Javier Saldana

86

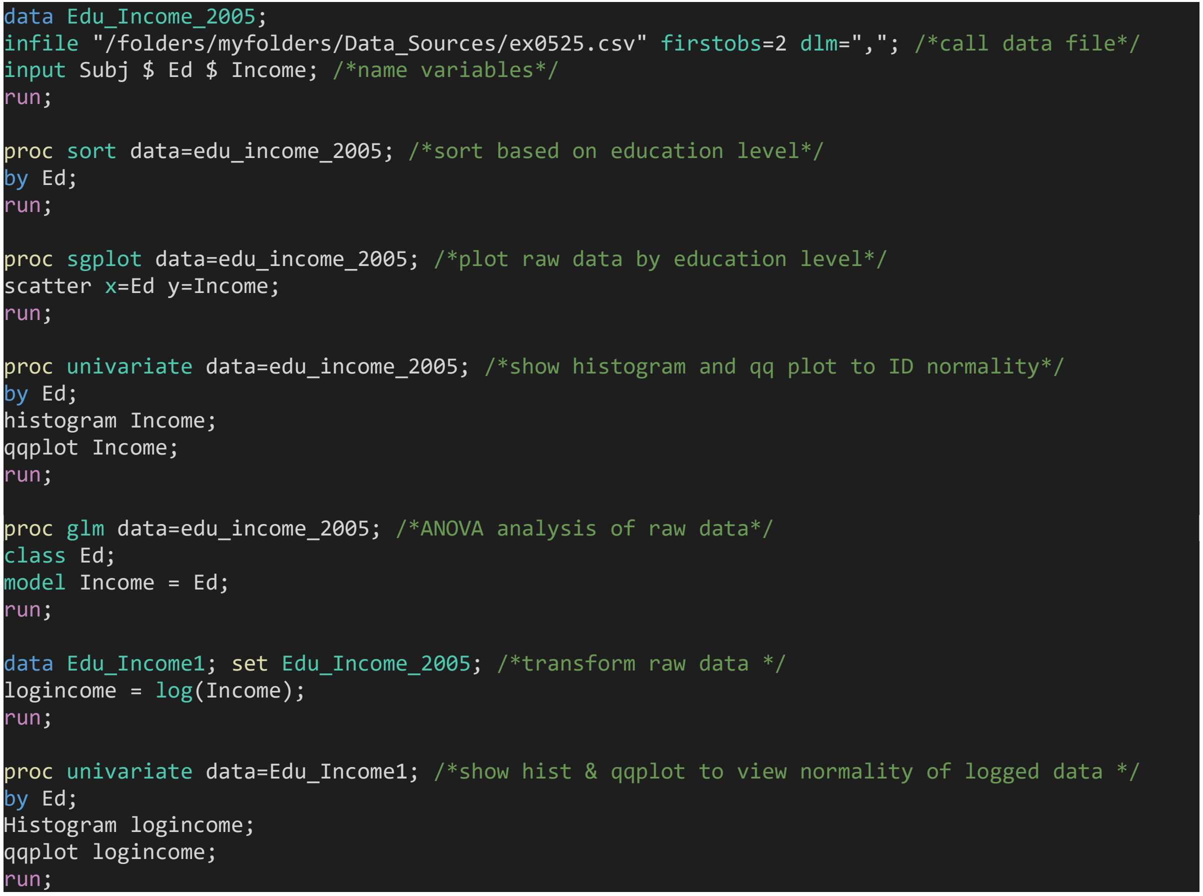
Unit 5 HW

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, if you may see the value **<<12**. Note that **<<12 = 12**.]

