# STA 2101/442 Assignment 9[1]

Questions 1, 3 and 5 use R. Please bring your printouts to the quiz on Friday November 17th. The non-computer questions on this assignment are practice for the quiz, and are not to be handed in. Please do the problems using the formula sheet as necessary. A copy of the formula sheet will be distributed with the quiz if necessary. As usual, you may use anything on the formula sheet unless you are directly asked to prove it.

1. 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. We will adopt a Poisson regression model with a linear model for the natural log of $\lambda_i$. Data are given in the file http://www.utstat.toronto.edu/~brunner/data/legal/awards.data.txt.

   The variables are Student 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. Your model has no product terms, for now.

      i. 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 `contrasts`.

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

      iii. 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. Give your answer in terms of model parameters ($\beta$ quantities).

      iv. 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. Give your answer in terms of model parameters ($\beta$ quantities).

      v. 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. Give your answer in terms of model parameters ($\beta$ quantities).

vi. This model could be called a "proportional means" model, because for each fixed $x$, Expected $y$s for any two categories are in the same proportion. For example, the expected number of awards for a student in the academic program might be twice the expected number of awards of a student in the vocational program with the same general knowledge score. Assuming the model is correct, if you plotted the three curves relating academic knowledge score to expected number of awards, would the curves be parallel?

vii. 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. Some of the questions below ask for estimation, while others ask for hypothesis tests. For the estimation questions, give numbers. For the hypothesis test questions, 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 `summary` 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.

ix. Give an estimate and an approximate (large-sample) 95% confidence interval for the expected number of awards won by students in the Academic programme with a score of 80 on the knowledge test. Please do *not* use the `predict` function to get the standard error, though you can use it to check your work. Your answer is a set of three numbers.

2. In the *Heart attack data* (which you will analyze later), a sample of middle-aged men who had heart attacks were classified into three groups. Either they died of the first heart attack, or they died during the next 10 years, or they were still alive 10 years after the first attack. This is the response variable. Potential explanatory variables include age, blood pressure, and family history of heart disease (Yes-No). Let's just consider these for now. For interpretability, make the probability of being alive 10 years later the denominator in each generalized logit.

   (a) Write the multinomial logit model for these data. How many generalized logits do you have? Of course you must have a regression equation for each one.

   (b) Solve for the probabilities in terms of the *beta* values in your model. Show your work.

   (c) Make a table with two rows, one for Family history = Yes, and one for Family history = No. In each row, write *two* probability ratios. Let's call then "relative risks." (The relative risk of dying in a particular way is the probability of dying that way divided by the probability of living.)

   (d) Controlling for age and blood pressure, the relative risk of dying in the first heart attack is _____ times as great for those with a family history of coronary heart disease.

   (e) Controlling for age and blood pressure, the relative risk of dying in the next 10 years after the first heart attack is _____ times as great for those with a family history of coronary heart disease.

3. The file http://www.utstat.toronto.edu/~brunner/data/illegal/attack.data.txt. contains the *Heart attack data*, in which a sample of middle-aged men who had heart attacks were classified into three groups. Either they died of the first heart attack, or they died during the next 10 years, or they were still alive 10 years after the first attack. This is the response variable. Please make the probability of being alive 10 years later the denominator in your generalized logits.

   The variables are

   - AGE AT ENTRY TO STUDY
   - AVERAGE DIASTOLIC BLOOD PRESSURE
   - SERUM CHOLESTEROL
   - NUMBER OF CIGARETTES PER DAY (Self report)
   - HEIGHT IN INCHES
   - WEIGHT IN POUNDS
   - FAMILY HISTORY OF CORONARY HEART DISEASE
   - EDUCATION
   - OUTCOME

   Instead of height and weight, let's use Body Mass Index (BMI), defined as

$$\text{BMI} = 703 \times \frac{\text{weight}}{\text{height}^2}.$$

   A BMI under 18.5 suggests that the person is underweight, while a value over 25 may indicate that the person is overweight. The first full model (the biggest one) will include all available explanatory variables, except that height and weight will be replaced by BMI.

   (a) Fit the model, meaning estimate the parameters.
      i. Test whether *any* of the explanatory variables are useful in predicting the response variable. This is one big test. Give the value of the test statistic, the degrees of freedom, and the $p$-value. The test statistic and $p$-value are on your printout, but the degrees of freedom are not. In plain language, what do you conclude?
      ii. We should probably just give up, but let's proceed anyway for practice. If there is any hope, it looks like a model with just age, cholesterol level, and family history of heart disease. So carry out a simultaneous test of all the other explanatory variables. What is your full model? What is your reduced model? Give the value of the test statistic, the degrees of freedom, and the $p$-value. In plain language, what do you conclude?

   (b) Based on the results of the last test, I am willing to consider the model with just age, cholesterol level, and family history of heart disease. For that model, it is possible to reject the null hypothesis that the regression coefficients for all the explanatory variables equal zero? What is your full model? What is your reduced model? Give the value of the test statistic, the degrees of freedom, and the $p$-value. In plain language, what do you conclude?

(c) Now for this model with three explanatory variables, test each of the explanatory variables controlling for the other two. That's three tests. For each one, what is your reduced model? Give the value of the test statistic, the degrees of freedom, and the $p$-value. In plain language, what do you conclude?

(d) Overall, what is your assessment of this analysis?

4. 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. *Treatment group includes the waiting list control condition.*

   (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.

   (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 $\mathbf{L}$ 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.

5. Telephone sales representatives use computer software to help them locate potential customers, answer questions, take credit card information and place orders. Twelve sales representatives were randomly assigned to each of three new software packages the company was thinking of purchasing. The data for each sales representative include the software package (1, 2 or 3), sales last quarter with the old software, and sales this quarter with one of the new software packages. Sales are in number of units sold.

   The data are in the file
   http://www.utstat.toronto.edu/∼brunner/data/legal/sales.data.txt.
   The explanatory and response variables are what you would think.

   (a) Fit a full model in which the slopes and intercepts of the regression lines relating sales last quarter to sales this quarter might depend on the kind of software the sales representatives are using.

   (b) Carry out an ordinary $F$-test to determine whether the effect of software type on sales depends on the representative's performance last quarter. Be able to state your conclusion in plain, non-statistical language.

   (c) Estimate the slopes of the three regression lines. Base the estimates on numbers from your printout. I don't see how you can do this without making a table.

   (d) Carry out tests to answer these questions. If they are already on the output of `summary`, use that.

      i. Are the slopes for Software 1 and 2 different?

      ii. Are the slopes for Software 1 and 3 different?

      iii. Are the slopes for Software 2 and 3 different?

      Protecting the three tests with a Bonferroni correction at the joint 0.05 significance level, what do you conclude? Plain language is not necessary, but you should say what happened.

   (e) The average (sample mean) performance last quarter was 76.56 (please use exactly this number). We are interested in whether the three software packages differ in their effectiveness for sales representatives with average performance last quarter.

      i. Estimate expected performance this quarter for sales representatives with average performance last quarter. These three numbers should appear on your printout.

      ii. State the null hypothesis in symbols.

      iii. Carry out the $F$-test.

      iv. In plain language, what do you conclude?