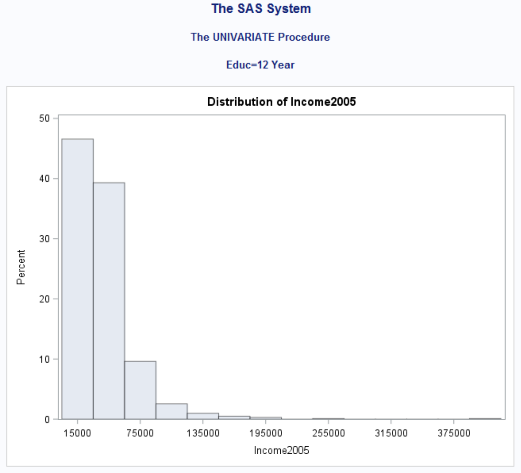
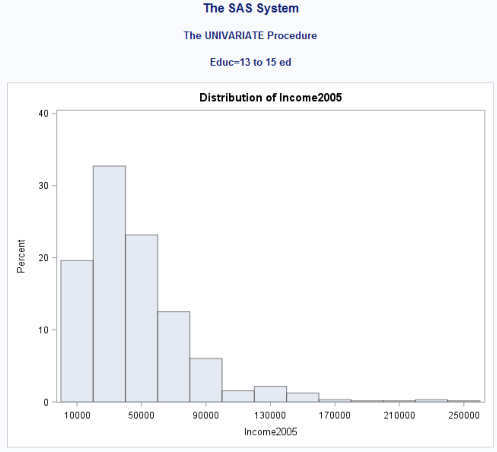
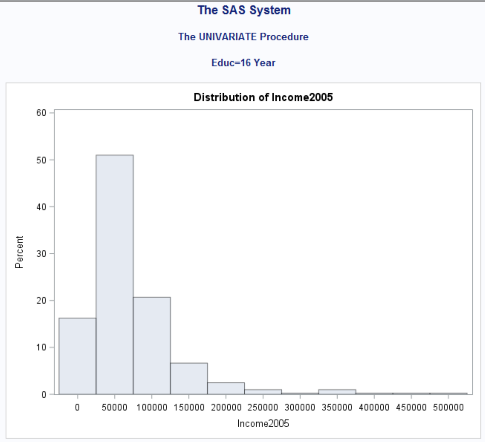
**Unit 5 HW**

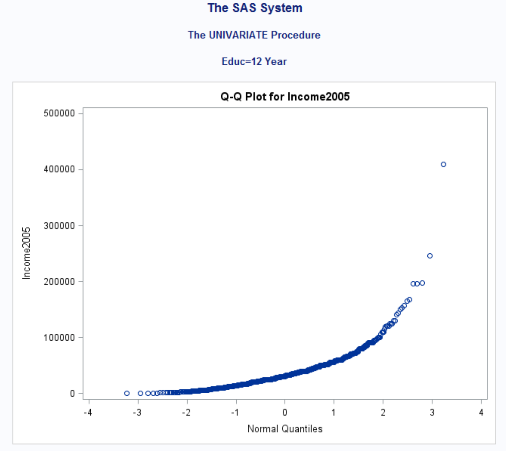
1. **Simply Answer Question 25 on pg. 147** **from the Statistical Sleuth (read it!):**

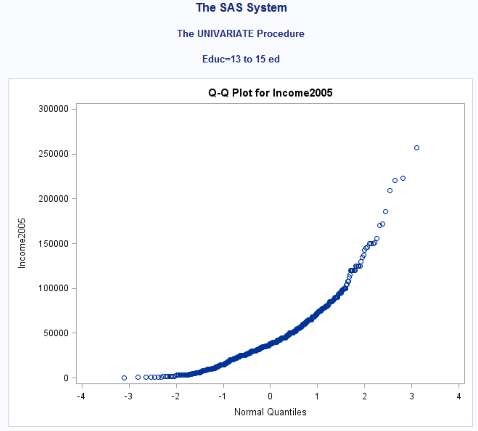
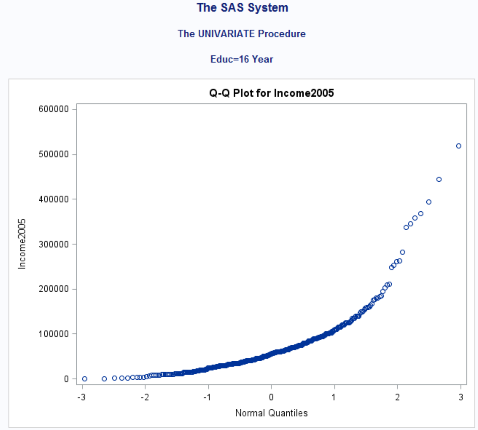
*Plot the raw data, and also plot the data after a log transform. After a log transform, do the data satisfy the assumptions better?* The data is in ex0525.csv or ex0525.xlsx. Perform this analysis in SAS. [Depending on where you find the data set, you may see the value **<<12**. Note that **<<12 = 12**.]

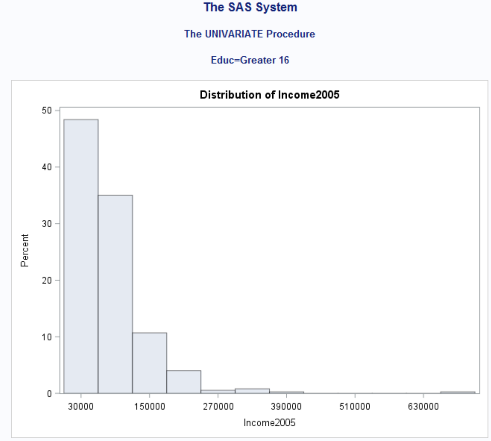
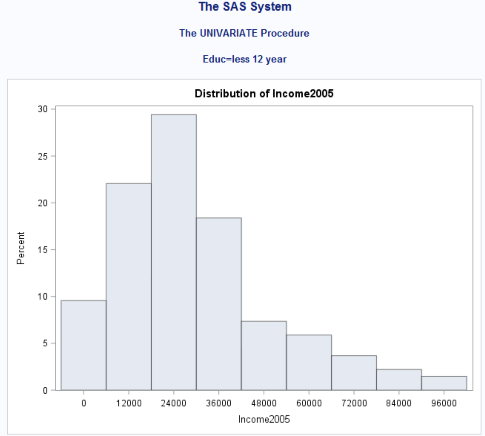
**RAW Data plot**