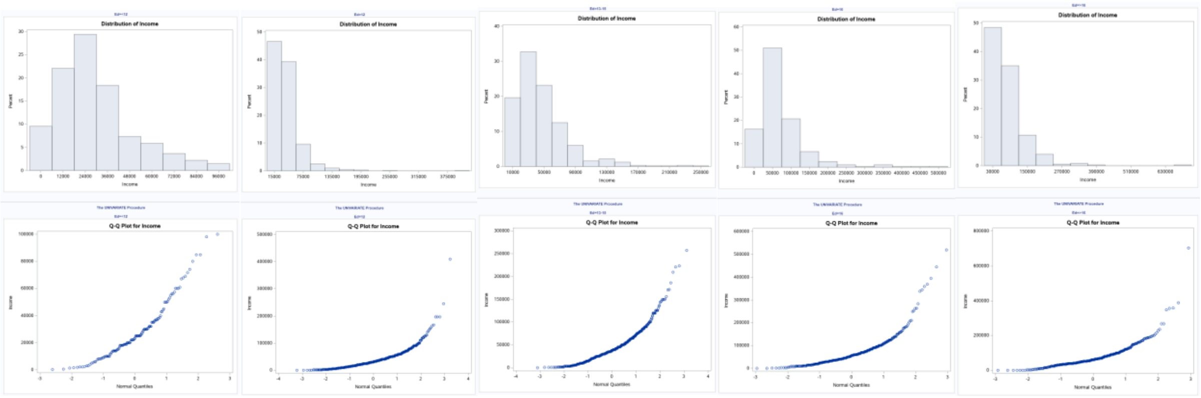
34/35

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 had 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?

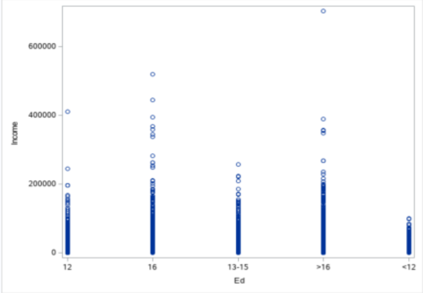
Based on the analysis conducted below, there is enough evidence to support that at least one education group's mean income is different than the other(s) ($F_{statistic}$ = 89.61; p-value = <0.0001). Once the data is transformed, there is insufficient evidence to suggest it is not normally distributed. As a result, it meets the assumptions much better than the raw data (assuming samples are all independent within each other and amongst one another).

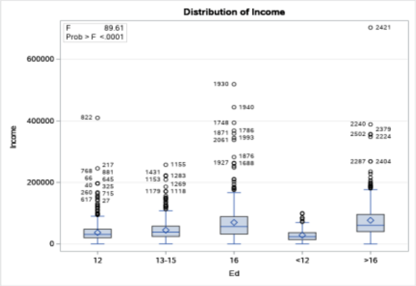


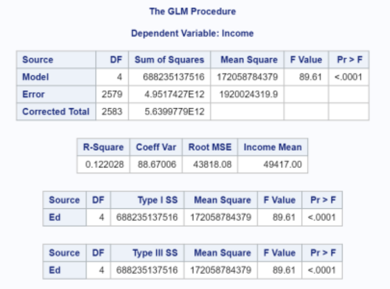
```
data Edu_Income_2005;
infile "/folders/myfolders/Data_Sources/ex0525.csv" firstobs=2 dlm=","; /*call data file*/
input Subj $ Ed $ Income; /*name variables*/
run;

proc sort data=edu_income_2005; /*sort based on education level*/
by Ed;
run;

proc sgplot data=edu_income_2005; /*plot raw data by education level*/
scatter x=Ed y=Income;
run;

proc univariate data=edu_income_2005; /*show histogram and qq plot to ID normality*/
by Ed;
histogram Income;
qqplot Income;
run;

proc glm data=edu_income_2005; /*ANOVA analysis of raw data*/
class Ed;
model Income = Ed;
run;

data Edu_Income1; set Edu_Income_2005; /*transform raw data */
logincome = log(Income);
run;

proc univariate data=Edu_Income1; /*show hist & qqplot to view normality of logged data */
by Ed;
Histogram logincome;
qqplot logincome;
run;
```



Javier Saldana







The GLM Procedure

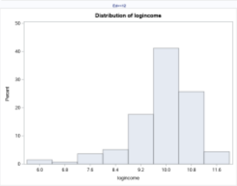
Dependent Variable: Income

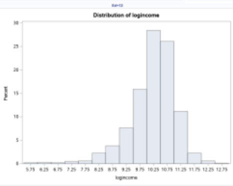
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	688235137516	172058784379	89.61	<.0001
Error	2579	4.9517427E12	1920024319.9		
Corrected Total	2583	5.6399779E12			

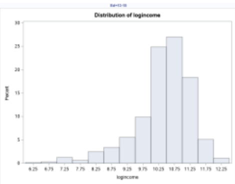
R-Square	Coeff Var	Root MSE	Income Mean
0.122028	88.67006	43818.08	49417.00

