

STA441 Assignment 8

Quiz in Tutorial on Tuesday March 13th

1. Suppose you have a data set with a quantitative response variable Y , quantitative explanatory variables X_1 and X_2 , and factors A , B and C . Factor A has 3 levels (categories), factor B has 2 levels and factor C has 3 levels. You see this SAS code;

```
proc glm;  
class A B C;  
model Y = X1 X2 A|B|C;
```

- Is this an analysis of covariance? Answer Yes or No. If Yes, what are the covariates?
 - Indicate how you would define dummy variables for the factors, using effect coding. That's the setup with 1, 0 and -1. Use names like a_1 , a_2 , and so on for your dummy variables.
 - Write $E(Y|\mathbf{X}=\mathbf{x})$ for a regression model with your dummy variables, equivalent to the `proc glm` model above.
 - For the overall F-test produced by `proc glm`,
 - State the null hypothesis in terms of β quantities from your regression model.
 - How many (numerator) degrees of freedom will be in your test?
 - For each of the following effects in the `proc glm` model, state the null hypothesis in terms of β quantities from your regression model. You need not show any work. Just write down the answers.
 - Main effect of A .
 - Main effect of B .
 - Main effect of C .
 - A by B interaction.
 - A by C interaction.
 - B by C interaction.
 - A by B by C interaction.
 - Controlling for the covariates, are there any differences among the 18 treatment means? State the null hypothesis you would test in order to answer this question. Your answer should be in terms of β quantities from your regression model.
2. Now let's make that last example a little smaller, and try cell means coding. Again the quantitative response variable is Y . This time there is one quantitative covariate X , and factors A and B . Factor A has 3 levels (categories), factor B has 2 levels. The SAS code is

```
proc glm;  
class A B;  
model Y = X A|B;
```

- Indicate how you would define dummy variables, using cell means coding. That's the setup with just zeros and ones, and no intercept. You need to make a table with 6 rows, and a column for each dummy variable.
- Write $E(Y|\mathbf{X}=\mathbf{x})$ for a regression model with your dummy variables, equivalent to the `proc glm` model above. This equation has all 7 β quantities.
- Add one more column to your table, showing $E(Y|\mathbf{X}=\mathbf{x})$ for each treatment combination. My answer has only two β quantities in each row.
- Make a 3x2 (A by B) table, and write $E(Y|x)$ in each cell. This will help you answer the next question.
- For each of the following effects in the `proc glm` model, state the null hypothesis in terms of β quantities from your regression model. You need not show any work. Just write down the

answers.

- i. Main effect of A .
 - ii. Main effect of B .
 - iii. A by B interaction.
 - f. Perhaps the clearest test of a factor is to test the null hypothesis that it has no relationship with the response variable within *any* combination of the other factors. This is the null hypothesis of "conditional independence." That is, conditionally upon the values of all the other factors, the factor in question is independent of y .
 - i. What is the null hypothesis for testing whether factor A is related to y within any level of factor B ? This is the null hypothesis of conditional independence of A and y given B . State the null hypothesis in terms of β quantities from your regression model.
 - ii. What is the null hypothesis for testing whether factor B is related to y within any level of factor A ? This is the null hypothesis of conditional independence of B and y given A . State the null hypothesis in terms of β quantities from your regression model.
3. Consider a two-factor analysis of variance in which each factor has two levels. Use this regression equation for the problem:

$$E[Y|X=\mathbf{d}] = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_1 d_2$$

where d_1 and d_2 are dummy variables.

- a. Make a two-by-two table showing the four treatment means in terms of β values. Use *effect coding* (the scheme with the minus ones). In terms of the β values, state the null hypothesis you would use to test for
 - i. Main effect of the first factor
 - ii. Main effect of the second factor
 - iii. Interaction
 - b. Make another two-by-two table showing the four treatment means in terms of β values. Use *indicator dummy variables* (just zeros and ones). In terms of the β values, state the null hypothesis you would use to test for
 - i. Main effect of the first factor
 - ii. Main effect of the second factor
 - iii. Interaction
 - c. Which dummy variable scheme do you like more for this purpose? Why?
4. I know this is pretty gruesome, but the data are real -- from the U of T School of Dentistry.

An experiment in dentistry seeks to test the effectiveness of a drug (HEBP) that is supposed to help dental implants become more firmly attached to the jaw bone. This is an initial test on animals. False teeth were implanted into the leg bones of rabbits, and the rabbits were randomly assigned to receive either the drug or a saline solution (placebo). Technicians administering the drug were blind to experimental condition.

Rabbits were also randomly assigned to be "sacrificed" after either 3, 6, 9 or 12 days. At that time, the implants were pulled out of the bone by a machine that measures force in newtons and stiffness in newtons/mm. For both of these measurements, higher values indicate more healing. A measure of "pre-load stiffness" in newtons/mm is also available for each animal. This may be another indicator of how firmly the false tooth was implanted into the bone, but it might even be a covariate. Nobody can seem to remember what "preload" means, so we'll ignore this variable for now.

The data are available in the file [bunnies.data.txt](#). The variables are

1. Identification code

2. Time (3,6,9,12 days of healing)
3. Drug (1=HEBP, 0=saline solution)
4. Stiffness in newtons/mm
5. Force in newtons
6. Preload stiffness in newtons/mm

Please do the following.

