed with the ANOVA to test for differences in mean log income (median income) between the five levels of education. Note: this is risky assumption, as it turns out the sample is a random sample of households in which all members of the household were recruited into the survey. More pertinent information on the sample can be found in the first paragraph of the “Sampling Procedures” section that can be found by following this link:** [**https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/sample-design-screening-process**](https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/sample-design-screening-process)**.**

**Step 1 - Hypotheses (2 points):**

**All median incomes are the same across education levels.**  
 **At least one pair of income medians are different between education levels.**

**Step 2 - Identification of Critical Value: You may skip step 2 (critical value) in ANOVA settings, although one could be found (and the comparison to the F statistic should match the p-value’s comparison to alpha).**

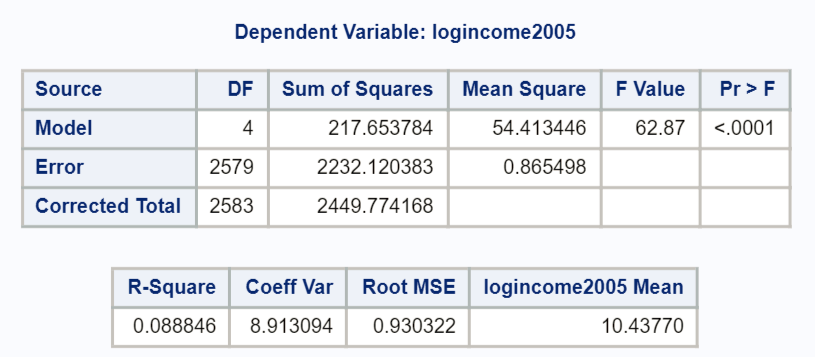
**Step 3 - Value of Test Statistic (2 points):**

**Step 4 - Give p-value (2 points):**

**Step 5 - Decision (2 points): Reject**

**Step 6 - Conclusion (5 points): There is strong evidence to suggest that at least one of the median incomes (median, not mean, because we used a log transform) for a particular education level is different from the others ( from a pure ANOVA).**

\*To perform ANOVA on log transformed data;  
proc glm data = incomedata;  
class educ;  
model logincome2005 = educ;  
run;

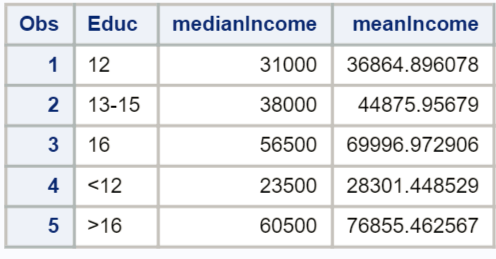


##Here is how to answer the problem using R  
##Read in the data, note your directory will be different  
  
edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 5/HW/ex0525.csv')  
  
edu$log.income <- log(edu$Income2005)  
  
edu.anova <- aov(log.income ~ Educ, data=edu)  
summary(edu.anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## 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

**To answer the second part of the question in the textbook, we will look at the difference of means here given the table, although a difference in medians could also be provided. Several SAS procedures will produce means and medians, including proc univariate (as coded above).**

\*There are many possibilities to produce means and medians… you can get even fancier code that will compute the percentage difference between each consecutive jump in education;  
proc means data = incomedata nway;  
class educ;  
var income2005;  
output out = incomesummary median = medianIncome mean = meanIncome;  
run;  
proc print data = incomesummary;  
var Educ medianIncome meanIncome;  
run;



* **The differences are as follows:**  
  + **(1 point) The estimated difference in mean income between those with less than a high school education and those with a high school degree only is $8,563.45 ($36,865.90 - $28,301.45), which is a 30.3% ($8,563.45/$28,301.45) increase in means from less than high school to only high school levels of education. The estimated difference in median incomes between those with less than a high school education and those with a high school degree only is $7,500 ($31,000 - $23,500), which is a 31.9% ($7,500/$23,500) increase in medians from less than high school to only high school levels of education.**
  + **(1 point) The estimated difference in mean income between those with only a high school education and those with only some college is $8,011, with a 21.7% increase in means from only high school to some college only. The estimated difference in median income between those with only a high school education and those with only some college is $7,000, with a 22.6% increase in medians from only high school to some college only.**
  + **(1 point) The estimated difference in mean income between those with only some college and those with only a college degree $25,121, with a 56.0% increase in means from only some college to only a college degree. The estimated difference in median income between those with only some college and those with only a college degree $18,500, with a 48.7% increase in medians from only some college to only a college degree.**
  + **(1 point) The estimated difference in mean income between those with only a college degree and those with more than 16 years of education (more than a college degree) is $6,858, with a 9.8% increase in means from only college to more than college. The estimated difference in median income between those with only a college degree and those with more than 16 years of education (more than a college degree) is $4,000, with a 7.1% increase in medians from only college to more than college.**