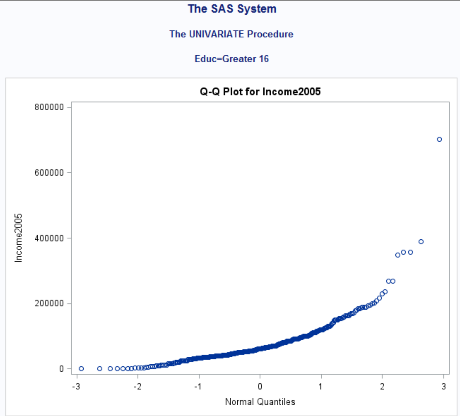
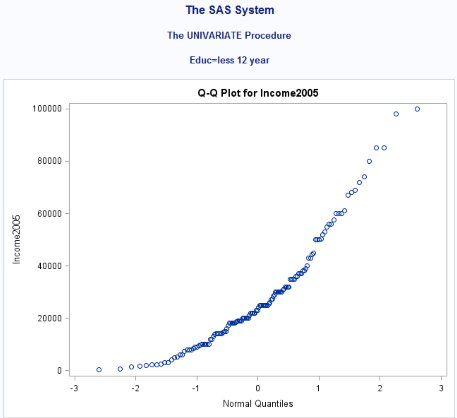




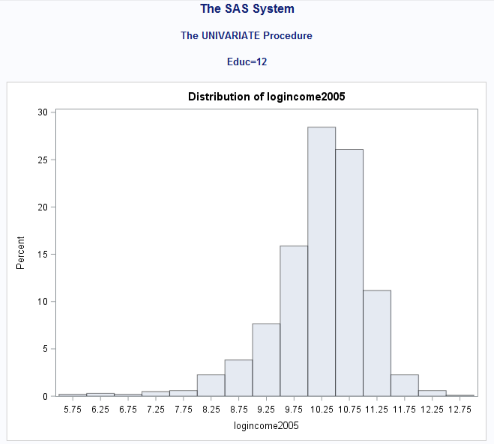
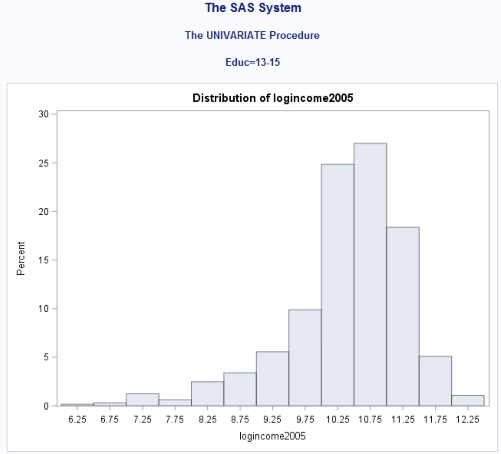
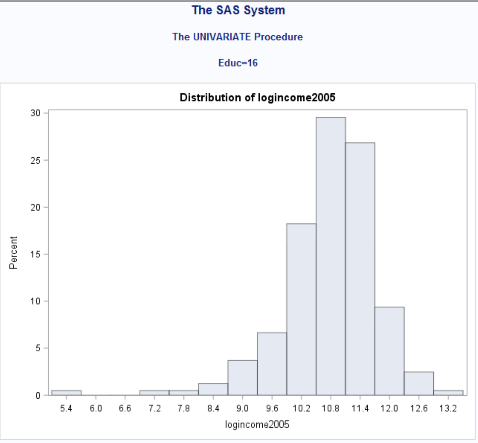


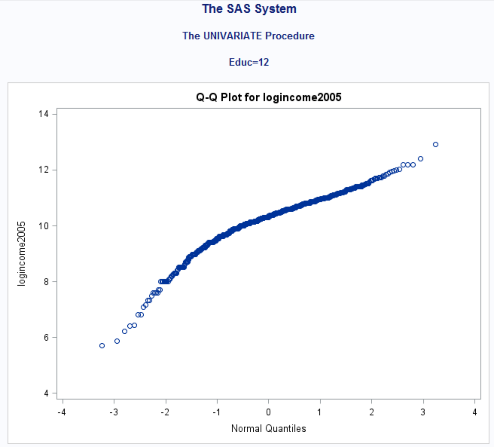


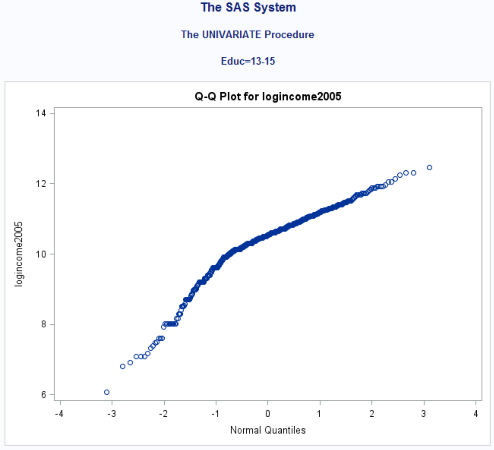
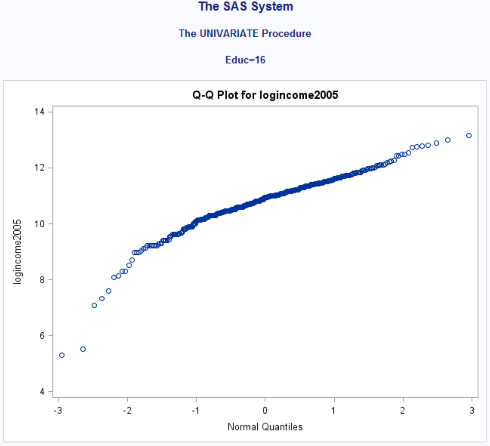


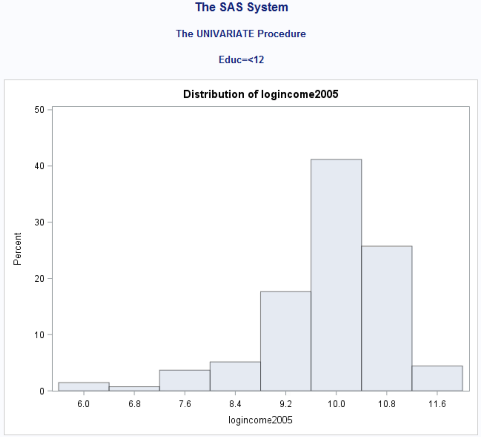
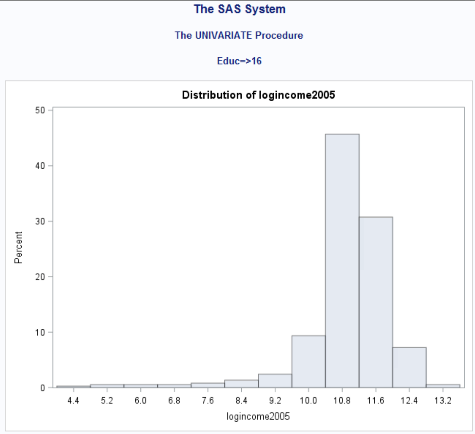


