Homework #5 Solutions

## 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.

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):

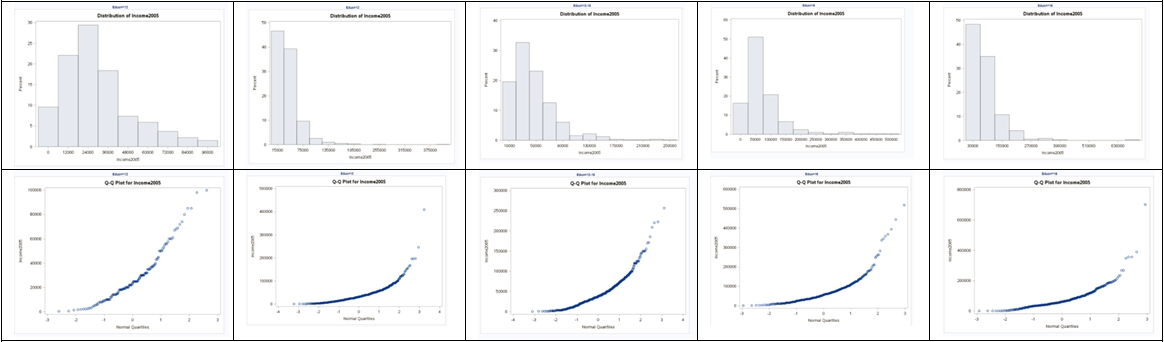
1. Please also identify .
2. Also specify the mean square error and how many degrees of freedom were used to estimate it.
3. Provide the code to perform the ANOVA in R and a screen shot of the output.

**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.**

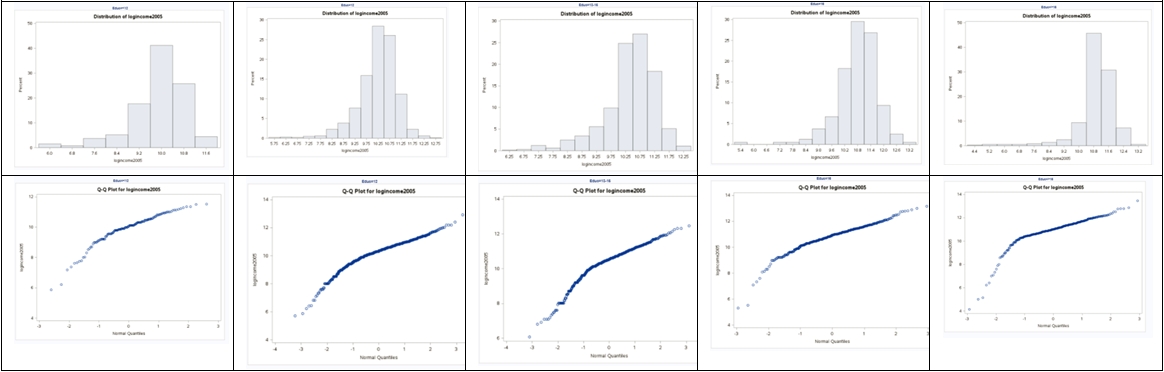
**Normality (3 points): The histograms and QQ plots below (of original data) appear to each show strong evidence of right skew, and thus provide evidence against coming from a normal distribution. This is not unexpected, as income data is often right skewed. However, each group has a sample size greater 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;



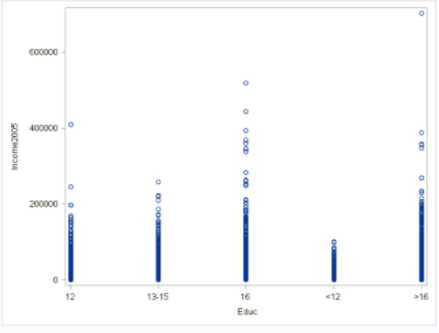
\*Perform a log transform;  
data incomedata;  
set incomedata;  
logincome2005 = log(income2005);  
run;