**Scope of Inference (5 points): this is an observational study, and thus, we cannot assign causal inference to this relationship. (Education does not necessarily cause the difference in income.) The NLSY is a random sample of households and, thus, is a random sample but not a simple random sample of subjects in the desired population. Inference can be generalized to the population of areas sampled in the United States, although one should be wary of the standard deviations and standard errors estimated here. Cluster sampling of households was employed, which introduces dependency/correlation at the cluster (household) level. We will address this adjustment/calculation later in the Sampling Course.**

**(1 points)**   
**(2 points)**

## Question 2 (30 points total)

2.Use thebachelor’s degree group (16) and the more than bachelor’s degree group (>16) to do the analysis below:

a. Evaluate if the Normality assumptions holds for the data.

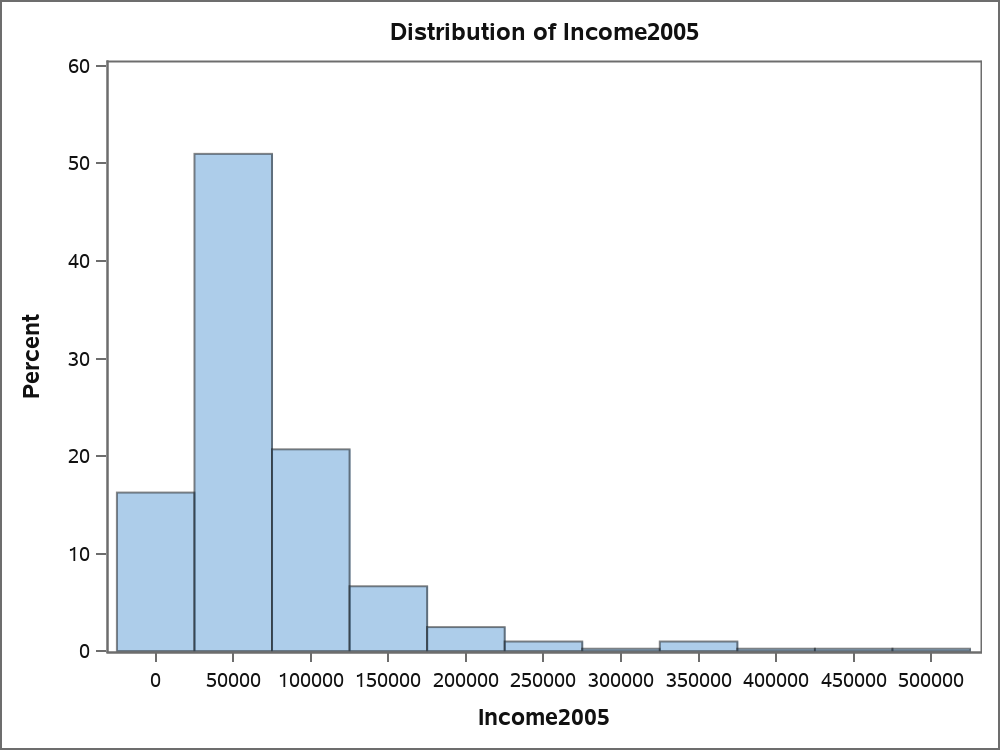
b. Evaluate if the equal standard deviation assumption holds for the data.