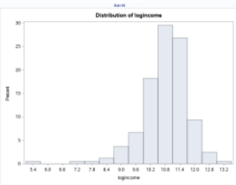
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Ed	4	688235137516	172058784379	89.61	<.0001

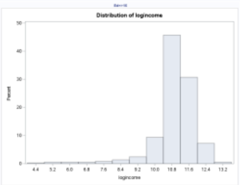
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Ed	4	688235137516	172058784379	89.61	<.0001

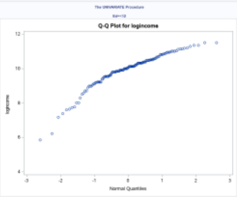


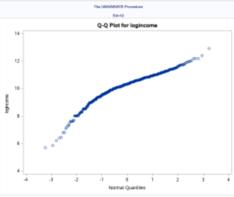


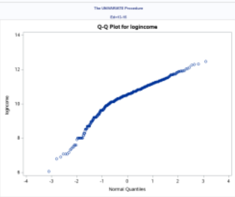


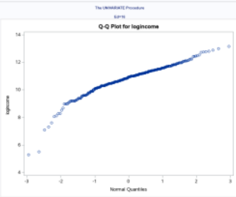


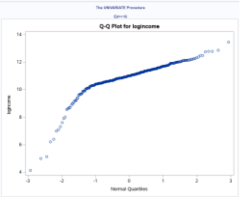


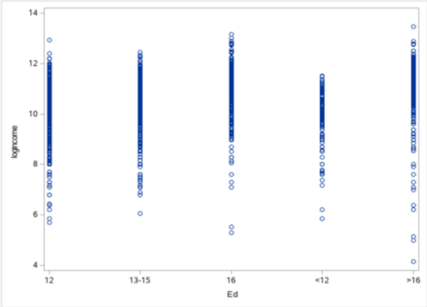


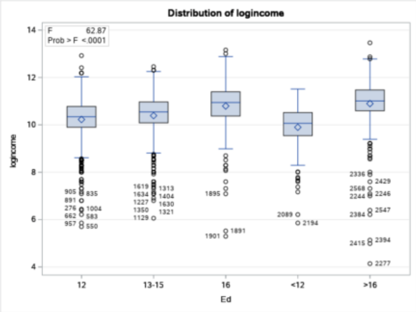


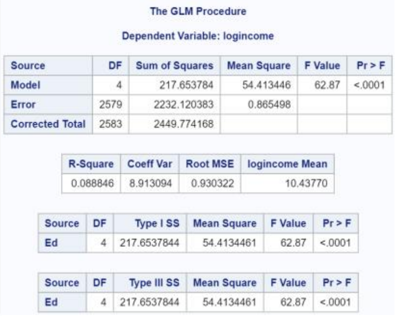












The GLM Procedure

Dependent Variable: logincome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

R-Square	Coeff Var	Root MSE	logincome Mean
0.088846	8.913094	0.930322	10.43770

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Ed	4	217.6537844	54.4134461	62.87	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Ed	4	217.6537844	54.4134461	62.87	<.0001

$H_0: \mu_{<12} = \mu_{12} = \mu_{13-15} = \mu_{16} = \mu_{>16}$
$H_A: At\ least\ one\ mean\ \neq$
F-statistic = 89.61
p-value = <0.0001
Reject H_0
The evidence suggests that at least one mean is different than another.

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.

Is there a difference in the income means of the education levels (<12, 12, 13 – 15, 16, >16)? This analysis will test if the income means of the education levels are different.

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.

The plot of the raw data indicates the data has similar standard deviations. Yet, the histograms show strong evidence that the data is not normally distributed. In order to correct this issue, we transform the data (log). The review of the transformed data shows the normally distribution issue has been corrected and the standard deviation is also similar. Considering all we know about the sampling procedure is that it was gathered randomly, we'll assume the observations are independent within and amongst the groups. The data meets the assumptions of the ANOVA test, and we will proceed with it.

