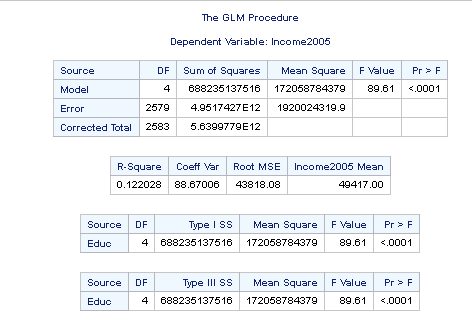
Sean Kennedy – Statistical Foundations for Data Science HW 5

## Question 1:

1. **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 had completed by 2006: 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?

Running the analysis in SAS to 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?*



Given the extremely high F value (89.51) and extremely small p-value (p<0.0001) – it is highly likely that at least one of the means in the groups of education levels is different.

*By how many dollars or by what percent does the mean or median for each of the last four categories:*

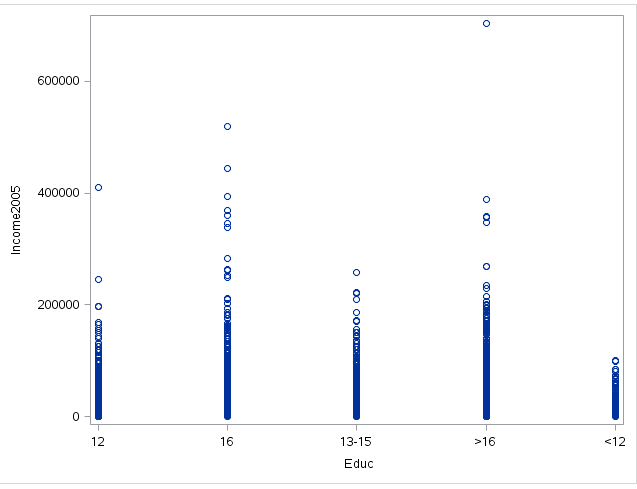
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | AverageIncome | MedianIncome | Mean Income Change | Median Income Change |
| <12 | 28,301 | 23,500 |  |  |
| 12 | 36,865 | 31,000 | 8,563 | 7,500 |
| 13-15 | 44,876 | 38,000 | 8,011 | 7,000 |
| 16 | 69,997 | 56,500 | 25,121 | 18,500 |
| >16 | 76,855 | 60,500 | 6,858 | 4,000 |

## 6 Step Analysis:

Address assumptions of ANOVA:

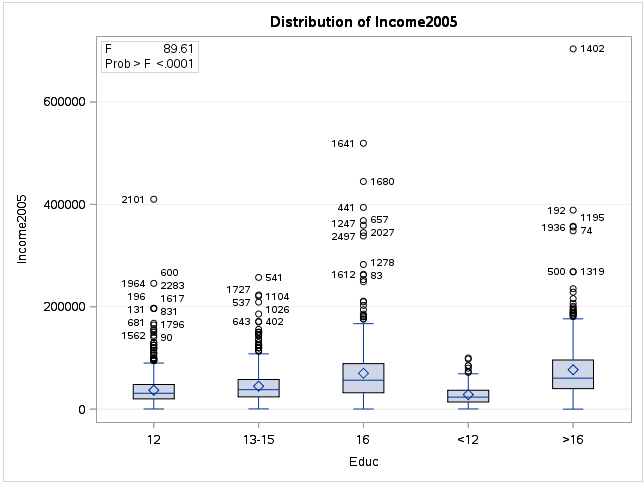
* Is data normal?