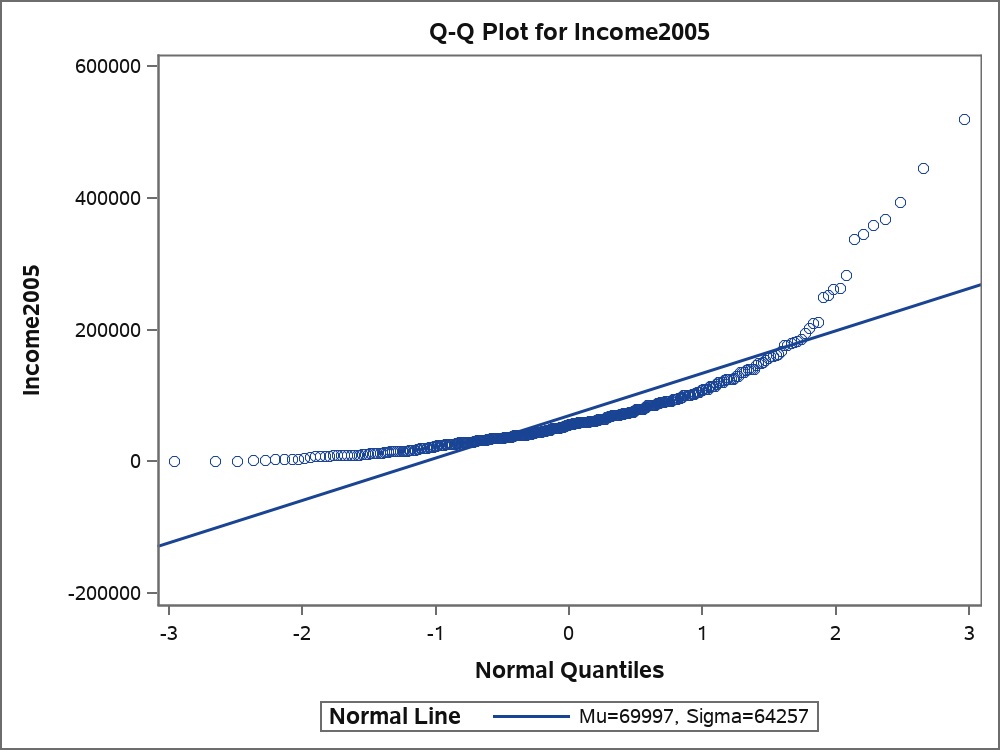
c. Why do any deviations from Normality in this data not restrict an ANOVA with this data?

d. Manually construct the ANOVA table showing all sums of squares and the F test.

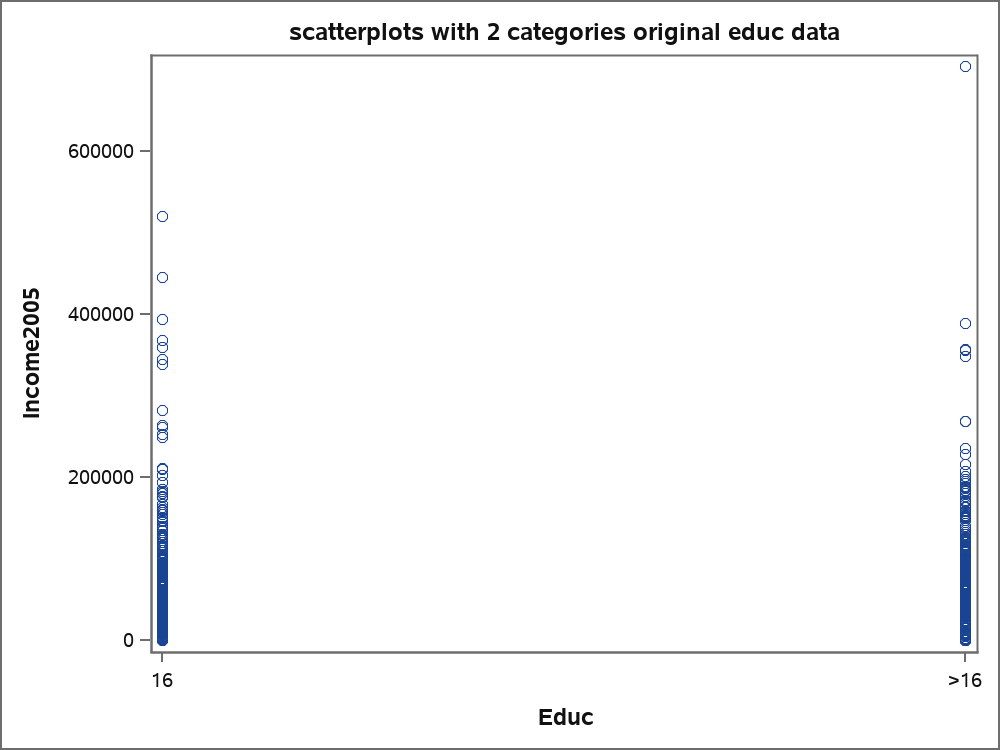
**e.** Compareyour final ANOVA table to that produced by SAS and do a 6-step complete analysis.