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

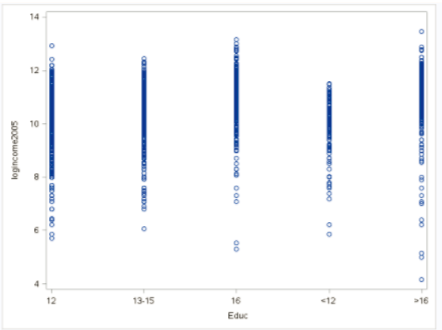


**Equal Standard Deviations (3 points): It appears that the original data shows evidence against equal standard deviations in the scatter plot. We were given in the problem that we are able to assume that the standard deviations between the groups are equal (homoscedasticity) for log transformed data. 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).**

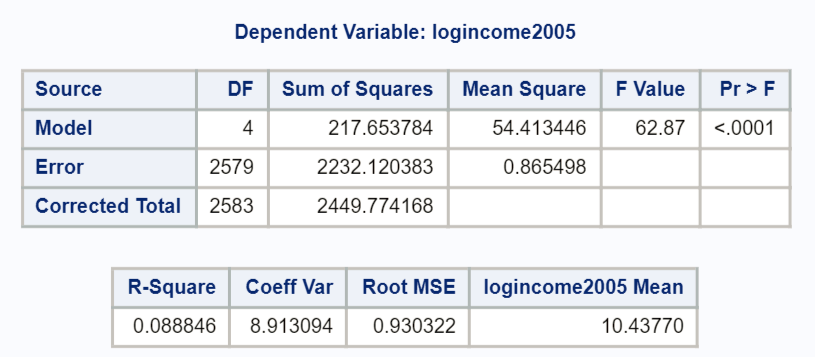
**Step 3 - Value of Test Statistic (2 points):**

**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;

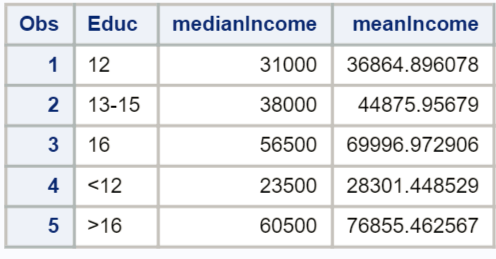


##Here is how to answer the problem using R  
##Read in the data, note your directory will be different  
  
edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 5/HW/ex0525.csv')  
  
edu$log.income <- log(edu$Income2005)  
  
edu.anova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

Use an extra sum of squares F-test (BYOA: Build Your Own ANOVA!) to use all the data (to increase the degrees of freedom and thus the power of the test) to compare only the bachelor’s degree group (16) income to the more than bachelor’s degree group (>16) income. Show your final ANOVA table and your 6-step complete analysis. You will need to assume that the standard deviations of the log-transformed data are again equal to proceed here. A two-sample t-test between these two groups (assuming equal standard deviations on logged data) yields a p-value of **0.1648** (try it!), but it only uses 778 degrees of freedom (from a pooled t-test). Make note again of how many degrees of freedom were used to estimate the pooled standard deviation in your extra sum of squares test. You may use SAS or R.

**Problem (1 point): Test whether there is a difference in income between those with a bachelor’s degree and those with more than a bachelor’s degree.**

*Note: alternatively, you could list the full and reduced model as is stated below:*

**Full Model:**   
**Reduced Model:**

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

**Step 2 - Identification of Critical Value: You may skip step 2 (critical value) in ANOVA settings, although the correct critical value is .**

**Step 3 - Value of Test Statistic (0 points, this is built in to the BYOA):**

**Step 4 - Give p-value (0 points, this is built in to the BYOA):**

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

**Calculation of the 95% CI:**

**Step 6 - Conclusion (5 points): There is not sufficient evidence to suggest that the median incomes of the bachelor’s degree group (16) and post bachelor’s degree group (>16) are different. A 95% confidence interval for this difference is . Now, a 95% CI for the multiplicative change in median incomes between these two groups is . Note that the multiplicative change of 1 is within our interval, consistent with the decision to fail to reject the null hypothesis. We will get to this confidence interval in Chapter 6, but see if you can reverse engineer this!**

(The confidence interval above is not necessary for full credit, as it is covered later in chapter 6.)