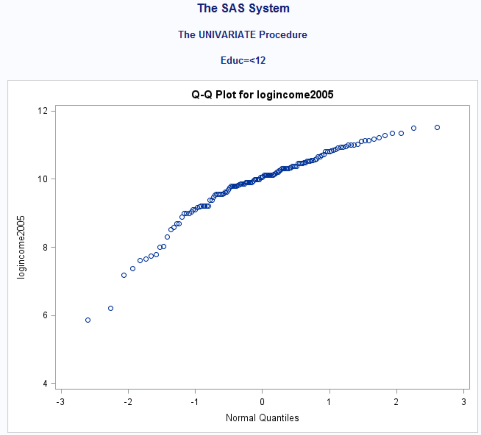
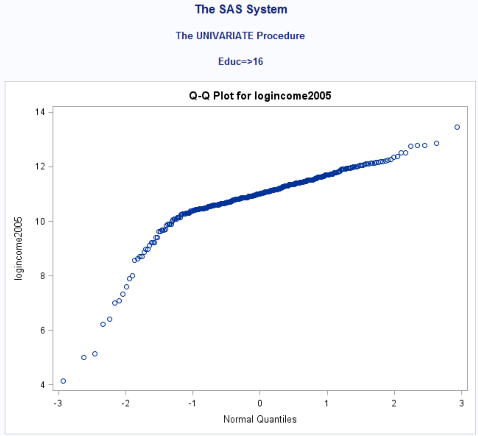
**Log Transformed Data plot**











**After a log transform, do the data satisfy the assumptions better?**

Yes, the log transformation does satisfy the assumption better than raw data.

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?)

ADDITIONAL THINGS TO INCLUDE (for the logged data):

1. Please also identify R2
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.

*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 Ho to discover WHICH pairs have evidence of different means / medians.*

**25. Education and Future Income.** The data file ex0525 contains annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005 (see Exercise 22 in Chapter 2). The data set also includes a code for the number of years of education that each individual has completed by 2006: <12, 12, 13-15, 16 and >16. 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? By how many dollars or by what percent does the mean or median for each of the last four categories exceed that of the next lowest category?

**Problem:** 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:**

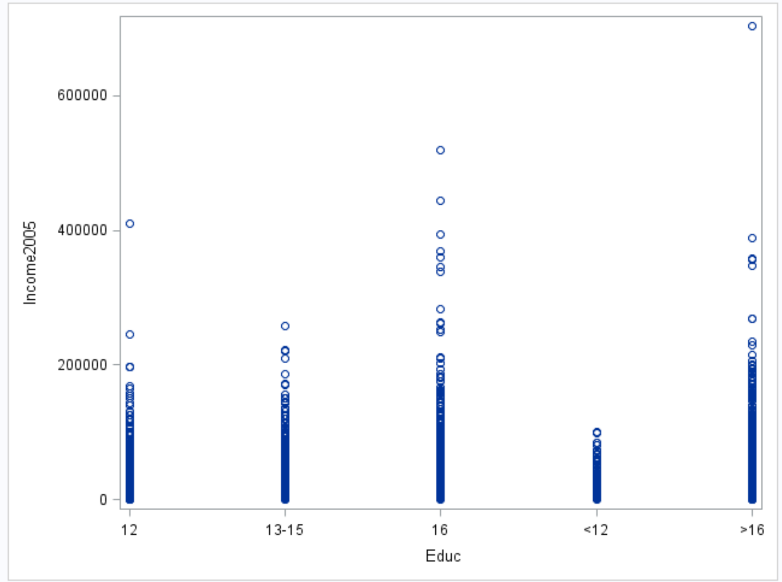
The assumptions of the ANOVA are: the incomes in each educational group come from a normal distribution, the variances of these normal distributions are equal, the data are independent with each group and the data independent between each group.

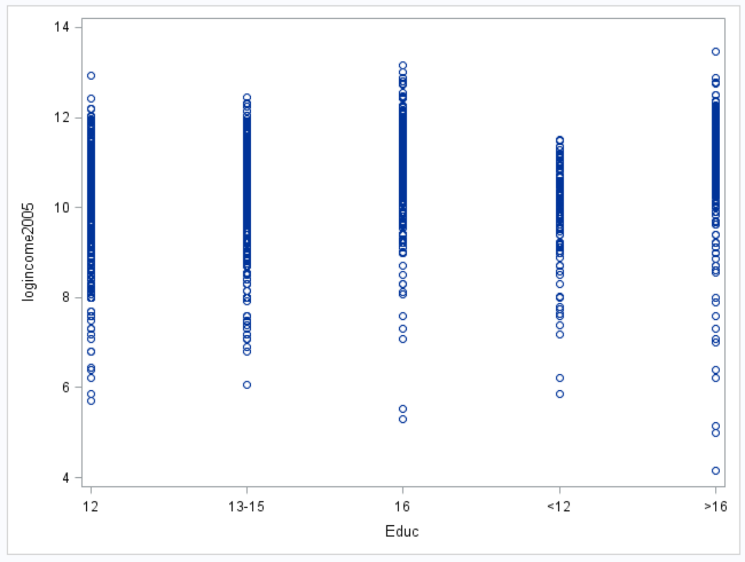
**Normality:**

The histograms and QQ plots below(log transformed data) appear to each show strong evidence of normality.

**Equal standard deviations:**

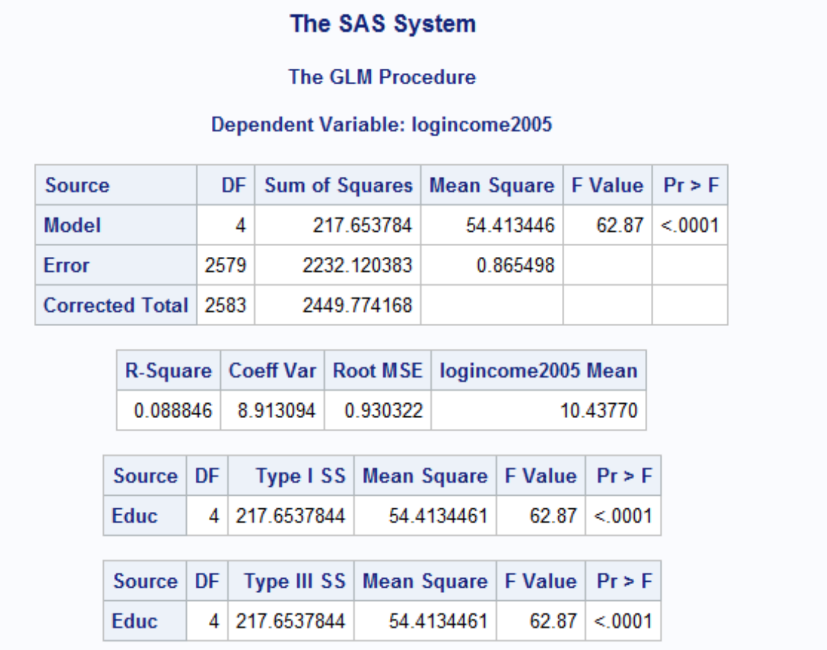
The standard deviations for log look normal rather than the raw data.