3. Conduct the Test. (An example is in the UNIT 5 PowerPoint.)

$H_0: \mu_{<12} = \mu_{12} = \mu_{13-15} = \mu_{16} = \mu_{>16}$ $H_A: At\ least\ one\ mean\ \neq$
F-statistic = 62.87 R^2 = 0.088846 p-value = <0.0001 MSE: 54.413446 (df = 4)

Javier Saldana

Reject H_0

4. Write a conclusion. (An example is in the UNIT 5 PowerPoint.)

The evidence overwhelmingly suggests that at mean incomes of the different education levels are different (p-value <0.0001 from an ANOVA F-test).

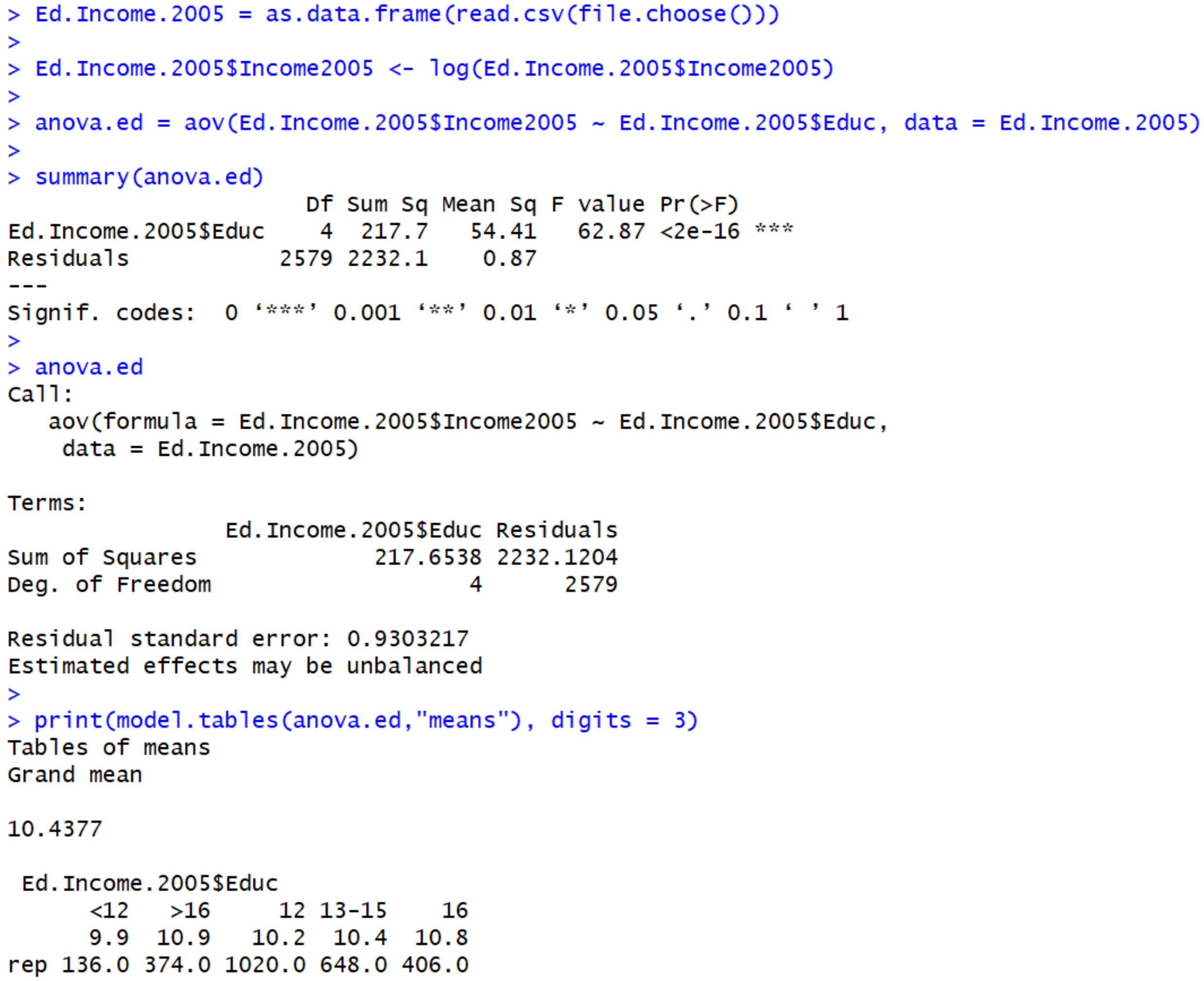
5. State the Scope. (Can we generalize to the entire population or just the sample that was taken? Is there a causal relationship present?)

