Unit 6 HW

1. Simply Answer Question 25 on pg. 147 from the Statistical Sleuth!
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**.]

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 UNIT 6 PowerPoint.)
4. Write a conclusion. (An example is in the UNIT 6 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?)

ADDITIONAL THINGS TO INCLUDE (for the logged data):

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

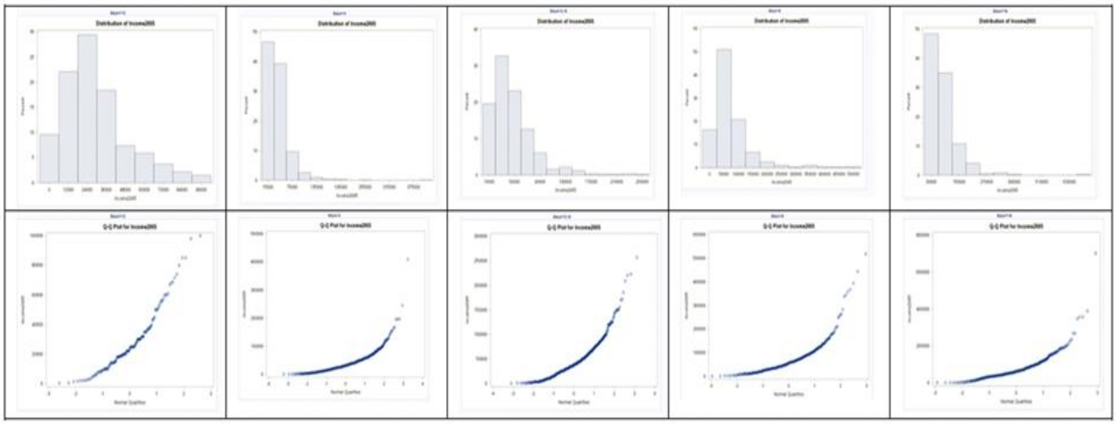
Looking to the future! This is not an additional problem. Just FYI: The next step will be to look at these pairwise if we reject the H_o to discover WHICH pairs have evidence of different means / medians.

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?

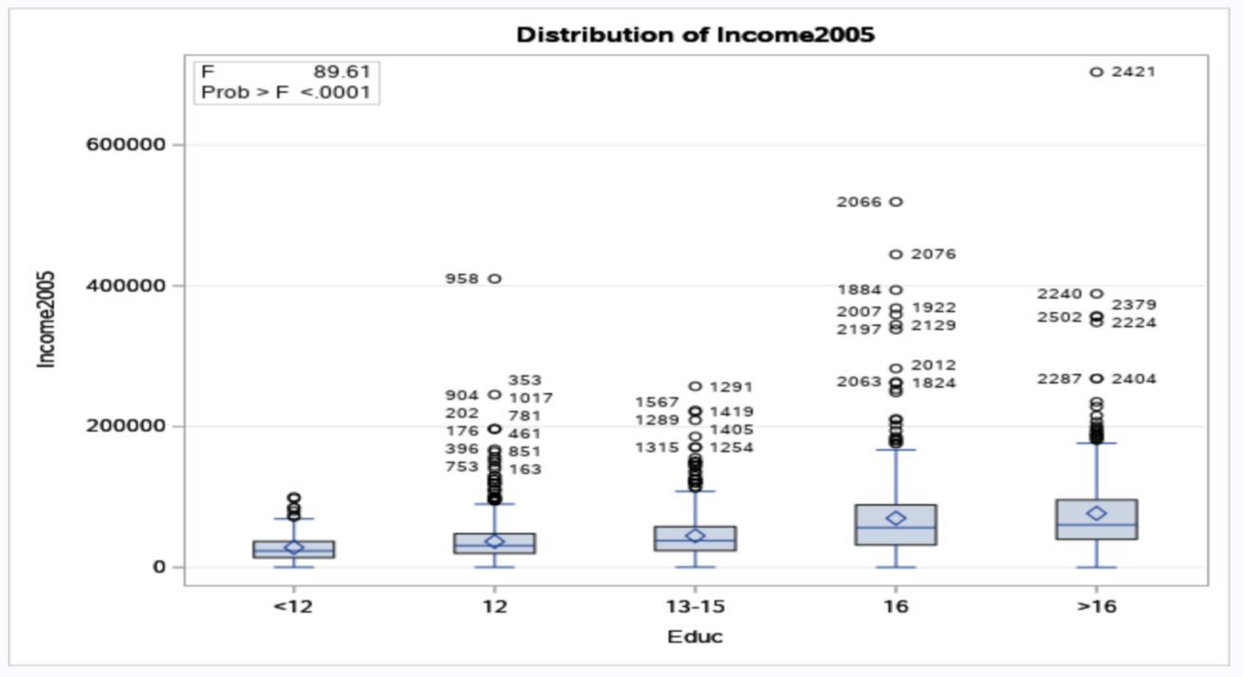
Assumptions: The Assumptions of the ANOVA are the incomes in each educational group come from a normal distribution, the variances of these normal distributions are equal, the data are independent within each group and the data re independent between each group.