**Scope of Inference (5 points): As the results are not significant, we do not need to discuss assigning a causal inference in this relationship. 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.**

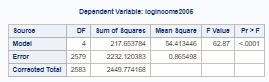
**DF Difference (5 points): note that in using the Extra Sums of squares, we were able to have 2579 degrees of freedom to estimate the pooled SD rather than only 778.**

**BYOA Table (10 points):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 1.9786 | 1.9786 | 2.286 | 0.13067 |
| Error | 2579 | 2232.120383 | 0.865498 |  |  |
| Total | 2580 | 2234.099010 |  |  |  |

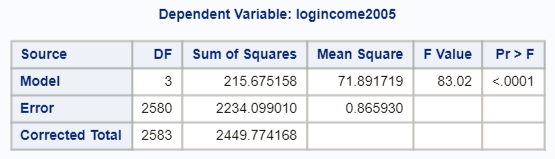
**SAS Code:**

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



\*Recode the data so that college and more than college are in the same group;  
Data incomedata;  
set incomedata;  
recodededuc = educ;  
if (educ = ‘>16’ | educ = ‘16’) then recodededuc = ‘16 or >16’;  
run;

\*To perform ANOVA on log transformed data (reduced model);  
proc glm data = incomedata;  
class recodededuc;  
model logincome2005 = recodededuc;  
run;



 SAS Code:

\*To find P-value from an F statistic and degrees of freedom (numerator and denominator) for building our own ANOVA;

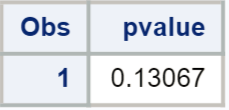
data pval;

pvalue = 1-probf(2.286, 1, 2579);

run;

proc print data = pval;

run;



\*To find critical value for an F test at alpha = 0.05;

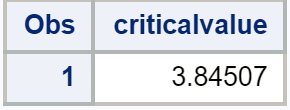
data critval;

criticalvalue = quantile("F", .95, 1, 2579);

run;

proc print data = critval;

run;



##Here is how to answer the problem using R  
##Combine the 16 and >16 groups  
edu$group.reduced <- ifelse(edu$Educ %in% c('>16', '16'), '16 plus',  
edu$Educ)  
  
##Full Model  
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

##Reduced Model  
edu.anova2 <- aov(log.income ~ group.reduced, data=edu)  
summary(edu.anova2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## group.reduced 3 215.7 71.89 83.02 <2e-16 \*\*\*  
## Residuals 2580 2234.1 0.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##Determine the critical value  
qf(0.95, 1, 2579)

## [1] 3.845067

##Determine the p-value  
pf(2.286, 1, 2579, lower.tail=F)

## [1] 0.1306685

## Question 3 (30 points total)

Now, suppose that you cannot assume the standard deviations are the same (for both the original or log transformed data). Conduct another complete analysis of the question in Chapter 5, problem 25 in Statistical Sleuth. Answer the question, “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?” This question should be answered in at least 1 or 2 sentences after providing a **complete analysis** without the assumption of equal standard deviations for the logged data (or for the original data). Perform the test in SAS or R.