Since this was an observational study, we are unable to establish a causal relationship between education and income. Since the group was selected from the longitudinal study in 1976 and we don't know how that group was obtained, we are only able to infer the results of this study to these participants.

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.*

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.

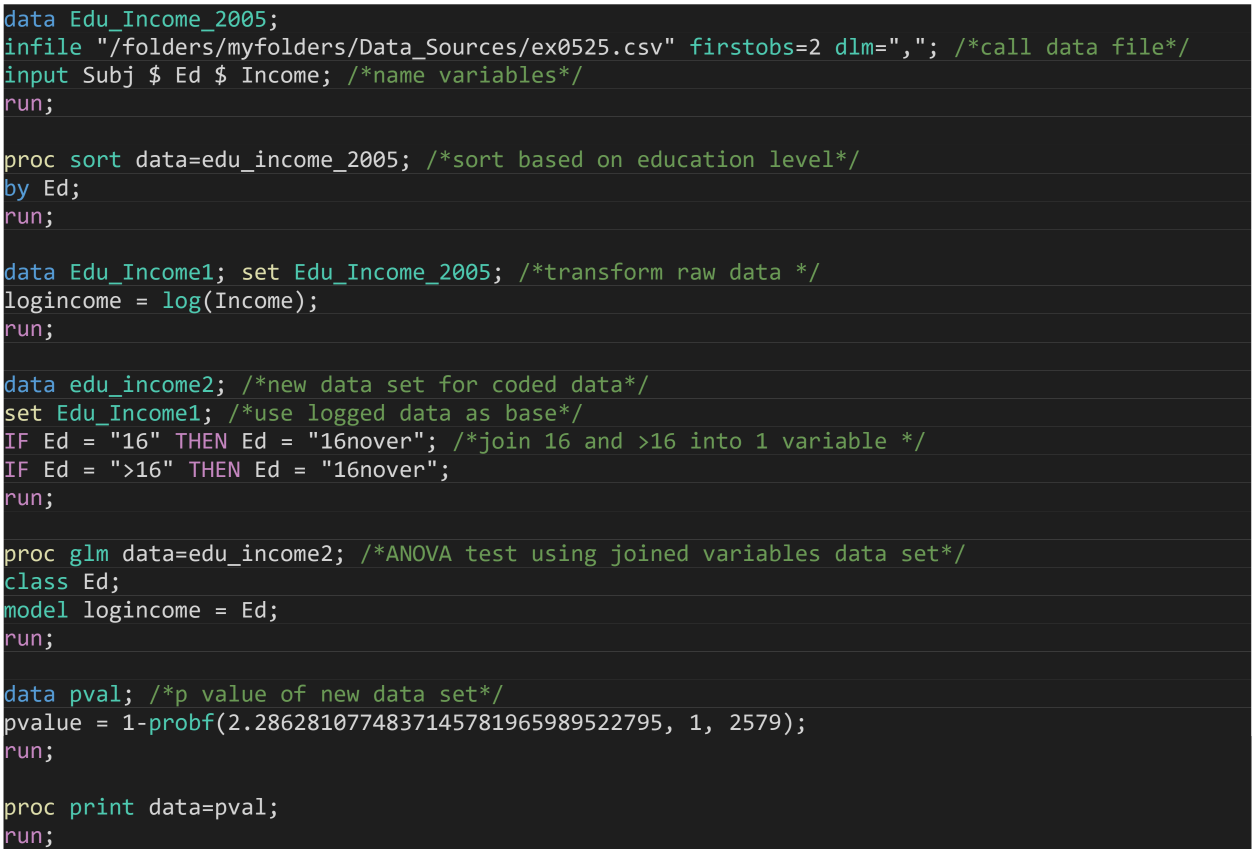


```
> Ed.Income.2005 = as.data.frame(read.csv(file.choose()))
>
> Ed.Income.2005$Income2005 <- log(Ed.Income.2005$Income2005)
>
> anova.ed = aov(Ed.Income.2005$Income2005 ~ Ed.Income.2005$Educ, data = Ed.Income.2005)
>
> summary(anova.ed)
                      Df Sum Sq Mean Sq F value Pr(>F)    
Ed.Income.2005$Educ    4  217.7   54.41   62.87 <2e-16 ***
Residuals           2579 2232.1    0.87                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> anova.ed
Call:
   aov(formula = Ed.Income.2005$Income2005 ~ Ed.Income.2005$Educ, 
    data = Ed.Income.2005)

Terms:
                Ed.Income.2005$Educ Residuals
Sum of Squares             217.6538 2232.1204
Deg. of Freedom                   4      2579

Residual standard error: 0.9303217
Estimated effects may be unbalanced
>
> print(model.tables(anova.ed,"means"), digits = 3)
Tables of means
Grand mean
         
10.4377 

 Ed.Income.2005$Educ 
      <12   >16     12 13-15    16
      9.9  10.9   10.2  10.4  10.8
rep 136.0 374.0 1020.0 648.0 406.0
```