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.

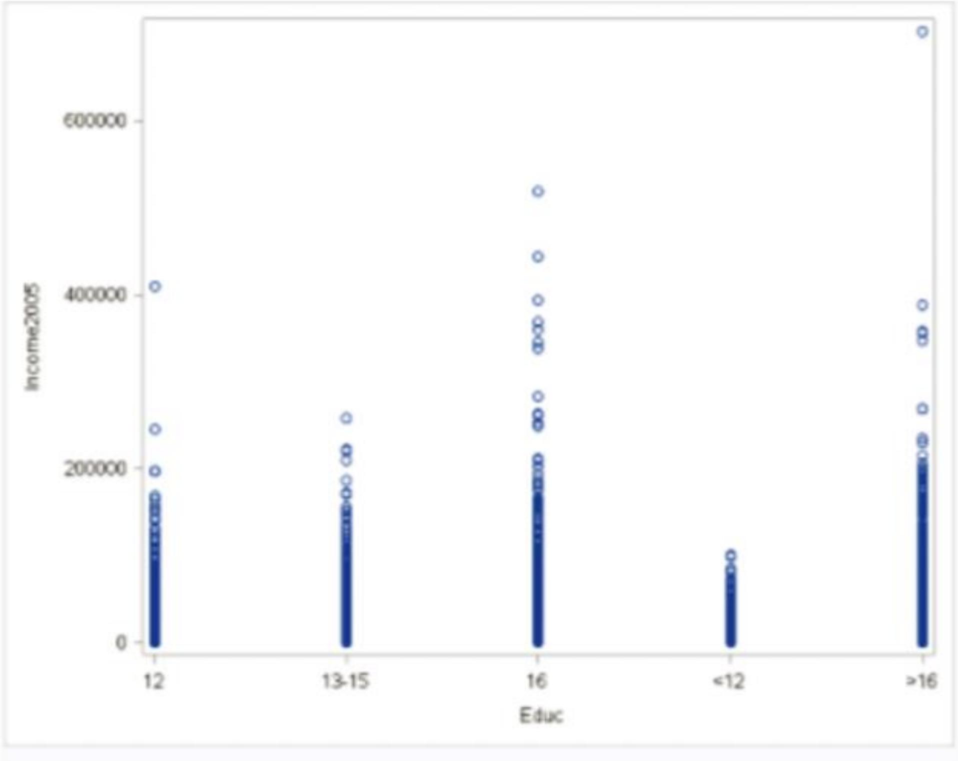
```
*To address ANOVA assumptions on original data with histograms and QQ plots;
proc univariate data = incomedata;
by educ;
histogram income2005;
qqplot income2005;
run;
```

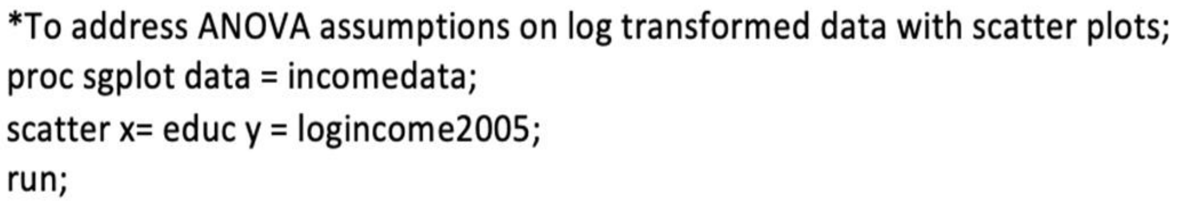


```
*ordering data by Educ by creating a column called EO;
data income;
set income;
if Educ eq '<12' then EO = 1;
 if Educ eq '12' then EO = 2;
 if Educ eq '13-15' then EO = 3;
 if Educ eq '16' then EO = 4;
 if Educ eq '>16' then EO = 5;
run;
proc sort data= income;
by EO;
run;
*ANOVA;
proc glm data = income order= DATA;
class Educ;
model income2005 = Educ;
run;
```

