# STA 2101/442f12 Assignment Nine[1]

Please bring your R printout for Question 2 to the quiz, and also your SAS log file and list file from the Question 3. It has to be the log file, not just a listing of the SAS program. The log file and the list file *must be from the same run of SAS.* Question 1 and the non-computer parts of the other questions are just practice for the quiz, and are not to be handed in. Any necessary formulas will be provided.

1. In a study comparing the effectiveness of different exercise programmes, volunteers were randomly assigned to one of three exercise programmes ($A$, $B$, $C$) or put on a waiting list and told to work out on their own. Aerobic capacity is the body's ability to process oxygen. Aerobic capacity was measured before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable was improvement in aerobic capacity. The explanatory variables were age (a covariate) and treatment group.

   (a) First consider a regression model with an intercept, and no interaction between age and treatment group.

      i. Make a table showing how you would set up indicator dummy variables for treatment group. Make Waiting List the reference category

      ii. Write the regression model. Please use $x$ for age, and make its regression coefficient $\beta_1$.

      iii. In terms of $\beta$ values, what null hypothesis would you test to find out whether, allowing for age, the three exercise programmes differ in their effectiveness?

      iv. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix.

      v. In terms of $\beta$ values, what null hypothesis would you test to find out whether Programme $B$ was better than the waiting list?

      vi. In terms of $\beta$ values, what null hypothesis would you test to find out whether Programmes $A$ and $B$ differ in their effectiveness?

      vii. Suppose you wanted to estimate the difference in average benefit between programmes $A$ and $C$ for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

      viii. Is it safe to assume that age is independent of the other explanatory variables? Answer Yes or No and briefly explain.

[1] Copyright information is at the end of the last page.

(b) Now consider a regression model with an intercept and the interaction (actually a set of interactions) between age and treatment.

i. Write the regression model. Make it an extension of your earlier model.

ii. Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of $\beta$ values, what null hypothesis would you test?

iii. Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of $\beta$ values, what null hypothesis would you test?

iv. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the **L** matrix. It is $r \times p$. What is $r$? What is $p$?

v. Suppose you wanted to know whether the difference in effectiveness between Programme $A$ and the Waiting List depends on the participant's age. In terms of $\beta$ values, what null hypothesis would you test?

vi. Suppose you wanted to estimate the difference in average benefit between programmes $A$ and $C$ for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

(c) Now consider a regression model *without* an intercept, but *with* possibly unequal slopes. Make a table to show how the dummy variables could be set up, and write the regression model. Again, please use $x$ for age and make its regression coefficient $\beta_1$. For each treatment condition, what is the conditional expected value of $Y$? The answer is in terms of $x$ and the $\beta$ values. Please put these values as the last column of your table.

i. Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of $\beta$ values, what null hypothesis would you test?

ii. Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of $\beta$ values, what null hypothesis would you test?

iii. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the **L** matrix. It is $r \times p$. What is $r$? What is $p$?

iv. Suppose you wanted to know whether the difference in effectiveness between Programme $A$ and the Waiting List depends on the participant's age. In terms of $\beta$ values, what null hypothesis would you test?

v. Suppose you wanted to estimate the difference in average benefit between programmes $A$ and $C$ for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

2. Awards received by students at a particular high school are thought to occur according to a Poisson process. That is, the numbers of awards received by students in one year are independent Poisson random variables, with mean $\lambda$ that may depend on characteristics of the student. From the Data Sets link on the course home page, you can find the Awards data. The variables are Sutudent identification code, Number of awards, Program (1=General, 2=Academic, 3=Vocational), and Score on a test of general academic knowledge. If you use `labels = c("General", "Academic", "Vocational")` in your `factor` statement, you will get nicer output.

   (a) Using `table`, make frequency table of number of awards. Does it look roughly normal?