20/30

Javier Saldana

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.



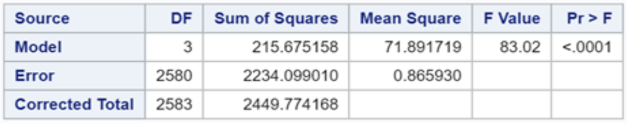
```
data Edu_Income_2005;
infile "/folders/myfolders/Data_Sources/ex0525.csv" firstobs=2 dlm=","; /*call data file*/
input Subj $ Ed $ Income; /*name variables*/
run;

proc sort data=edu_income_2005; /*sort based on education level*/
by Ed;
run;

data Edu_Income1; set Edu_Income_2005; /*transform raw data */
logincome = log(Income);
run;

data edu_income2; /*new data set for coded data*/
set Edu_Income1; /*use logged data as base*/
IF Ed = "16" THEN Ed = "16nover"; /*join 16 and >16 into 1 variable */
IF Ed = ">16" THEN Ed = "16nover";
run;

proc glm data=edu_income2; /*ANOVA test using joined variables data set*/
class Ed;
model logincome = Ed;
run;

data pval; /*p value of new data set*/
pvalue = 1-probf(2.2862810774837145781965989522795, 1, 2579);
run;

proc print data=pval;
run;
```

Is the mean income of bachelor's degree (16) and more than bachelor's degree (>16) different?

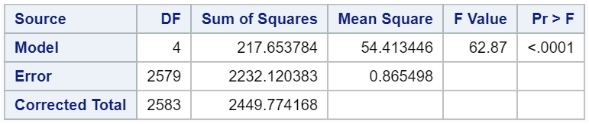
$$(H_0)\ Reduced\ Model\colon \mu_{>12}\ \mu_{12}\ \mu_{13-15}\ \mu_{16/>16}\ \mu_{16\ /\ >16}$$

$$(H_a)\ Full\ Model\colon \mu_{>12}\ \mu_{12}\ \mu_{13-15}\ \mu_{16}\ \mu_{>16}$$

$(H_0)\ Reduced\ Model\colon \mu\ \mu\ \mu\ \mu\ \mu$	$(H_0)\ Reduced\ Model\colon \mu\ \mu\ \mu\ \mu\ \mu$
$(H_a)\ Full\ Model\colon \mu_{>12}\ \mu_{12}\ \mu_{13-15}\ \mu_{16/>16}\ \mu_{16\ /\ >16}$	$(H_a)\ Full\ Model\colon \mu_{>12}\ \mu_{12}\ \mu_{13-15}\ \mu_{16}\ \mu_{>16}$



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	215.675158	71.891719	83.02	<.0001
Error	2580	2234.099010	0.865930		
Corrected Total	2583	2449.774168			



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.9787717	1.9787717	2.28628108	0.13064
Error	2579	2232.1202383	0.865498		
Corrected Total	2580	2234.099010			

F statistic = 2.2863
p-value = 0.13064
Mean Square Error = 1.9787717 (df = 1)

Javier Saldana

Fail to reject H_0

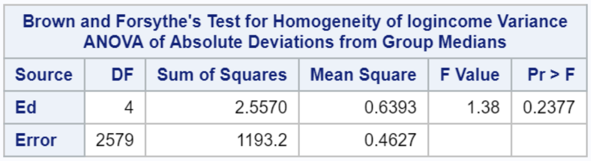
Evidence suggest that the mean incomes of 16-year of education is not different than the mean incomes of over 16-year education levels (p-value 0.13064 from an ANOVA)

20/20

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.

