**Unit 5 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. 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.

**Based on the plots the samples do not have normal distributions. Also, the sample standard deviations are not the same based on the boxplots.**

**The question states that the samples are random therefore we can** **assume they are independent within each sample.**

**The categories are mutually exclusive therefore we can assume that the observations in any one sample are independent of observations in other samples.**

**After log transformation the samples appear slightly more normal and the sample standard deviations are closer. There still appears to be some skew. The transformed samples still have strong curves in the qqplots.**

|  |  |
| --- | --- |
| ***Raw Data*** | ***Log Transformed Data*** |
| FILENAME REFFILE '/folders/myfolders/Datasets/ex0525.csv';  **PROC** **IMPORT** DATAFILE=REFFILE  DBMS=CSV  OUT=WORK.Education;  GETNAMES=YES;  **DATA** Education(DROP = Subject);  SET Education;  **proc** **Sort** data=Education;  by Educ;  **proc** **univariate** data=Education;  by Educ;  histogram;  qqplot Income2005;  /\* Boxplot \*/  **proc** **boxplot** data=Education;  plot Income2005\*Educ; | **data** lEducation;  set Education;  loggedIncome2005 = log(Income2005);  **DATA** WORK.LEDUCATION(DROP = Income2005);  SET WORK.LEDUCATION;  **proc** **Sort** data=lEducation;  by Educ;  **proc** **univariate** data=lEducation;  by Educ;  histogram;  qqplot loggedIncome2005;  **proc** **boxplot** data=WORK.LEDUCATION;  plot loggedIncome2005\*Educ; |

|  |  |
| --- | --- |
| ***Raw Data*** | ***Log Transformed Data*** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. State the Problem

**Test the claim that at least one of the five population distributions (corresponding to the different years of education) is different from the others.**

1. 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.

**Based on the plots the samples do not have normal distributions. The log transformation improved this though there still appears to be some skew. The transformed samples still have curves in the qqplots.**

**We have very different sample sizes ranging from 136 to 1020. An ANOVA f-test can mitigate this by using weighted averages.**

*When the samples have different numbers of observations, a weighted average is more appropriate where the estimates from larger samples are given more weight than those from smaller samples. The pooled estimate of variance, sp2 , is a weighted average of sample variances in which each sample variance receives the weight of its degrees of freedom.*

**The sample standard deviations are not the same based on the boxplots. The F test is not robust to differing standard deviations. The log transformation improved this.**

**The question states that the samples are random therefore we can assume they are independent within each sample.**

**The categories are mutually exclusive therefore we can assume that the observations in any one sample are independent of observations in other samples.**

**The log transform made the differences in SD more acceptable however skew remains. I would consider the F-test inappropriate in this case; however, we will proceed as instructed.**

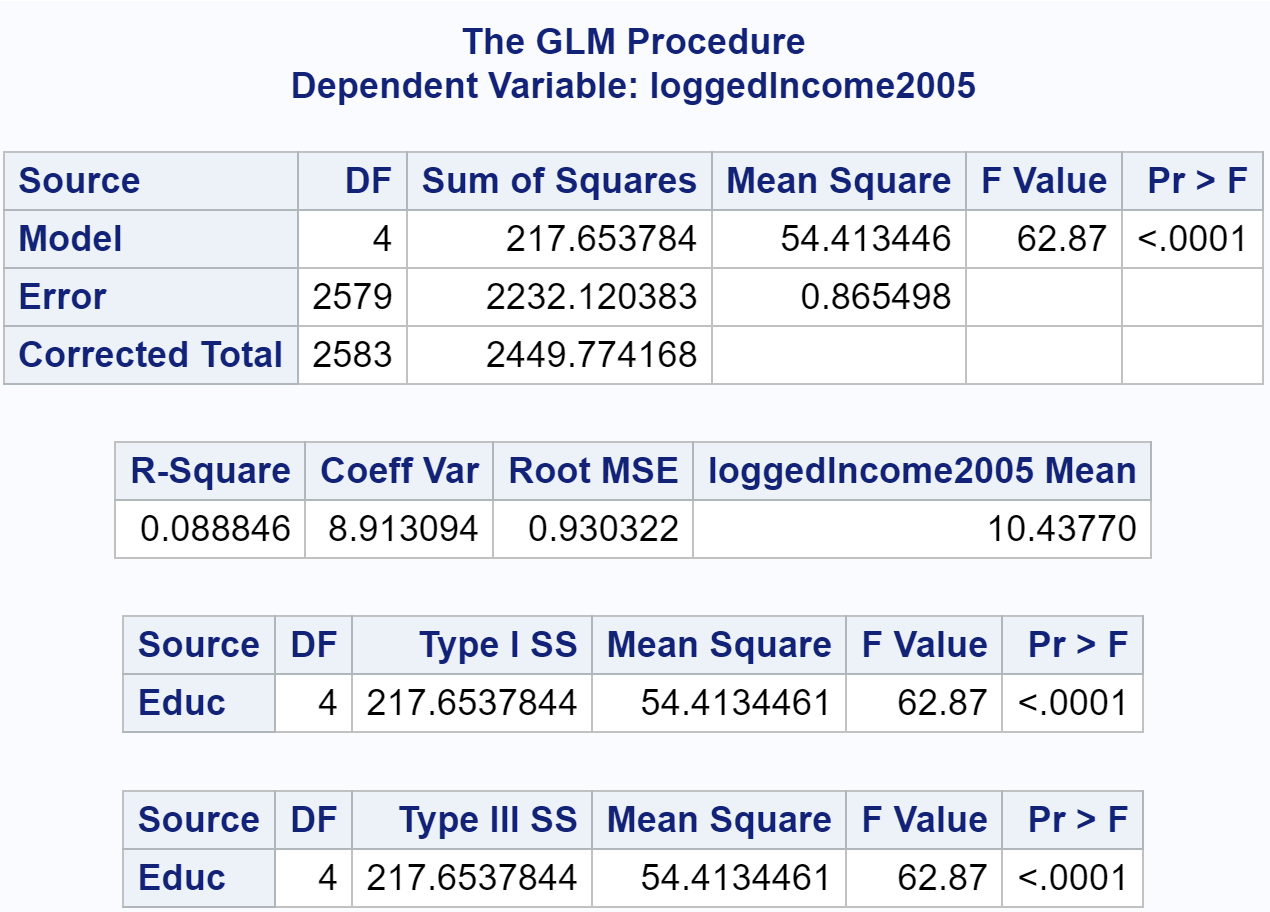
1. Conduct the Test. (An example is in the UNIT 5 PowerPoint.)
2. **H0: µ1= µ2 = µ3 = µ4 = µ5 (Equal Means)**
3. **Ha: At least 1 pair are different(Separate Means)**