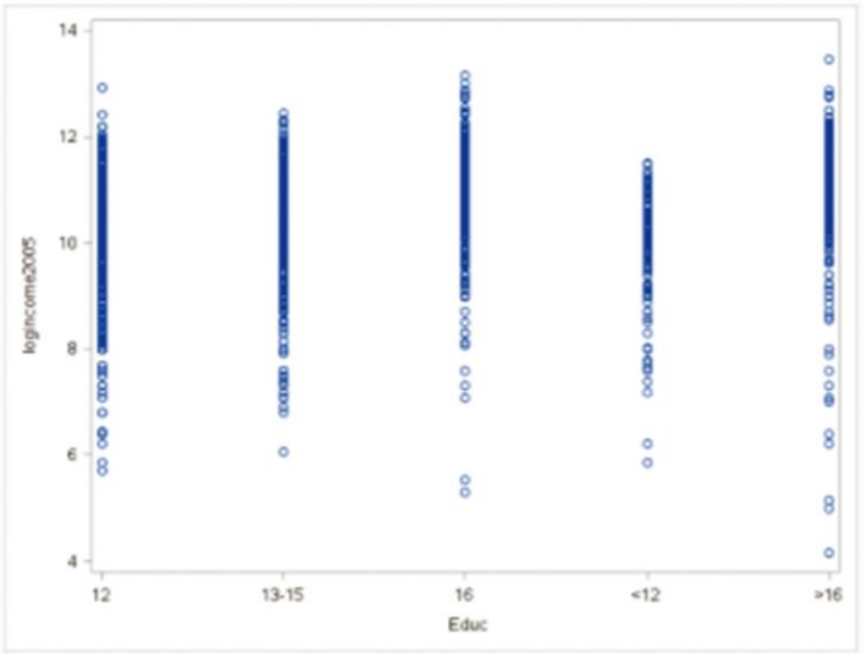


```
*To address ANOVA assumptions on original data with scatter plots;
proc sgplot data = incomedata;
scatter x= educ y = income2005;
run;
```





