**Data Analysis Project 1**

Spring 2017

**Before you begin:**

* You may use any reference materials you wish (textbook, class notes, lectures, labs, homework, practice exercises, etc.), including the computer/internet. However, you must work on this project **by yourself** and you **may not discuss this project with anyone**. Emails to the professor for clarification of questions are the only exception to the “no discussing” rule.

**General Instructions:**

* You are required to use Stata for all data analysis.
* Your final product will be a **written report** split into four parts (plus an appendix).
* A template for this report is provided separately (**Project1\_TEMPLATE.docx**). Please use the template as the file into which you write your answers.

**Submission Instructions:**

* The project is due by **Thursday March 9, 2017, 11:59pm.**
* **No late assignments will be accepted.**
* To submit, upload your project to the “Data Analysis Project 1” Dropbox on Carmen. Please submit your project as **one file**.
* Be sure to “sign” the Honor Code pledge in your submission.

Good luck!

**Honor Code Signature:**

Joshua Wu

I pledge on my honor that I have **neither given nor received unauthorized aid on this project**. Typing my name on the line above serves as a signature of this pledge.

**Part 1: Summarizing the Sample**

*Write-up:*

**There are a total of 301 people total in this sample, with 161 people in the acupuncture group and 140 people in the control group. Within the acupuncture group, 82.61% of the group are female and 94.41% are suffering from migraines. 34.16% of the acupuncture group have had 1-10 years of chronic headaches, 30.43% of the group have had 11-25 years of chronic headaches, and 35.40% of the group have had 25+ years of chronic headaches. The baseline severity score for the acupuncture group is 24.57609.**

**Within the control group, 85.71% of the group are female and 94.29% are suffering from migraines. 29.29% of the control group have had 1-10 years of chronic headaches, 30.71% of the group have had 11-25 years of chronic headaches, and 40.00% of the group have had 25+ years of chronic headaches. The baseline severity score for the acupuncture group is 26.71071.**

**Part 2: Regression Modeling**

*Write-up:*

**In the initial model, the adjusted R2 was 0.5318. Using backward selection, I first used a Partial-F test for the years2 and years3 group. When it returned the largest p-value in the set (0.3524), I removed years2 and years3 from the model. That resulted in an Adjusted R2 value of 0.5316. Though the Adjusted R2 was smaller, another step in backward selection might raise the value. Now the value with the highest p-value is sex, with a p-value of 0.330. After removing that, the Adjusted R2 value was raised to 0.5317. Seeing if the Adjusted R2 value would rise again, another step was conducted, this time removing migraine, with a p-value of 0.279. That resulted in a Adjusted R2 value of 0.5314, however, all predictors are now below p=0.05. That will be our final model.**

**Here are the interpretations for our predictors group, age and sev\_base:**

**Group: Controlling for age and sev\_base, there is a 4.61 unit difference in sev\_1yr between those who received acupuncture and those who did not receive acupuncture. This result is significant with a p-value of 0.000.**

**Age: Controlling for group and sev\_base, there is a 0.165 unit increase in sev\_1yr for every 1 year increase in age. This result is significant with a p-value of 0.006.**

**Sev\_Base: Controlling for group and age, there is a .710 unit increase in sev\_1yr for every 1 unit increase in sev\_base. This result is significant with a p-value of 0.000.**

**There is also a intercept interpretation that is not meaningful, since it is for people of age 0, which cannot happen.**

**Part 3: Checking Model Assumptions**

*Write-up:*

**The model assumptions are Linearity, Independence, Normality, and Equal Variance.**

**We’ll start with Normality.**

****

**Looking at this QQplot of residuals, we can see that the main body of the points lie on the line, therefore Normality has been met.**

**With Equal Variance,**

****

**We can see a fan as we go down the plot. This means that this model assumption of Equal Variance has not been met.**

**For Linearity, we will be performing CPR plots for Age and Sev\_Base, since those two are the continuous variables.**

****

**The plot above is for age, as we can see, the green line is curved at the ends, not lying on the blue line. However, most of the green line lies on the blue line, so linearity has been met.**

****

**The same however, cannot be said for sev\_base. There is a massive deviation between the blue and green lines, linearity has not been met. Perhaps if we square the value, they lines will approximately lie on top of eachother, however, as of right now, they do not.**

**As for independence, the assumption has been met. It does not seem that any of the predictors significantly affect one another in any way. Therefore, independence has been met.**

**Part 4: Looking for Effect Modification**

*Write-up:*

**With a p-value of 0.004, group effect seems to be dependent on baseline severity score. By using lincom, we also see that the slope of sev\_base for those in the acupuncture group and control group are different from 0. The current model is E(Sev\_1yr | Group, sev\_base, gsevb, age) = β0 + β1xGROUP + β2xSEV\_BASE + β3xGSEVB + β4xAGE. For someone with a baseline headache severity of 10, the estimated group difference is:**

**Sev\_base = 10, Group = 1:**

**-6.219208 + 1.359897x1 + .8151097x10 – .2338409x1x10 + 0.1468896xAGE**

**Sev\_base = 10, Group = 0:**

**-6.219208 + 1.359897x0 + .8151097x10 – .2338409x0x10 + 0.1468896xAGE**

**Difference = -0.979. The group difference in mean headache severity score at one year is -0.979 units.**

**For someone with a baseline headache severity of 40:**

**Sev\_base = 40, Group = 1:**

**-6.219208 + 1.359897x1 + .8151097x40 – .2338409x1x40 + 0.1468896xAGE**

**Sev\_base = 40, Group = 0:**

**-6.219208 + 1.359897x0 + .8151097x40 – .2338409x0x40 + 0.1468896xAGE**

**Difference = -7.99. The group difference in mean headache severity score at one year is -7.99 units.**

**Since the estimates are difference between the two groups, there is effect modification.**

**APPENDIX**

*All requested Stata code here (properly formatted so readable):*

**PART 1:**





















**PART 2:**

Initial





Final





**PART 4:**

****

****