**a. Examine plots for Normality Assumption – Not Normally distributed –right skewed, but dataset is large so proceed as if Normally distributed by CLT**



****

**b. Looks like equal standard deviations so do not have to take logs.**



**c. This is a large data set so we can use the central limit theorem for normality of group means.**

**d.** 

| e. Brown and Forsythe's Test for Homogeneity of Income2005 Variance ANOVA of Absolute Deviations from Group Medians | | | | | |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Sum of  Squares | Mean  Square | F Value | Pr > F |
| Educ | 1 | 33830522 | 33830522 | 0.01 | 0.9129 |
| Error | 778 | 2.201E12 | 2.8288E9 |  |  |

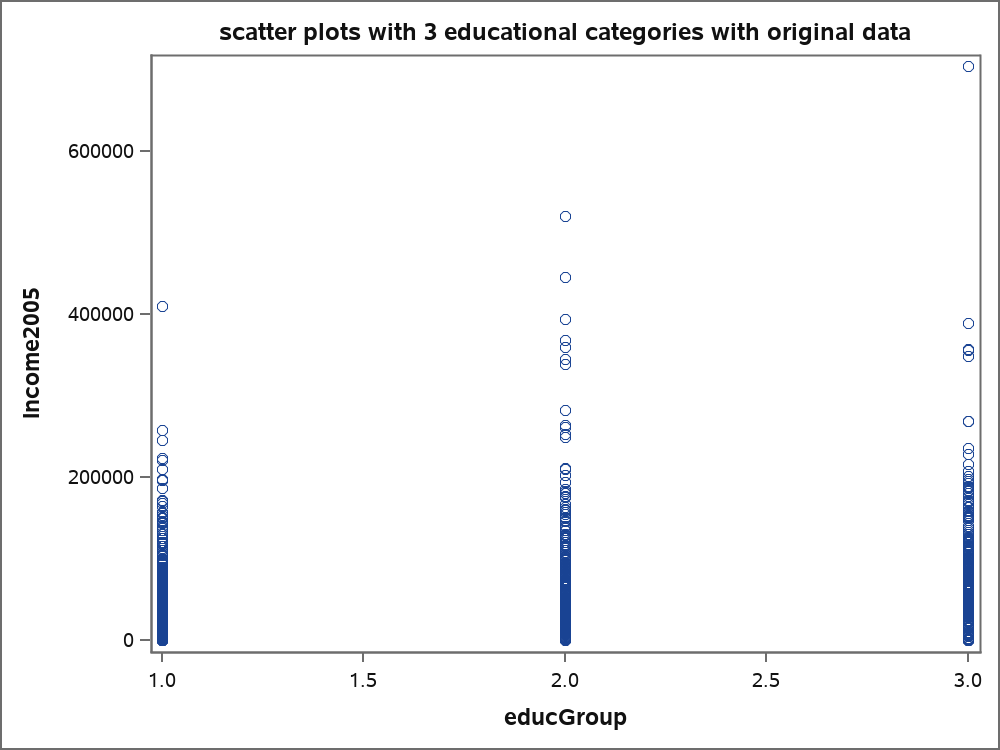
| Source | DF | Sum of  Squares | Mean Square | F Value | Pr > F |
| --- | --- | --- | --- | --- | --- |
| Model | 1 | 9157143280.6 | 9157143280.6 | 2.18 | 0.1403 |
| Error | 778 | 3.2689807E12 | 4201774622.9 |  |  |
| Corrected Total | 779 | 3.2781378E12 |  |  |  |

## Question 3 (30 points total)

3. Now add the subjects with < 16 years of education to the data set you used in question 2, so there are three educational levels. Conduct an ANOVA with SAS or R and show all comparisons between the three groups. Manually confirm all confidence intervals between groups using the T approach to comparisons of means in ANOVA. . For the manual calculations you will need to find the means of the three groups and the number of observations in each group.