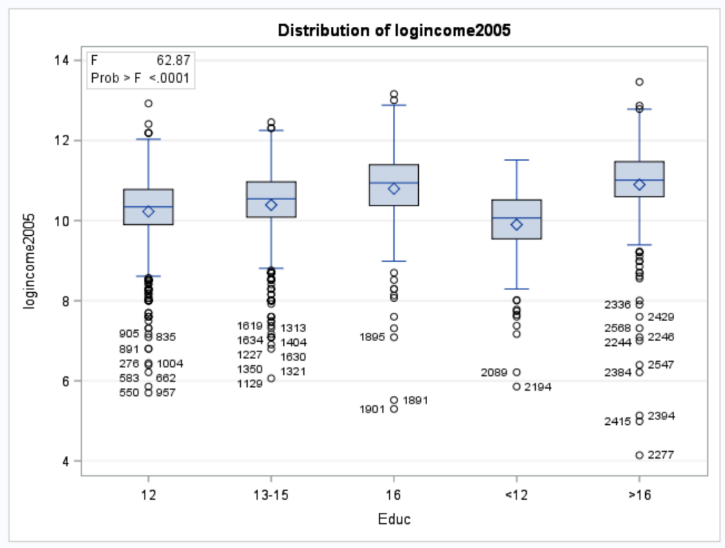
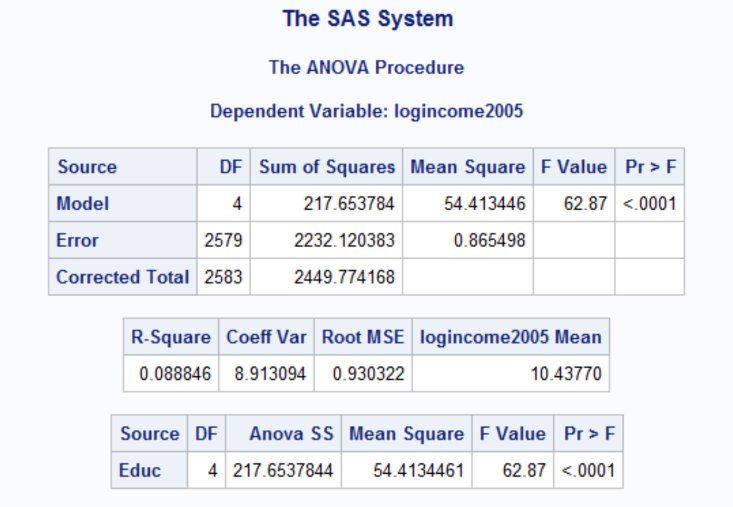




**Independence**

We will assume the data are independent, both between and within groups and proceed with ANOVA to test for differences in mean log income between the five levels of education.





Step 1: Hypothesis

H0: All median incomes are the same across education levels

Ha: At least one pair of income medians are different between education levels.

Step 2 Identification of critical value. Skip in ANOVA settings

Step 3 Value of Test Statistics: F=62.87

Step 4 – Give p value: P < 0.0001

Step 5 Decision: Reject Ho

Step 6 Conclusion: there is strong evidence to suggest that at least one of the median for a particular education level is different from the other (p <0.0001 from a pure ANOVA).

State the Scope Inference: The results of this study cannot be inferred to the generalized population beyond this study since the assignment was not randomized. This is till an observation study which cannot have cause and effect relationship attributed. But the results to convey that apply to the subjects of the study. It does have practical significance that more education can result in higher income by average. But there could be confounding variables that could explain why variances are occurring like charisma and type of jobs that subjects had.

Identify R2:

R^2 = 0.088

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?

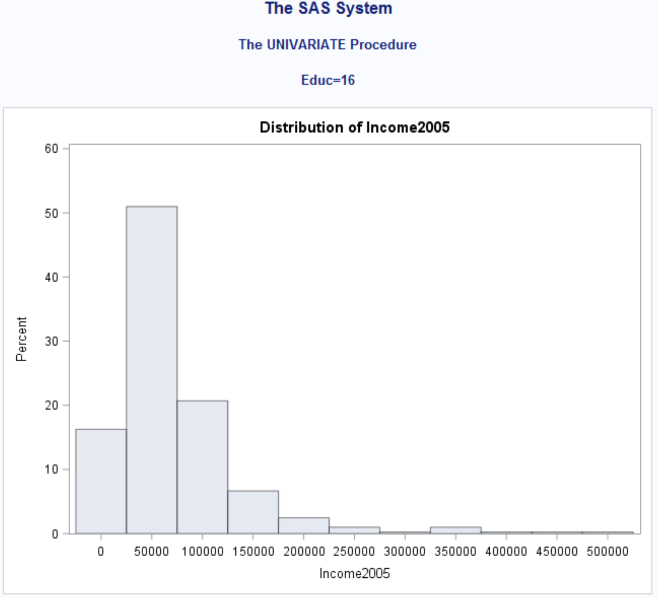
d. Manually construct the ANOVA table showing all sums of squares and the F test.

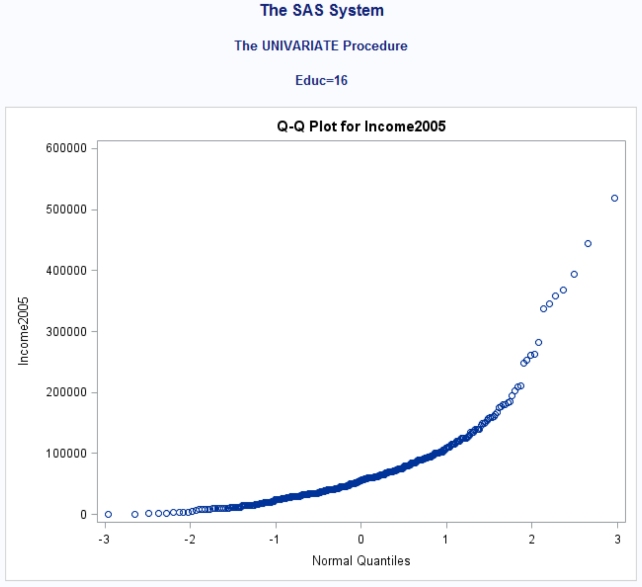
**e.** Compareyour final ANOVA table to that produced by SAS and do a 6-step complete analysis.