**Proc** **GLM** data= WORK.LEDUCATION;

class Educ;

model loggedIncome2005 = Educ;

means Educ;



1. **Critical value:** *You can skip this step for ANOVA*
2. **F statistic = 62.87**
3. **P-value < .0001**
4. **Reject H0. The evidence suggests that at least** **1 pair of the group distributions are different (P-value < .0001 from an ANOVA) at the significance level of .05**.
5. Write a conclusion. (An example is in the UNIT 5 PowerPoint.)

**There is convincing evidence to suggest that distributions of income of at least 1 pair of the 5 education groups are different (p-value < .0001; analysis of variance F-test).**

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

**If we assume the investigator had control over the samples we can draw inference to the population.**

*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):

1. Please also identify R2

**0.088846**

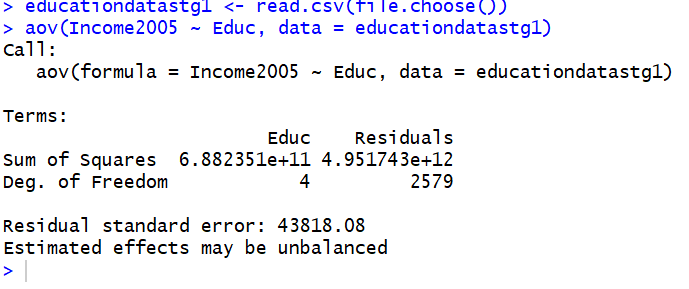
1. Also specify the mean square error and how many degrees of freedom were used to estimate it.

**0.865498, df=2579**

1. Provide the code to perform the ANOVA in R and a screen shot of the output.

educationdatastg1 <- read.csv(file.choose())

aov(Income2005 ~ Educ, data = educationdatastg1)



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) mean income to the graduate degree group (>16) mean 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 yields a p-value of .1403 (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.

1. **H0: µ16= µ>16**
2. **Ha: µ16<> µ>16**

|  |  |
| --- | --- |
| **data** Educ16;  set WORK.LEDUCATION;  if Educ ne ">16" then  if Educ ne "16" then OthersModel = "Other";  else OthersModel = "16yr";  else OthersModel = "More16yr";  **Proc** **GLM** data= Educ16;  class OthersModel;  model loggedIncome2005 = OthersModel;  **Proc** **GLM** data= Educ16;  class Educ;  model loggedIncome2005 = Educ; | |
|  |  |
|  | |

1. **Critical value: You can skip this step for ANOVA**
2. **F statistic = 17.29591, df=2581**
3. **P-value < .0001**
4. **Reject H0.**

Conclusion **There is convincing evidence to suggest that the means are different between the 16 year and >16 year groups (P-value < .0001) at the significance level of .05**

Scope **If we assume the investigator had control over the samples we can draw inference to the population.**

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

**Test the claim that at least one of the five population distributions (corresponding to the different years of education) is different from the others.**

Assumptions

**Based on the plots the samples do not have normal distributions. The log transformation improved this though there still appears to be some skew. The transformed samples still have curves in the qqplots.**

**We have very different sample sizes ranging from 136 to 1020. An ANOVA f-test can mitigate this by using weighted averages.**

**There is strong visual evidence that data have unequal variance. We will proceed under this assumption and run the Welch’s ANOVA which is robust to unequal variances.**

**The question states that the samples are random therefore we can assume they are independent within each sample.**

**The categories are mutually exclusive therefore we can assume that the observations in any one sample are independent of observations in other samples.**

Test

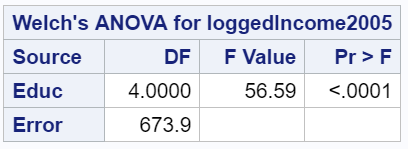
1. **H0: µ1= µ2 = µ3 = µ4 = µ5 (Equal Means)**
2. **Ha: At least 1 pair are different(Separate Means)**

**Proc** **GLM** data=Educ16;

class Educ;

model loggedIncome2005= Educ;

means Educ / hovtest=bf Welch;



1. **Critical value:** *You can skip this step for ANOVA*
2. **F statistic = 56.59**
3. **P-value < .0001**
4. **Reject H0. The evidence suggests that at least 1 pair of the group means are different (P-value < .0001)**

Conclusion

**These data provide overwhelming evidence that mean of income of at least 1 pair of the 5 education groups are different (p-value < 0:0001; Welch’s analysis of variance test).**

Scope

**If we assume the investigator had control over the samples we can draw inference to the population**

**There is convincing evidence that income changes as the level of education changes. (p-value < 0:0001; analysis of variance F-test)**