**Because this is a large dataset we use the Central Limit Theorem for the Normality of group means.**

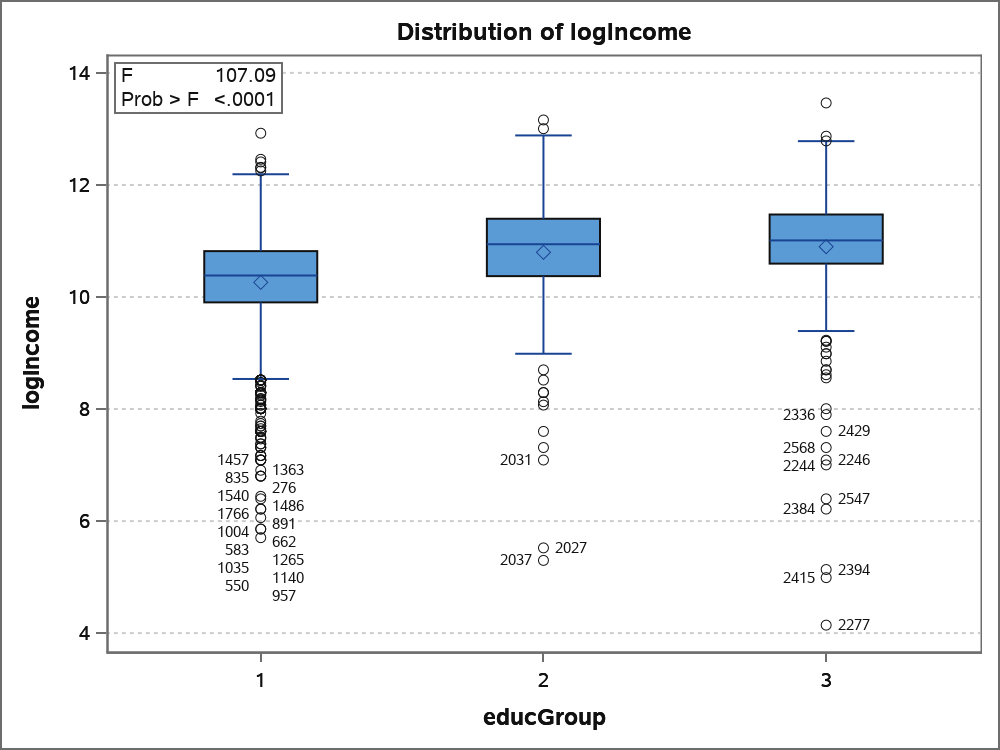
**We do not have equality of variances so see if taking logs helps.**



| Brown and Forsythe's Test for Homogeneity of Income2005 Variance ANOVA of Absolute Deviations from Group Medians | | | | | |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Sum of  Squares | Mean  Square | F Value | Pr > F |
| educGroup | 2 | 2.039E11 | 1.02E11 | 81.07 | <.0001 |
| Error | 2581 | 3.246E12 | 1.2578E9 |  |  |

**Taking logs does help adjust for equality of variances so do analysis with log of Income2005.**

| Brown and Forsythe's Test for Homogeneity of logIncome Variance ANOVA of Absolute Deviations from Group Medians | | | | | |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Sum of  Squares | Mean  Square | F Value | Pr > F |
| educGroup | 2 | 0.5446 | 0.2723 | 0.59 | 0.5554 |
| Error | 2581 | 1194.8 | 0.4629 |  |  |



**Analysis using Anova and t-tests with log of Income2005.**

| Source | DF | Sum of  Squares | Mean Square | F Value | Pr > F |
| --- | --- | --- | --- | --- | --- |
| Model | 2 | 187.714612 | 93.857306 | 107.09 | <.0001 |
| Error | 2581 | 2262.059556 | 0.876428 |  |  |
| Corrected Total | 2583 | 2449.774168 |  |  |  |

**There are differences between groups 3 and 1, 2 and 1, 1 and 3, and 1 and 2 from looking at confidence intervals using t-values. Group 1 is < 16 group, Group 2 is 16 group, and Group 3 is > 16 group.**

**NOTE: For R code for questions 2 and 3, see file session5.R. Replace data set and values for the response variables with logs as appropriate.**