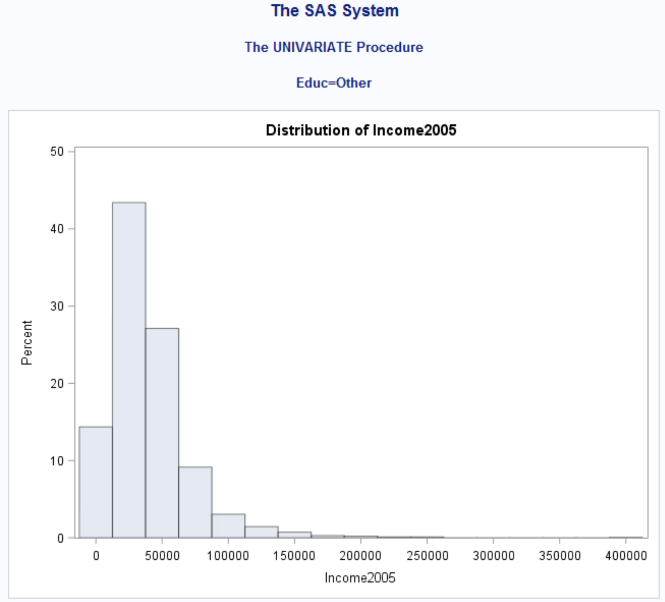
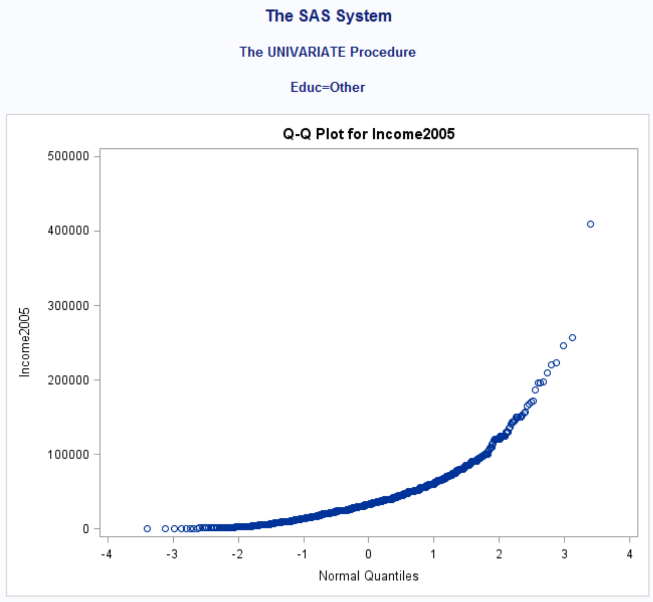
**a) Normality:**

The histograms and QQ plots below(log transformed data) appear to each show strong evidence of normality.

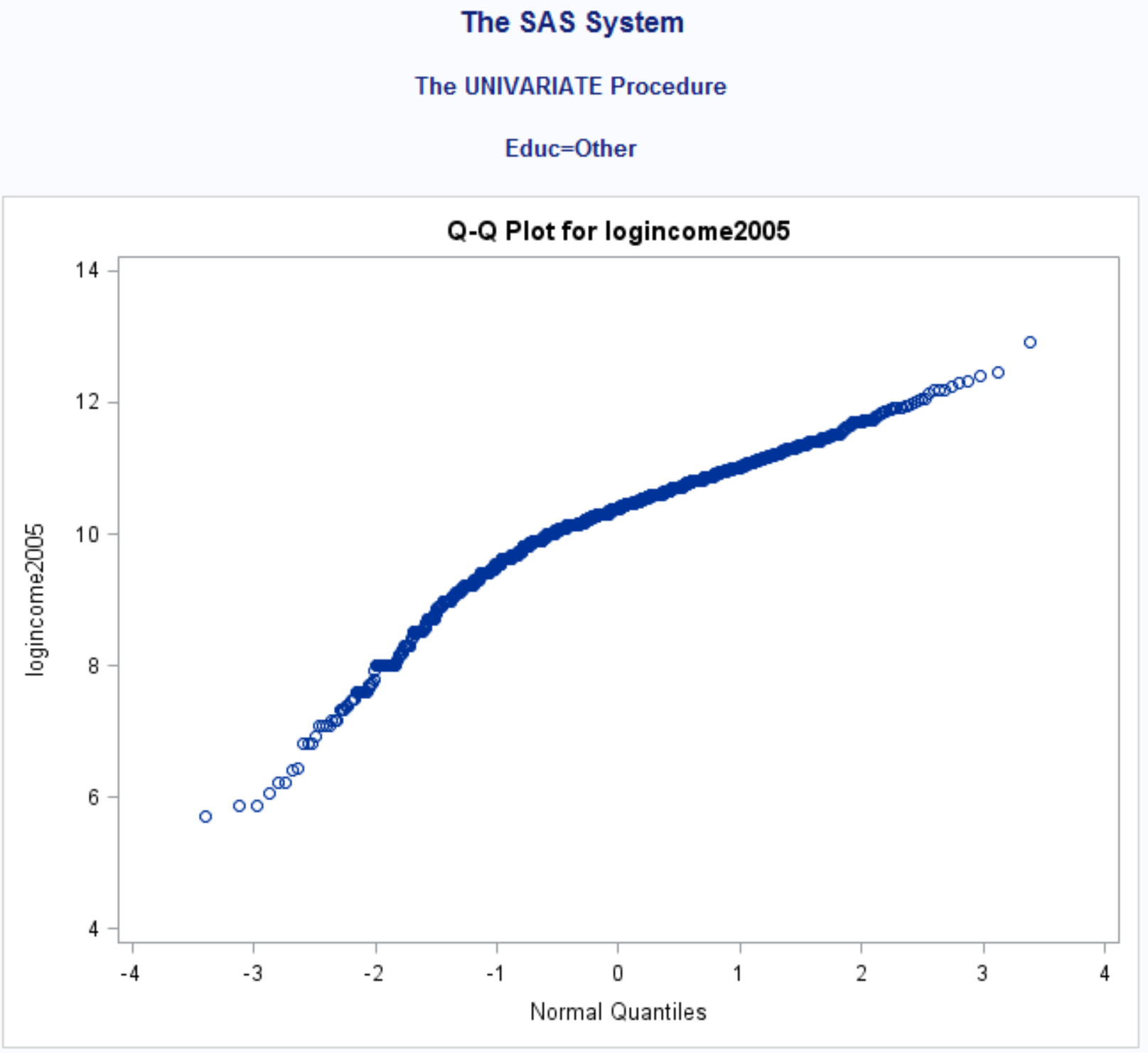
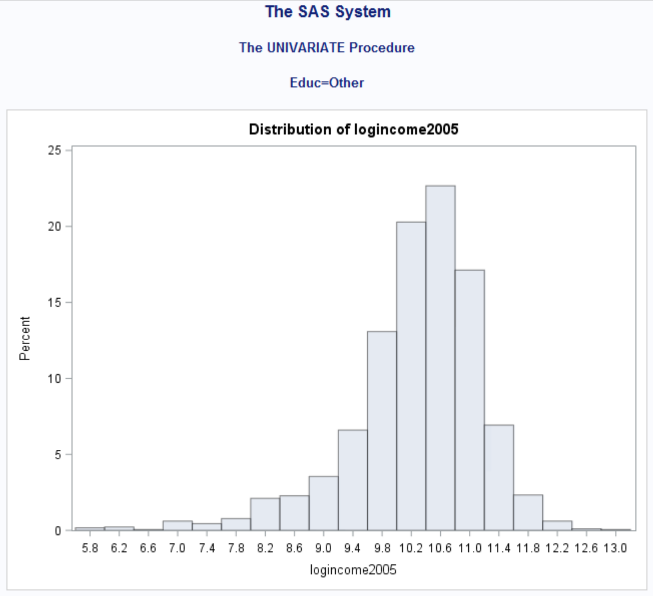
**Raw data**