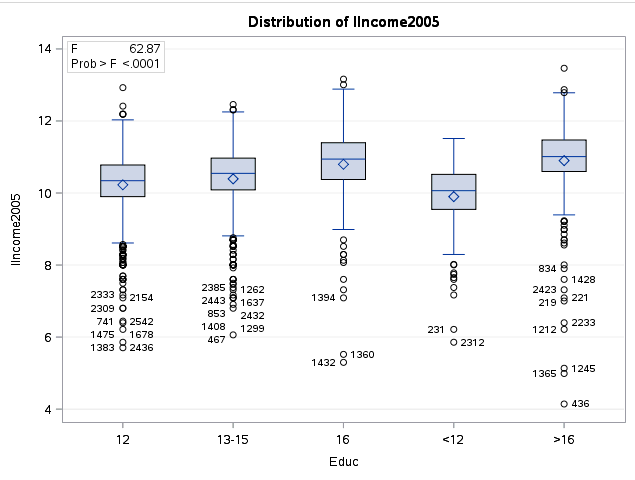
*Visual inspection of raw and log x-formed data*



Dataset appears relatively skewed and also contains several outliers – but sample size is large enough to invoke the CLT and assume normality. The ANOVA test is robust to the assumption of normality when N is sufficiently large.

* Are SDs equal?





Standard deviation of log-transformed data appears more uniform from visual inspection of box plots above.

* Independence within groups and between?

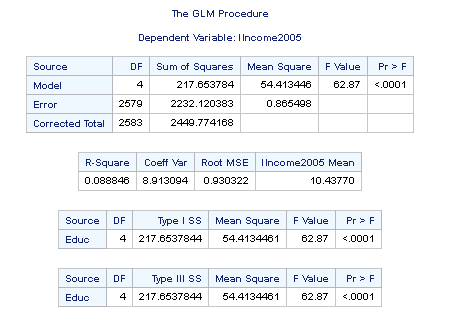
Since this is a random sample of data there is no reason to assume any dependence within members of a group or between the 5 groups themselves.

# State the Problem:

We wish to determine if a random sample of individuals from the National Longitudinal Survey of 1979 exhibit meaningful differences in their 2005 income based on the number of years of education undertaken. The age groups were defined as less than 12 years (assumed not to have graduated HS), 12 years (HS graduate), 13-15 years (some college), 16 years (college graduate) and more than 16 years (graduate school or longer). First – we will conduct an ANOVA test on the log transformed data to assess whether the mean salary of at least one of the 5 groups is significantly different than the other 4. At which point, we may conduct further analysis to determine which mean(s) of the 5 are significantly different.

Null Hypothesis: H0 (all means equal) m1=m2=m3=m4=m5

Alternate Hypothesis: Ha (at least one mean is different)



F-Statistic: 62.87

P-Value: <.0001

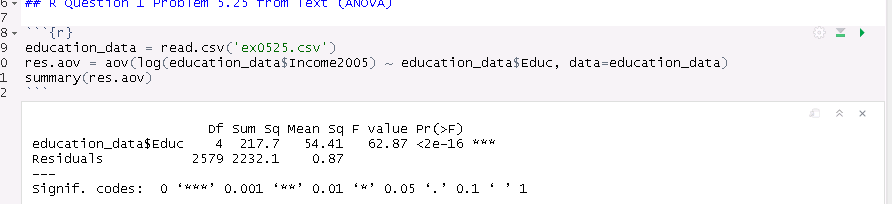
# Statistical Conclusion:

Given the extremely high F-statistic and very low p-value (p < 0.0001) we reject the null hypothesis of equal means and find strong evidence in favor of at least one of the means being different (probability of observing F-Value of 62.87 given 4 model df and 2579 error df is less than 1 in 10,000 by random chance). Note that the R-Square value for this test is abnormally low (0.088) since the extra sum of squares term (217.65: the amount of variance explained by the model) is a fraction of the overall sum of squares (2449.78: the total amount of variance in the model).

# Scope of Inference:

Since this was an observational study, no causal inference can be made regarding a single mean being different from the group – but – there is significant statistical evidence to suggest that at least one of the means is different than the others. This result can be generalized to a larger population given that the selection of participants was random. Given the low R-square value – further analysis is suggested.

# R-Code:



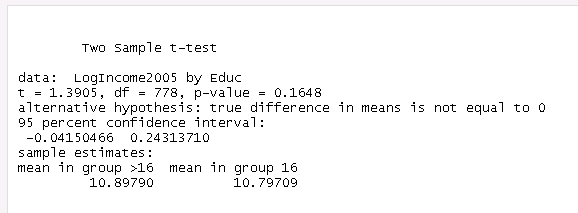
## Question 2:

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.