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



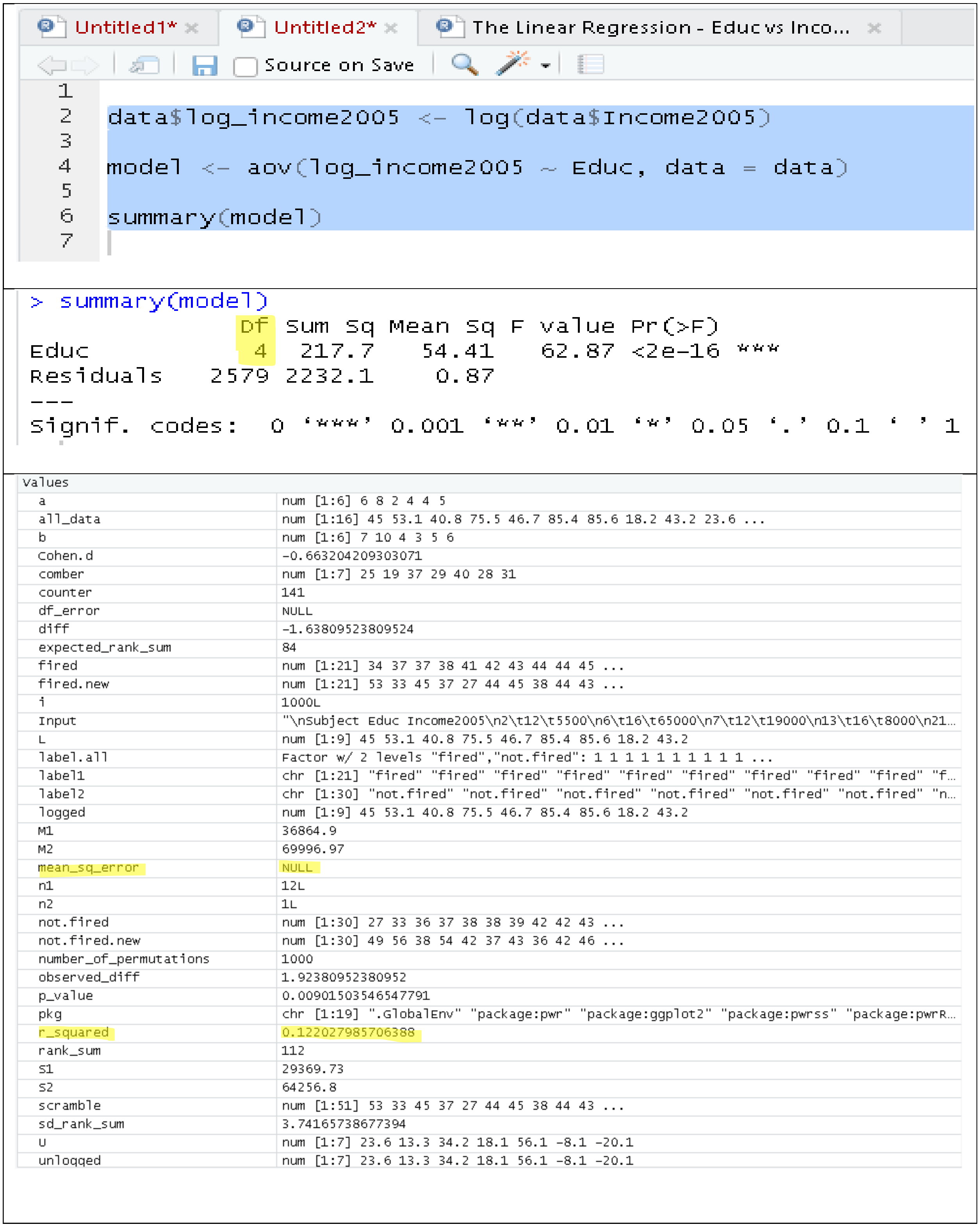
We will proceed with an ANOVA to assess differences in mean log income (or median income) across the five educational categories, operating under the assumption of data independence both between and within these groups. This is a somewhat precarious assumption given that the dataset comprises a random sample of households where all household members were included in the survey. Additional details about the sampling methodology can be found in the "Sampling Procedures" section, accessible through the following link: https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/sample-design-screening-process

Scope of the Analysis

The analysis was conducted on a dataset comprising a random sample of households, with all household members included in the survey. Due to this sampling method, caution should be exercised when generalizing these results to the broader population. The inclusion of entire households may introduce dependencies among observations, potentially violating the assumption of independence crucial for the validity of the ANOVA test applied in this study. It's important to note that while the ANOVA can identify statistically significant differences in mean income across various education levels, it does not establish a causal relationship. The observed differences indicate an association between education level and income, but they do not confirm that higher education causes

higher income. Additional studies, particularly those employing experimental or quasi-experimental designs, would be needed to make causal inferences.

Additional things to include:



```
data$log_income2005 <- log(data$Income2005)

model <- aov(log_income2005 ~ Educ, data = data)

summary(model)
```

```
> summary(model)
              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
```