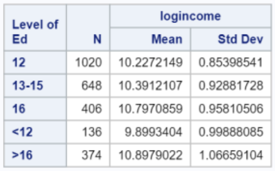
Problem - Is there a difference in the income means of the education levels (<12, 12, 13 – 15, 16, >16)? This analysis will test if the income means of the education levels are different.

Assumptions - The plot of the raw data indicates the data has similar standard deviations. Yet, the histograms show strong evidence that the data is not normally distributed. In order to correct this issue, we transform the data (log). The review of the transformed data shows the normally distribution issue has been corrected and the standard deviation is also similar. Considering all we know about the sampling procedure is that it was gathered randomly, we'll assume the observations are independent within and amongst the groups. For the sake of the homework, we will assume the groups have different standard deviations even after log transformation. With the violation of standard deviations, a Kruskal-Wallis Test is appropriate.

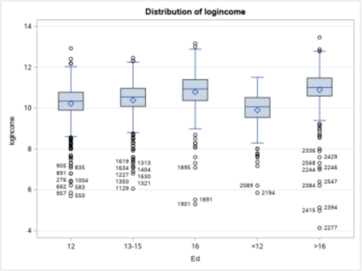


Brown and Forsythe's Test for Homogeneity of logincome Variance
ANOVA of Absolute Deviations from Group Medians

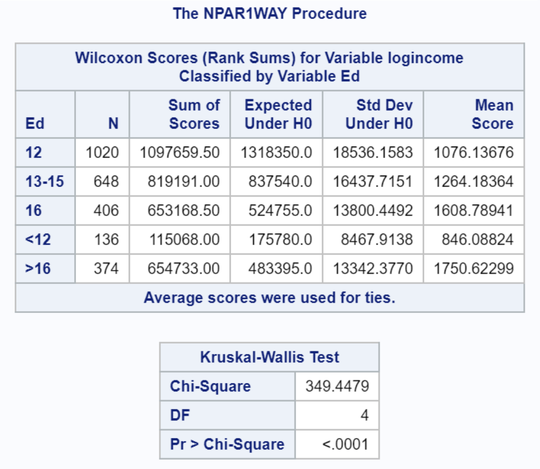
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Ed	4	2.5570	0.6393	1.38	0.2377
Error	2579	1193.2	0.4627		



Level of Ed	N	logincome	
		Mean	Std Dev
12	1020	10.2272149	0.85398541
13-15	648	10.3912107	0.92881728
16	406	10.7970859	0.95810506
<12	136	9.8993404	0.99888085
>16	374	10.8979022	1.06659104


Distribution of logincome

Test



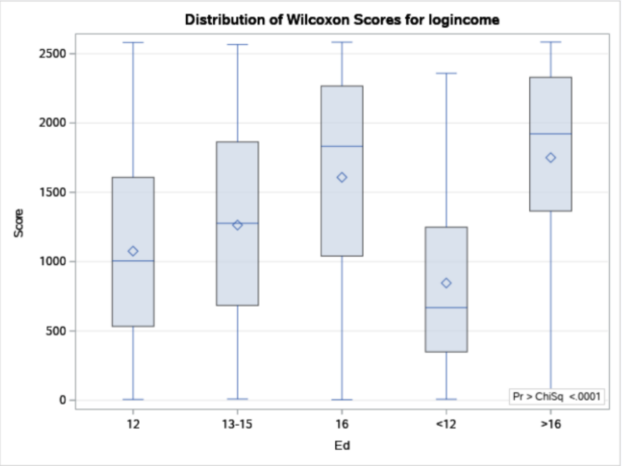
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable logincome
Classified by Variable Ed

Ed	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
12	1020	1097659.50	1318350.0	18536.1583	1076.13676
13-15	648	819191.00	837540.0	16437.7151	1264.18364
16	406	653168.50	524755.0	13800.4492	1608.78941
<12	136	115068.00	175780.0	8467.9138	846.08824
>16	374	654733.00	483395.0	13342.3770	1750.62299

Average scores were used for ties.