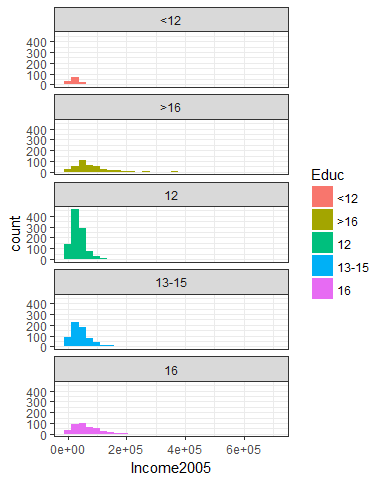
**Problem (3 points): How strong is the evidence that at least one of the five population distributions of education level has a different mean (or median, depending on your choice of test) income than any of the others?**

**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 independent between each group.**

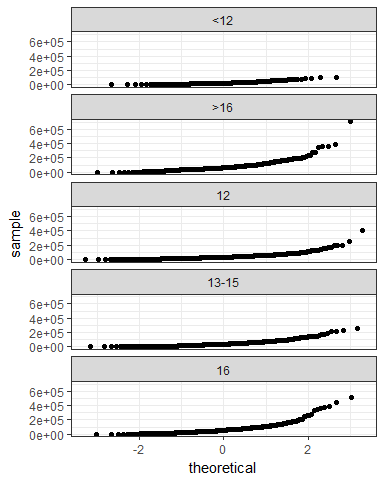
**Normality (3 points): this is exactly the same as in question 1 (plots below are generated using R, for reference).**

##There are multiple ways to do this, but the easiest is to use  
##the very popular ggplot package. First, install if necessary:  
##install.pacakges(ggplot2)  
  
library(ggplot2)   
  
##Histograms, Raw Data  
ggplot(data = edu, aes(x=Income2005)) +  
geom\_histogram(aes(fill=Educ)) +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

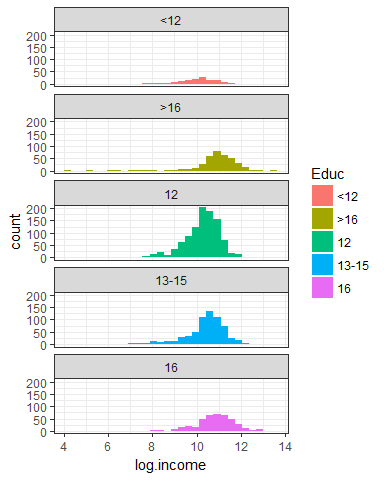


##QQ Plots, Raw Data  
ggplot(data = edu, aes(sample=Income2005)) +  
stat\_qq() +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

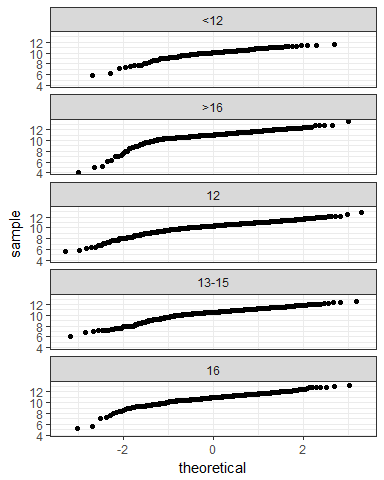


##Histograms, Logged Data  
ggplot(data = edu, aes(x=log.income)) +  
geom\_histogram(aes(fill=Educ)) +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

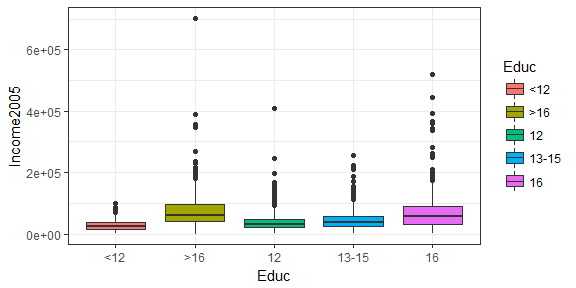


##QQ Plots, Logged Data  
ggplot(data = edu, aes(sample=log.income)) +  
stat\_qq() +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