Values	
a	num [1:6] 6 8 2 4 4 5
all_data	num [1:16] 45 53.1 40.8 75.5 46.7 85.4 85.6 18.2 43.2 23.6 ...
b	num [1:6] 7 10 4 3 5 6
Cohen.d	-0.663204209303071
comber	num [1:7] 25 19 37 29 40 28 31
counter	141
df_error	NULL
diff	-1.63809523809524
expected_rank_sum	84
fired	num [1:21] 34 37 37 38 41 42 43 44 44 45 ...
fired.new	num [1:21] 53 33 45 37 27 44 45 38 44 43 ...
i	1000L
Input	"\nSubject Educ Income2005\n2\t12\t5500\n6\t16\t65000\n7\t12\t19000\n13\t16\t8000\n21…
L	num [1:9] 45 53.1 40.8 75.5 46.7 85.4 85.6 18.2 43.2
label.all	Factor w/ 2 levels "fired","not.fired": 1 1 1 1 1 1 1 1 1 1 ...
label1	chr [1:21] "fired" "fired" "fired" "fired" "fired" "fired" "fired" "fired" "fired" "f…
label2	chr [1:30] "not.fired" "not.fired" "not.fired" "not.fired" "not.fired" "not.fired" "n…
logged	num [1:9] 45 53.1 40.8 75.5 46.7 85.4 85.6 18.2 43.2
M1	36864.9
M2	69996.97
mean_sq_error	NULL
n1	12L
n2	1L
not.fired	num [1:30] 27 33 36 37 38 38 39 42 42 43 ...
not.fired.new	num [1:30] 49 56 38 54 42 37 43 36 42 46 ...
number_of_permutations	1000
observed_diff	1.92380952380952
p_value	0.00901503546547791
pkg	chr [1:19] ".GlobalEnv" "package:pwr" "package:ggplot2" "package:pwrss" "package:pwrR…
r_squared	0.122027985706388
rank_sum	112
S1	29369.73
S2	64256.8
scramble	num [1:51] 53 33 45 37 27 44 45 38 44 43 ...
sd_rank_sum	3.74165738677394
U	num [1:7] 23.6 13.3 34.2 18.1 56.1 -8.1 -20.1
unlogged	num [1:7] 23.6 13.3 34.2 18.1 56.1 -8.1 -20.1

2. 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 **.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.

Step 1: State the Hypotheses:

Null Hypothesis (H0): The mean log-transformed income for the group with bachelor's degrees (16 years of education) is the same as for the group with more than a bachelor's degree (>16 years).
Alternative Hypothesis (H1): The means are not the same.

Step 2: Define the Analysis Plan:

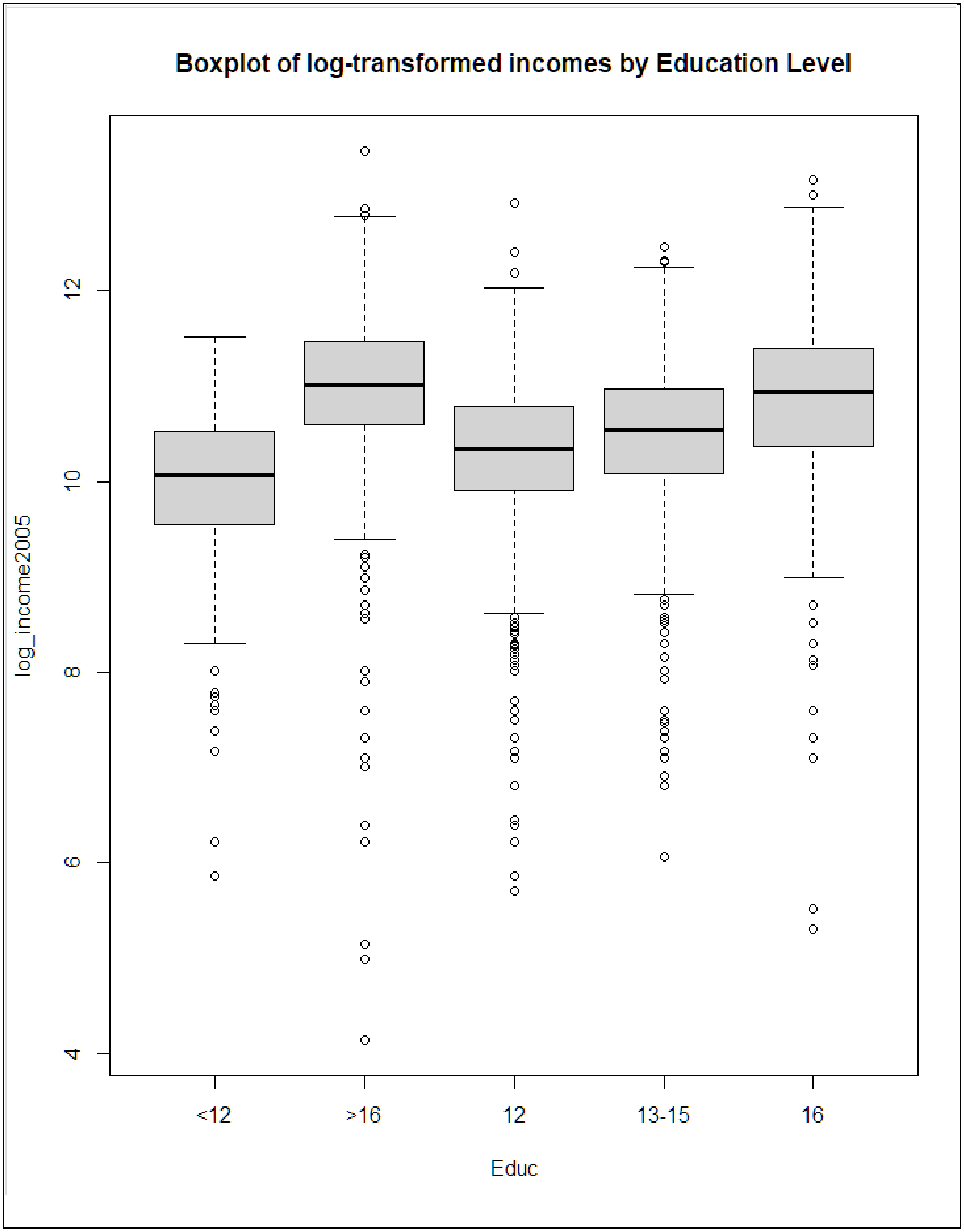
We'll use an F-test for comparing two variances, with a significance level of 0.05.