# State the Problem:

Having already established that there is at least one group amongst the five groups detailed in the previous analysis using a one-way ANOVA test on the log-transformed data, we will now build our own ANOVA model (extra sum of squares F-Test) to test whether or not the mean (LOG) incomes of the groups 16 and >16 are different.

Let’s run a simple 2 sample t-test in R to create a baseline test and improve upon its power by leveraging the full model in our bespoke ANOVA.



## ANOVA

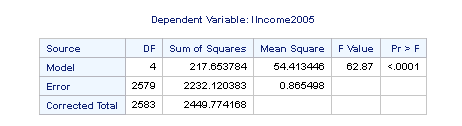
**Assuming equal SD requirement to be true**

Null Hypothesis: H0 (all means equal) m1=m2=m3=m4=m5

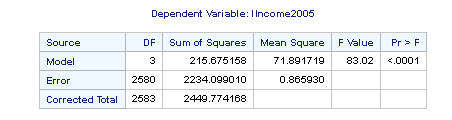
Alternate Hypothesis: Ha m4 and m5 are different (groups 16 and >16 are different)

Roll Your Own ANOVA using SAS/EXCEL:

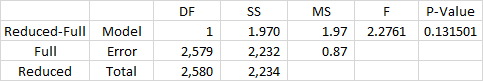
Full Model – Equal Means – H0



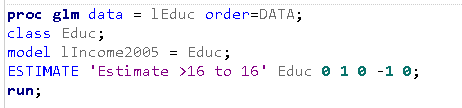
Reduced (recode groups)

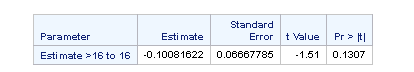


Difference between two models (Reduced – Full) gives us the analysis we are looking for in the QOI



Running the same analysis as a linear combination of groups (no recoding, just reweighting) yields identical results to serve as confirmation of the BYOA process above.





Statistical Conclusion:

Given a 95% CL (alpha = 0.05) and one sided p-value of 0.1307 (df = 2,580) there is not enough evidence to show that the mean income of group 16 was significantly different than that of the group >16.

We used an ANOVA test to increase the power or our standard pooled SD t-test to leverage the extra degrees of freedom in the full dataset.

Scope of Inference:

In regard to the first question of interest – whether or not there is at least one group amongst the 5 that has a significantly different mean income than the group average – we found overwhelming evidence that it is indeed the case that one group differs. As to the second QOI, whether or not the mean income of groups with a college degree (16 years) and those with more education than a college degree (>16 years) differs – the answer is no. By performing a series of ANOVA tests – we were able to verify the results of the two sample pooled SD t-test (df = 778) in a more powerful way (ANOVA df = 2580). Since the sample was the result of random selection, the analysis should extend to the general population. But since it was not a controlled experiment – there can be no causal inference as far as education level and mean income.

### Question 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.

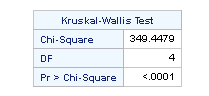
State the Problem:

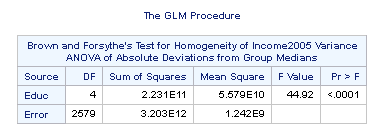
Same as above – but no longer assuming equal standard deviations forces us to adopt a non-parametric tests on the median: Brown/Forsythe ANOVA, Kruskall Wallis Ranksum tests:

H0: medians of each group are equal

Ha: at least one is different

Running both tests yields the following:





Statistical Conclusion:

Reject H0 - Both test show string evidence (p-values <0.0001) that at least one of the medians is different than the others.

Scope Of Inference:

Since the sample was the result of random selection, the analysis should extend to the general population. But since it was not a controlled experiment – there can be no causal inference as far as education level and mean income. Having performed ANOVA tests on the mean log(Income) and come to the statistical conclusion that there is a group with different means, we were able to confirm this result by using a non-parametric test on the medians of the non-log-transformed data.