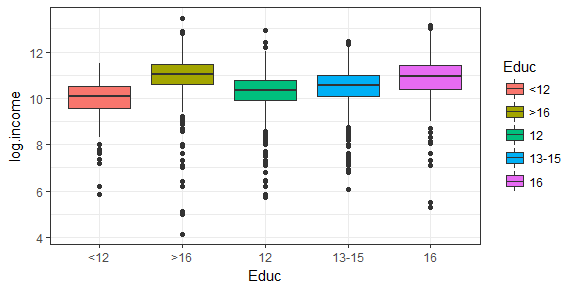


**Equal Standard Deviations (3 points): again, the plots are the same as in question 1 (plots below are generated using R, for reference). It appears that the original data show evidence against equal standard deviations in the scatter plot. We were given in the problem that we are NOT able to assume that the standard deviations between the groups are equal (homoscedasticity) for log transformed data.**

##Boxplots, Raw Data  
ggplot(data = edu, aes(x=Educ, y=Income2005)) +  
geom\_boxplot(aes(fill=Educ)) +  
theme\_bw()



##Boxplots, Logged Data  
ggplot(data = edu, aes(x=Educ, y=log.income)) +  
geom\_boxplot(aes(fill=Educ)) +  
theme\_bw()



**Independence (3 points): We will assume the data are independent, both between and within groups.**

**Because the CLT applies but the standard deviation assumption is violated, we have a few options. First, a Welch’s ANOVA can be used on the original data, or the nonparametric Kruskal-Wallis test can be used on the original data. Technically, a Welch’s ANOVA could be used on log transformed data, but that makes little sense if we still can’t overcome the equal standard deviation assumption violation by logging the data. Logging would only “fix” the normality issue, and with such large sample sizes, the sample means are likely to be normally distributed via the Central Limit Theorem. Still, it is a judgement call.**

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

**Option 1: Welch’s ANOVA**  
 **All mean incomes are the same across education levels.**  
 **At least one pair of income means are different between education levels.**

**Option 2: Kruskal-Wallis**  
 **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: May be omitted here.**

**Step 3 - Value of Test Statistic (2 points): For Welch’s ANOVA: ; for Kruskal-Wallis:**

**Step 4 - Give p-value (2 points): for both Welch’s ANOVA and Kruskal-Wallis**

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

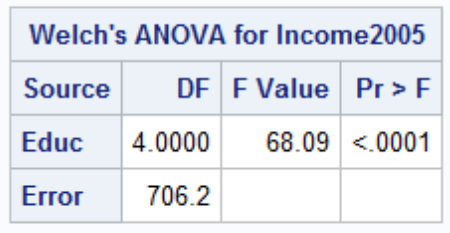
**Step 6 - Conclusion (5 points):**

**Welch’s ANOVA: At the level of significance, there is strong evidence () that the mean incomes between at least 1 pair of education levels are different.**

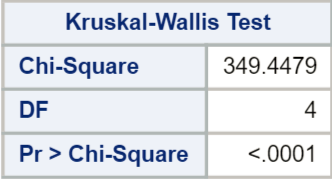
**Kruskal-Wallis: At the level of significance, there is strong evidence () that the median incomes between at least 1 pair of education levels are different.**

**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 dependence/correlation at the cluster (household) level. We will address this adjustment/calculation later in the Sampling Course.**

\*For Welch’s ANOVA on original data (code for assumptions was run in Q1);  
proc glm data = incomedata;  
class educ;  
model income2005 = educ;  
means educ / hovtest = bf welch;  
run;



\*For Kruskal-Wallis test on original data;  
proc npar1way data = incomedata;  
class educ;  
var income2005;  
run;



##Using R  
##Welch's ANOVA  
oneway.test(Income2005 ~ Educ, data=edu, var.equal=F)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Income2005 and Educ  
## F = 68.089, num df = 4.00, denom df = 706.18, p-value < 2.2e-16

##Kruskal-Wallis  
kruskal.test(Income2005 ~ Educ, data=edu)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Income2005 by Educ  
## Kruskal-Wallis chi-squared = 349.45, df = 4, p-value < 2.2e-16