STA 312f22 Assignment Ten[1]

Please bring your R printouts to the quiz. The non-computer questions are practice for the quiz on Friday Dec. 2nd, and are not to be handed in. **Bring a calculator to the quiz**.

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 λ that may depend on characteristics of the student. You can find the data here. 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.

      i. Write a regression equation for $\log(\lambda)$. There should be no product terms (yet).

      ii. 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.

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

      iv. 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.

      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 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 Vocational program with the same score on the general knowledge test.

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

      viii. 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 β values in the full model.

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

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 χ^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.

(d) Finally, test the proportional means assumption with a likelihood ratio test. Give the value of G^2, the degrees of freedom and the p-value. Do the ratios of expected numbers of awards appear to depend on the student's level of general knowledge?

2. Consider a multinomial logit model in which the outcome has four categories, and there is just one explanatory variable x. The x variable is *binary*, taking the values zero and one. Make the last response category the reference category, so its probability goes in the denominators of the probability ratios.

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

   (b) Suppose you wanted to test whether x is related to the outcome. What is the null hypothesis?

   (c) Write the probabilities in terms of the β values in your model. Put your answer in a 2×4 table, with one row for each value of x, and a column for each outcome.

3. 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 probability ratio.

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

   (b) Write the probabilities in terms of the β values in your model.

   (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.