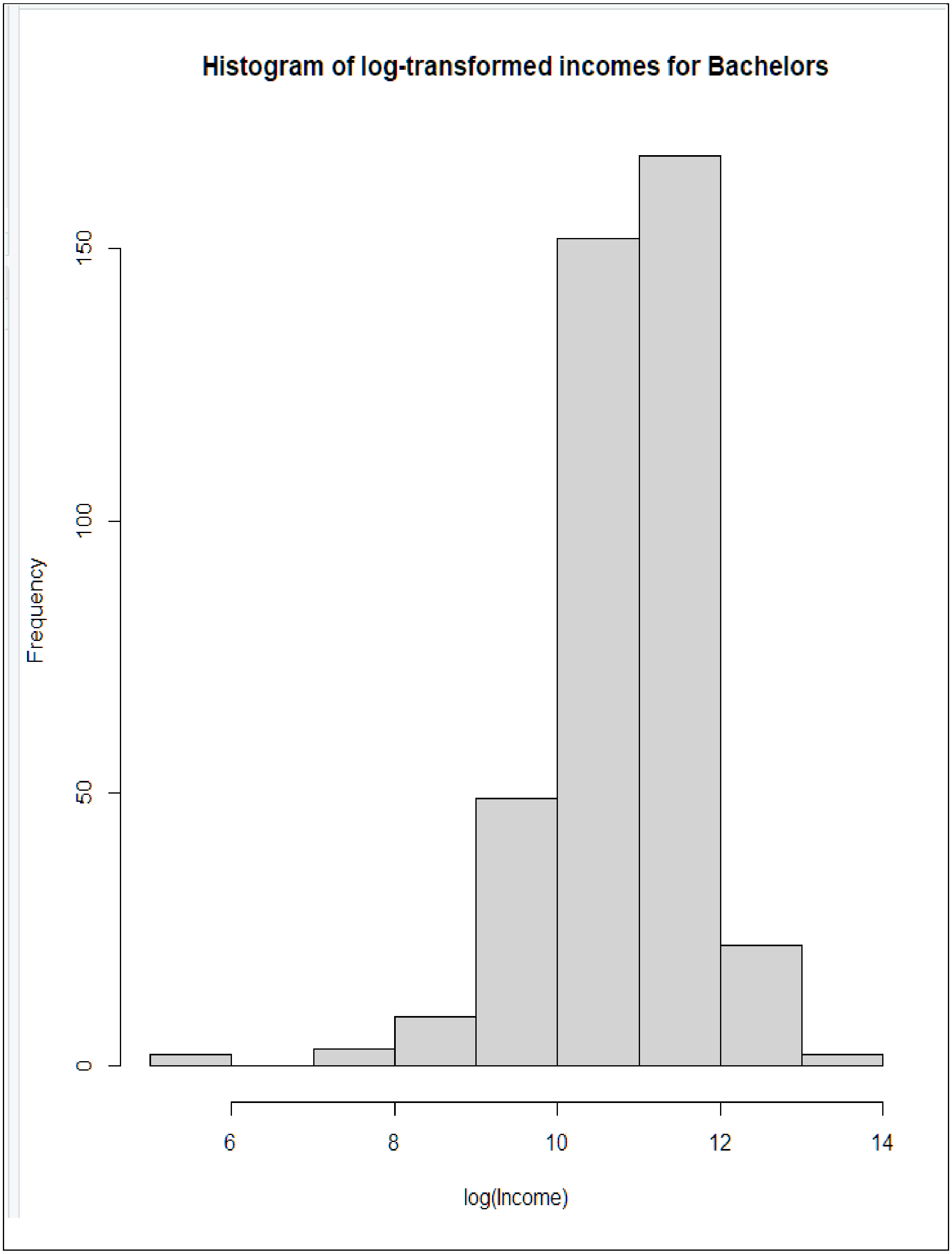
Step 3: Collect and Prepare Data:

Assuming data contains the log-transformed income (log_income2005) and education (Educ) variables.

Step 4 Perform Analysis:

```
summary(full_model)$r.squared
[1] 0.08884647
```


Boxplot of log-transformed incomes by Education Level
log_income2005
4
6
8
10
12
<12
>16
12
13-15
16
Educ



Step 5: Interpret the Results:

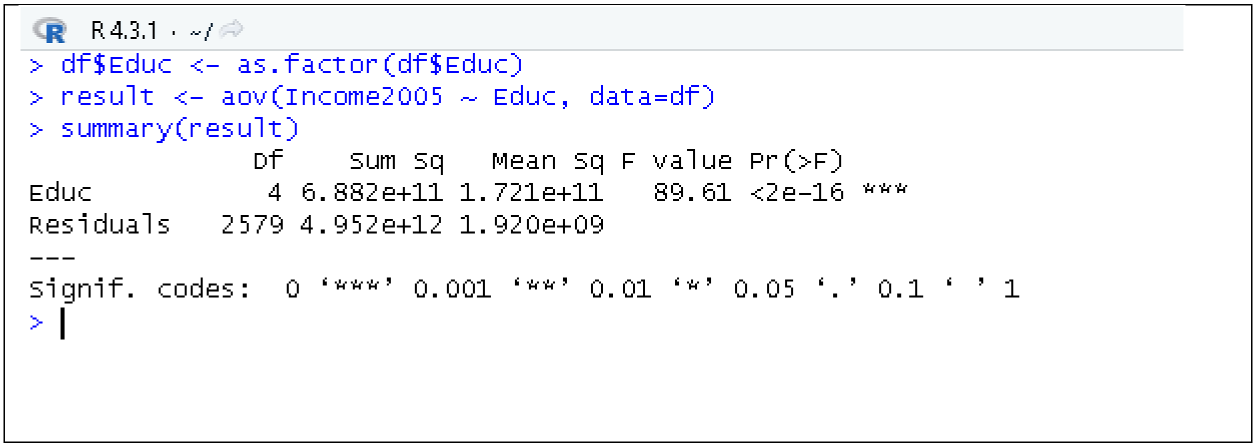
If the p-value is less than 0.05, reject the null hypothesis.

Step 6: State the Conclusion:

Based on the p-value, I can conclude that there is enough evidence to say that the mean income differs between the two groups.

3. 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.

The results suggest that the differences in mean income across the different education levels are statistically significant. Specifically, the p-value is less than 0.001 (indicated by <2e-16), which is well below the commonly used alpha level of 0.05. Therefore, I reject the null hypothesis, confirming that there is strong evidence that at least one of the education levels has a different mean income than the others. The F-value of 89.61 is also considerably high, suggesting that the effect is not just statistically significant, but also practically meaningful.



```
> df$Educ <- as.factor(df$Educ)
> result <- aov(Income2005 ~ Educ, data=df)
> summary(result)
              Df    Sum Sq   Mean Sq F value Pr(>F)    
Educ           4 6.882e+11 1.721e+11   89.61 <2e-16 ***
Residuals   2579 4.952e+12 1.920e+09                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```