- a. Classify the factors as within cases or between cases. See Lecture One Slide 29 for the definition.
- b. Use `proc freq` to find out how many rabbits are in each experimental condition.
- c. Using `proc glm`, conduct a two-way ANOVA, with force as the response variable. Use the `means` statement to get cell means and marginal means. Be prepared to answer the following questions about each of the significance tests that SAS produces by default (I count 4 default tests).
 - i. What is the value of the test statistic? The answer is a number from your printout.
 - ii. What proportion of the remaining variation is explained? Better use `proc iml`.
 - iii. What is the p-value? The answer is a number from your printout.
 - iv. Is the result statistically significant at the 0.05 level? Yes or No.
 - v. What, if anything, do you conclude? This is not the place for statistical jargon. "What do you conclude" means say something about the drug, healing, time -- something like that.
- d. Here are some questions you have already answered, partly.
 - i. Averaging across healing time, does the drug have an effect on implantation of the false teeth into the bone? Answer Yes or No. If Yes, is it possible to draw a directional conclusion without further testing?
 - ii. Averaging across drug versus placebo, does healing time have an effect on implantation of the false teeth into the bone? Answer Yes or No. If Yes, is it possible to draw a directional conclusion without further testing?
 - iii. Does the effect of drug depend on amount of healing time? Answer Yes or No. If Yes, is it possible to draw a directional conclusion without further testing?
- e. Make a table with one row for each treatment combination. Make columns showing the dummy variables for effect coding. That's the setup with 1, 0 and -1. You do *not* need to make columns for the product terms.
- f. Give $E(Y|X=x)$ for a regression model with both main effects and the interaction. Use your variable names from the table.
- g. In terms of the β values of your regression model, give the null hypothesis you would test in order to answer each of the following questions.
 - i. Averaging across time periods, is there a difference between the drug and placebo in mean force required to extract the tooth?
 - ii. Averaging across drug and placebo, does elapsed time affect the mean force required to extract the tooth?
 - iii. Does the effect of the drug depend upon elapsed time?
 - iv. Does the pattern of healing over time depend upon drug?
 - v. Is there a drug by time interaction?
- h. Now, make a table with a row for each treatment combination. Make columns showing how you would set up the dummy variables for cell means coding. That's the setup with just zeros and ones, and no intercept.
- i. Write $E(Y|X=x)$ for a regression model with your dummy variables. This equation has all 8 β quantities.
- j. Add one more column to your table, showing $E(Y)$ for each treatment combination in terms of your β quantities.
- k. Make a 2x4 (Drug by Time) table, and write $E(Y)$ in each cell. This will help you answer the next question.

- l. In terms of the β values of your regression model, give the null hypothesis you would test in order to answer each of the following questions.
 - i. Averaging across time periods, is there a difference between the drug and placebo in mean force required to extract the tooth?
 - ii. Averaging across drug and placebo, does elapsed time affect the mean force required to extract the tooth?
 - iii. Does the effect of the drug depend upon elapsed time?
 - iv. Does the pattern of healing over time depend upon drug?
 - v. Is there a drug by time interaction?
- m. Now please return to SAS. Using `proc reg` and cell means coding, test whether the drug has an effect at any time period. This is the main point of the study. You are testing the null hypothesis of conditional independence with one test. Obtain the F statistic and p-value. Do you reject H_0 ? Are the results statistically significant? What do you conclude?

To follow up, conduct tests to answer the following questions. You will Bonferroni-correct the four tests at a *joint* significance level of $\alpha=0.05$, and base any conclusions on Bonferroni-adjusted p-values.

- i. Is there a difference between Drug and Placebo at 3 days?
- ii. Is there a difference between Drug and Placebo at 6 days?
- iii. Is there a difference between Drug and Placebo at 9 days?
- iv. Is there a difference between Drug and Placebo at 12 days?

Be able to answer questions like these for each test.

- i. What is the value of the test statistic? The answer is a number.
 - ii. What is the Bonferroni-adjusted p-value? The answer is a number that you calculate with a calculator.
 - iii. Is the result statistically significant at the joint 0.05 level? Yes or No.
 - iv. What, if anything, do you conclude? This is not the place for statistical jargon. "What do you conclude" means say something about the drug, implantation of the false tooth into bone, healing, time -- something like that.
- n. Now test the same four hypotheses, except using the permutation approach to adjusting for multiple comparisons. Obtain adjusted p-values. Do your conclusions change?
 - o. Here is a question I will not ask on the quiz, but some people may find it interesting. The randomization tests you just did are fine as a partial follow-up to the test of equality among the 8 treatment combinations, but as a (complete) follow-up to the null hypothesis of conditional independence, they leave room for improvement. Why? If you had full control of the randomization process -- say, if you were writing your own code -- how would you do it better? Actually I think there is a way to do it properly with `proc multtest`, and I will show you later.
 - p. Finally, this was an animal trial. Based on the results of this experiment, do you recommend proceeding to clinical trials with humans? Answer Yes or No and briefly comment.

Please **bring your log file and your results file to the quiz**. As usual, answers to the questions are not to be handed in. They are just practice for the quiz. Please do not write anything on your printouts except your name and student number. It is okay to highlight the results file, but **do not write interpretations on your results file, or cause them to appear in any way (including comment statements) on your log file**.