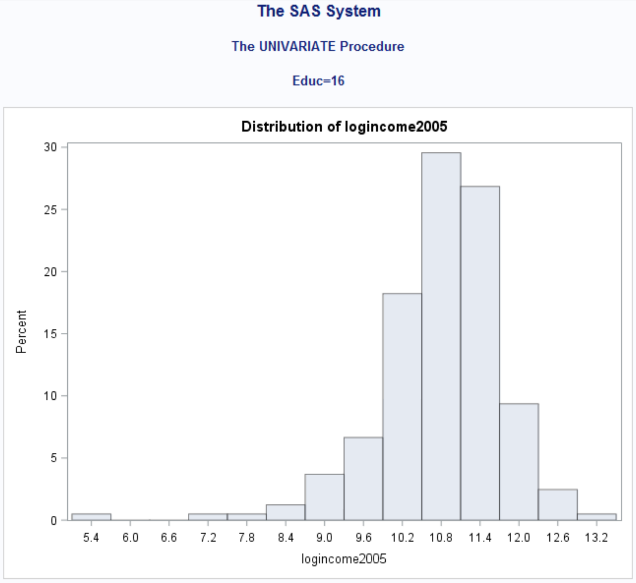


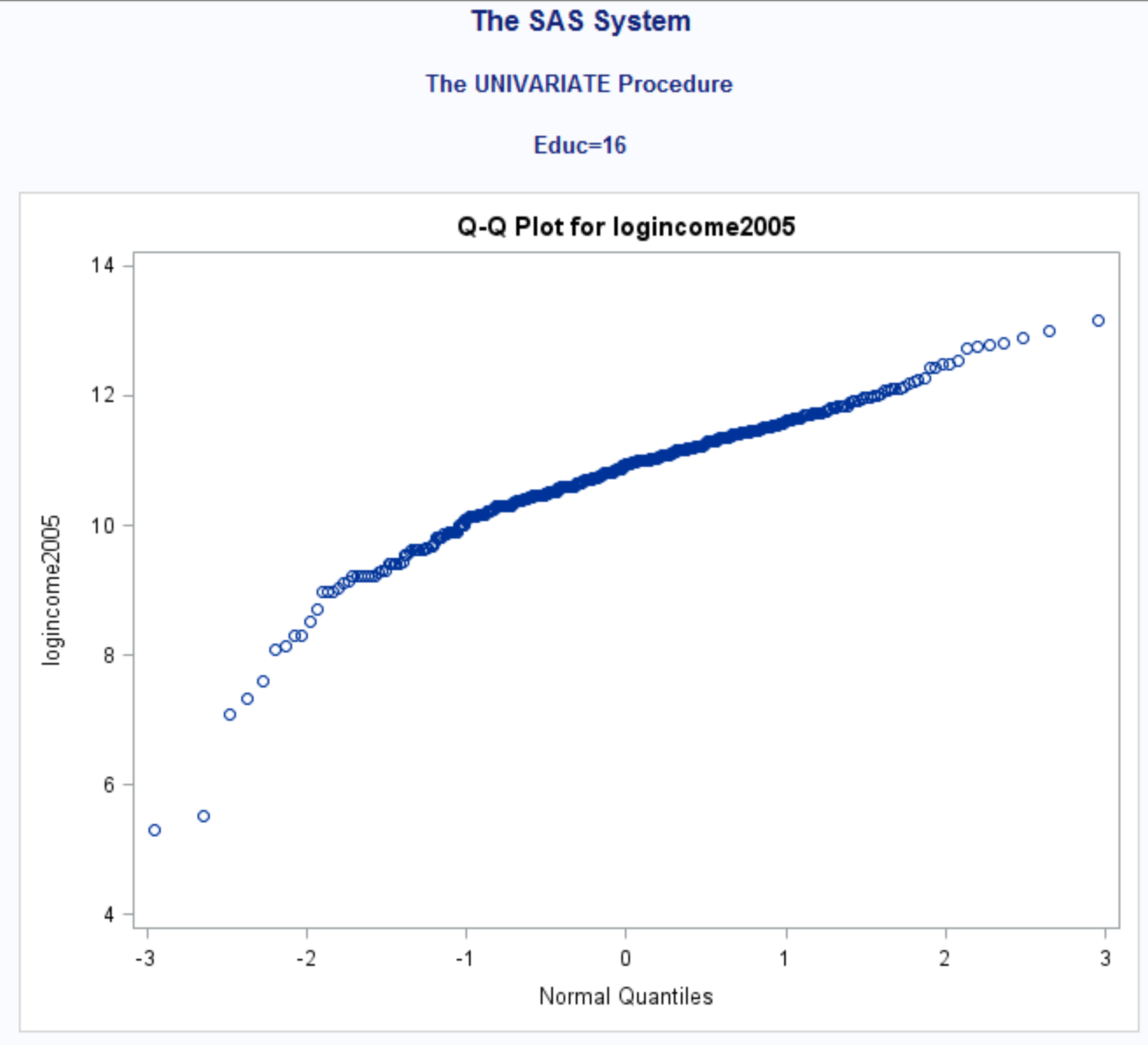




Log transformed:

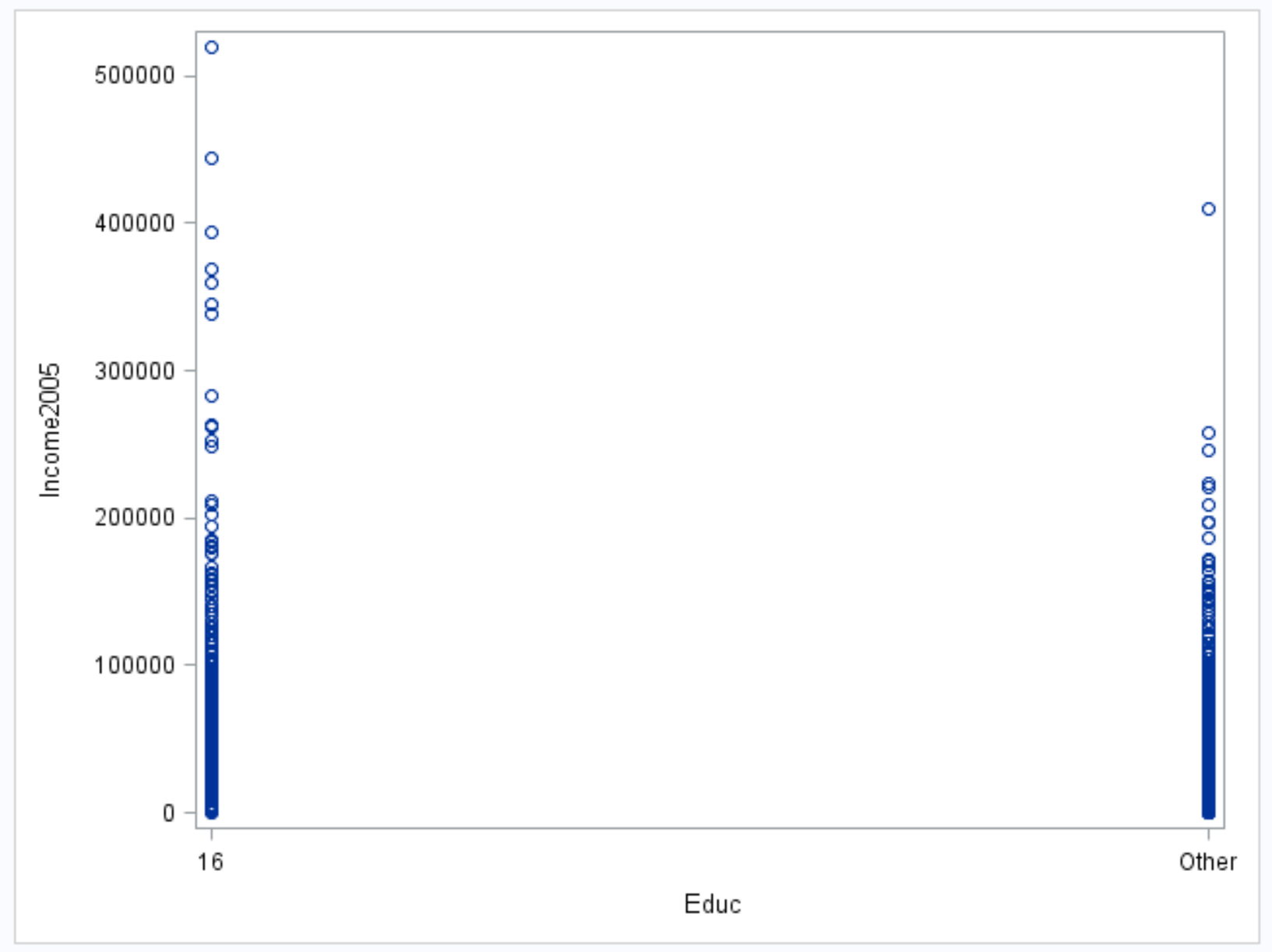
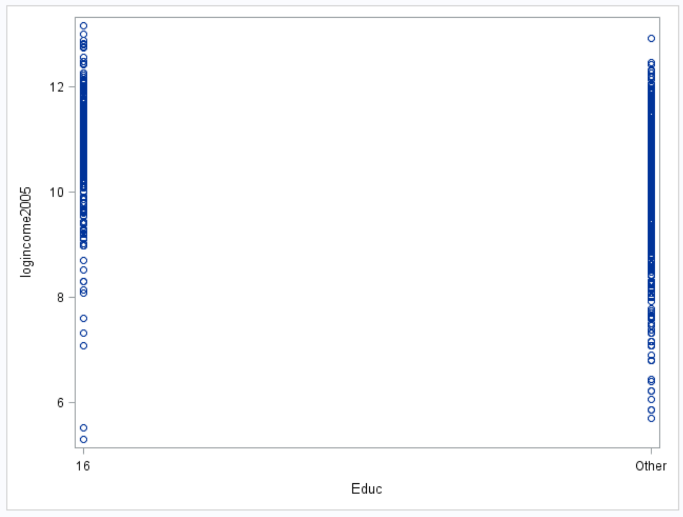






**Equal standard deviations:**

The standard deviations for log look normal rather than the raw data.