   (b) Consider a Poisson regression model, without actually fitting it yet.

      i. What is the linear predictor? There should be no product terms (yet).

      ii. What is the link function?

      iii. Make a table with 3 rows, one for each academic program. Make columns showing how R will define the dummy variables for the variable academic program. If you're not sure, you can check your answer with R.

      iv. Add another column to your table, showing the expected number of awards given score on the math test, for each academic program.

      v. The expected number of awards for a student in the Vocational program is _______ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test.

      vi. The expected number of awards for a student in the Academic program is _______ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test.

      vii. The expected number of awards for a student in the Academic program is _______ times as great as the expected number of awards for a student in the Vocational program with the same score on the general knowledge test.

      viii. Explain why this model could be called a "proportional means" model.

      ix. Suppose we wanted to test the proportional means assumption (and it is an assumption).

         A. Write a linear model for the log of the mean for the full model you would use.

         B. State the null hypothesis. It is a statement about the $\beta$ values in the full model.

         C. What is the reduced model?

         D. What are the degrees of freedom of this test?

   (c) Now fit the proportional means Poisson regression model to the awards data. For each question below, state the null hypothesis, give the value of the test statistic ($Z$ or $\chi^2$), the $p$-value, and be able to state the conclusion in plain language. Give a *directional* conclusion if possible, even though the test is non-directional.

i. Controlling for academic program, is score on the test of general knowledge related to the expected number of awards?

ii. Controlling for score on the test of general knowledge, do students in the Academic program get more awards on average than students in the General program?

iii. Controlling for score on the test of general knowledge, do students in the Vocational program get more awards on average than students in the General program?

iv. Do any of the explanatory variables matter? You could do this with a calculator from the default output if necessary, but do it with R and get the *p*-value.

v. Controlling for score on the test of general knowledge, do students in the Vocational program get the same number of awards on average as students in the Academic program? I can't get this from the default output.

vi. The expected number of awards for a student in the Vocational program is estimated to be _______ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test.

vii. The expected number of awards for a student in the Academic program is estimated to be _______ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test.

viii. The expected number of awards for a student in the Academic program is estimated to be _______ times as great as the expected number of awards for a student in the Vocational program with the same score on the general knowledge test.

3. The Chick Weights Data was originally an R dataset called `chickwts`. In this question, we will analyze it with SAS. There is a link to the data file from the Data Sets link on the course home page.

   In this study, newly hatched chickens were randomly assigned to one of six different feed supplements, and their weight in grams after 6 weeks was recorded. I did not show you how to do this in SAS, but you can read character-valued variables by ptting a $ sign after the variable name in your input statement.

   (a) Make sure a table of means, standard deviations and sample sizes for the 6 feed types is part of your output.

   (b) Test whether the six mean weights are different. Get the *F* statistic, degrees of freedom, *p*-value and proportion of explained variation.

   (c) You want to know which means are different from which other means. Carry out the multiple comparison procedure likely to be the most powerful in this situation.

Base your conclusions on the usual $\alpha = 0.05$ *joint* significance level for the family of tests. Of course when you state your conclusions in plain language, you would not mention the significance level or joint significance level. But to be honest, stating the conclusions in plain language isn't easy. The pattern is complicated.

(d) Test for differences among mean weights for the five feed types *excluding* horsebean.

  i. First, write the null hypothesis in terms of $\mu$ values.

  ii. Now obtain the $F$ statistic, degrees of freedom and $p$-value. Do you reject $H_0$ at $\alpha = 0.05$?

(e) Obtain a 95% confidence interval for the difference between the expected weight for chicks fed horsebean, versus the average of the other expected values. Your answer is a pair of numbers.

(f) Is the test of that last contrast different from zero as a Scheffé follow-up to the over-all one-factor analysis of variance?

(g) State the conclusion from that Scheffé in plain, non-statistical language.

(h) Would you advise a chicken farmer to purchase the Horsebean feed supplement if she wanted big fat chickens?

This assignment was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/appliedf12