Kruskal-Wallis Test	
Chi-Square	349.4479
DF	4
Pr > Chi-Square	<.0001


Distribution of Wilcoxon Scores for logincome

$H_0: \mu_{<12} = median_{12} = median_{13-15} = median_{16} = median_{>16}$

$H_A: At\ least\ one\ median\ \neq$

Chi-square = 349.4479

p-value = <0.0001

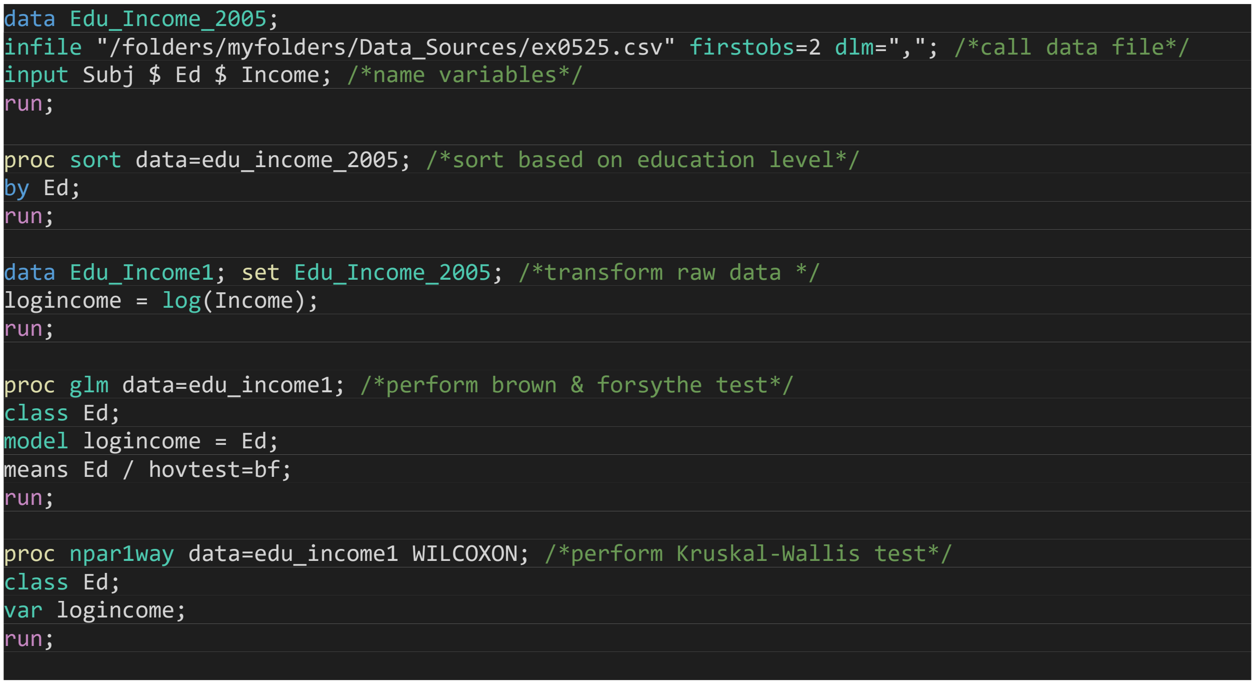
Reject H_0

The evidence suggests that at least one median is different than another (p-value = <0.0001 from Kruskal-Wallis Test).

Conclusion – There is strong evidence at the alpha = 0.05 level of significance (p-value < 0.0001 from Kruskal-Wallis Test) to support the claim that the median income for at least 1 educational level group is different that another. Further testing should be conducted to determine which group(s) has the different mean from the others.

Scope of Inference - Since this was an observational study, we are unable to establish a causal relationship between education and income. Since the group was selected from the longitudinal study in 1976 and we don't know how that group was obtained, we are only able to infer the results of this study to these participants.

Javier Saldana



```
data Edu_Income_2005;
infile "/folders/myfolders/Data_Sources/ex0525.csv" firstobs=2 dlm=","; /*call data file*/
input Subj $ Ed $ Income; /*name variables*/
run;

proc sort data=edu_income_2005; /*sort based on education level*/
by Ed;
run;

data Edu_Income1; set Edu_Income_2005; /*transform raw data */
logincome = log(Income);
run;

proc glm data=edu_income1; /*perform brown & forsythe test*/
class Ed;
model logincome = Ed;
means Ed / hovtest=bf;
run;

proc npar1way data=edu_income1 WILCOXON; /*perform Kruskal-Wallis test*/
class Ed;
var logincome;
run;
```