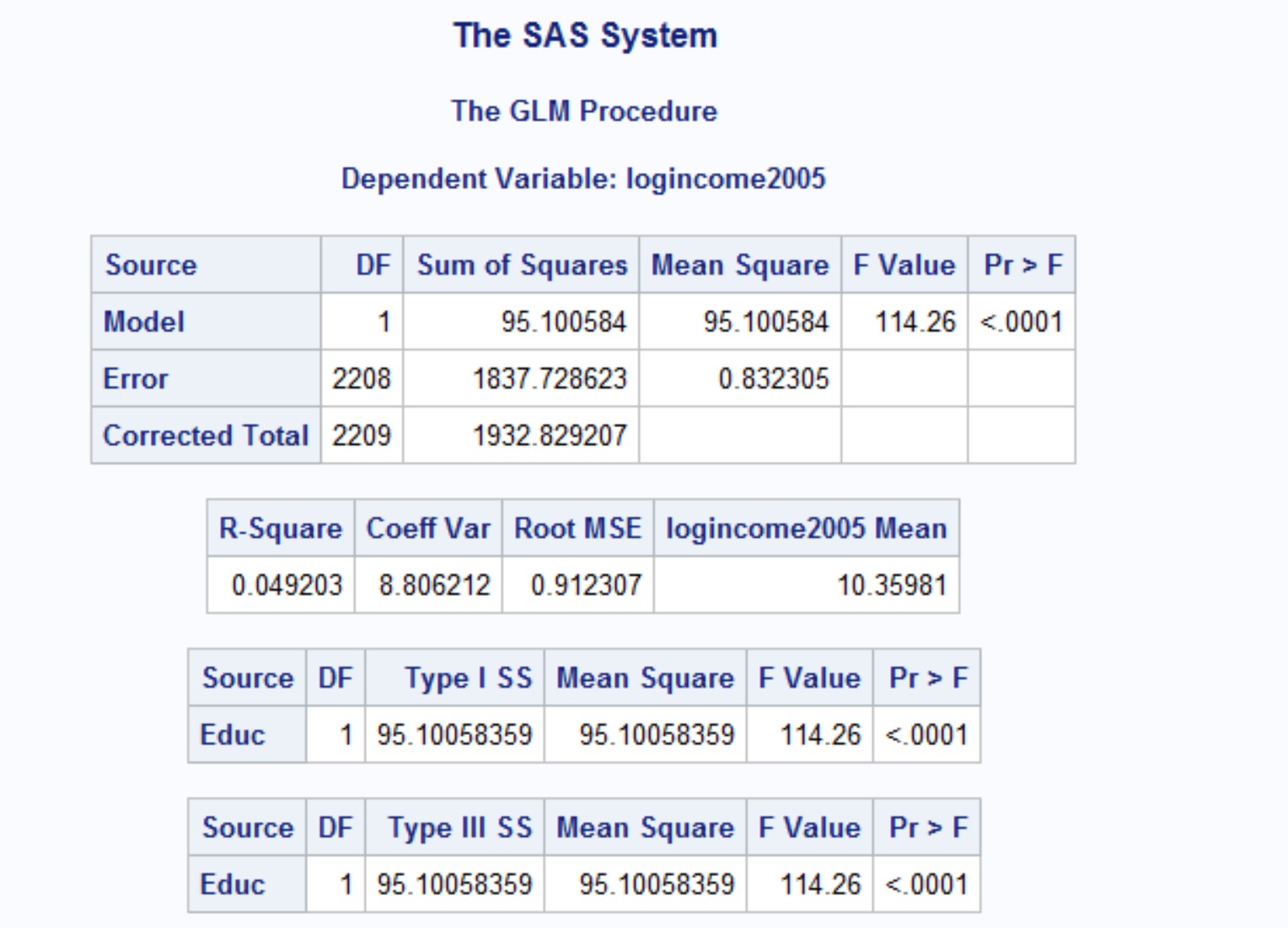
**Independence**

We will assume the data are independent, both between and within groups and proceed with ANOVA to test for differences in mean log income between the five levels of education.

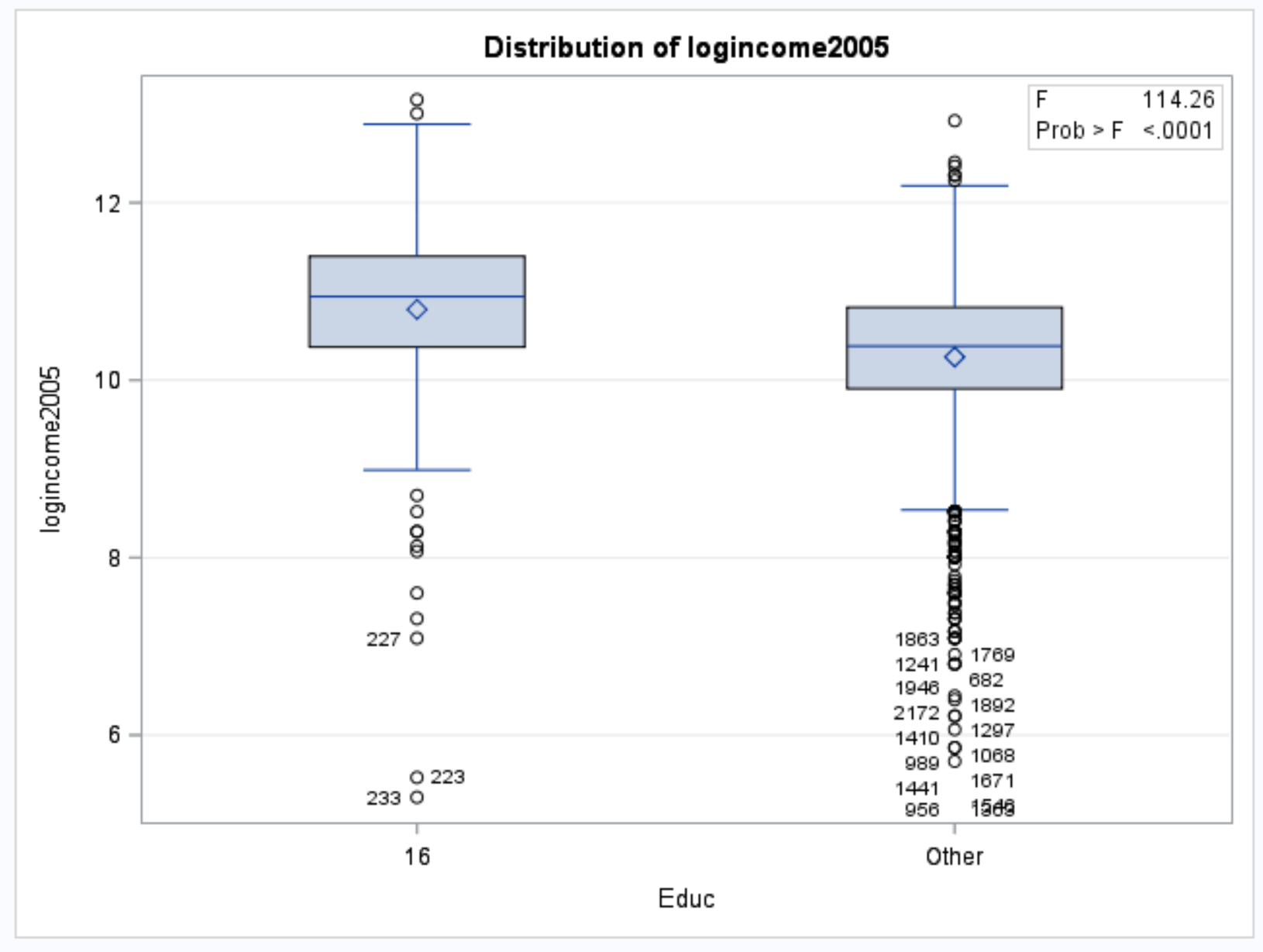
**c. Why do any deviations from Normality in this data not restrict an ANOVA with this data?**

This is to see what is causing the variance away from normality.

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



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | Mean Sq | F | Prob F |
| Model | 1 | 95.100584 | 95.1005 | 114.36 | <.0001 |
| Error | 2208 | 1837.7286 | 0.83205 |  |  |
| Total | 2209 | 1932.82 |  |  |  |



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

Step 1: Hypothesis

H0: u= ued16 = uother

Ha: At least one pair of income medians are different between education levels.

Step 2 Identification of critical value. Skip in ANOVA settings

Step 3 Value of Test Statistics: F=114.86

Step 4 – Give p value: P < 0.0001

Step 5 Decision: Reject Ho

Step 6 Conclusion: there is strong evidence to suggest that at least one of the median for a particular education level is different from the other (p <0.0001 from a pure ANOVA).

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.

