STA 2101/442 Assignment Ten[1]

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. Data are given in the file `awards.data`. There is a link from the course home page in case the one in this document does not work.

   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. 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 R.

      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 (β 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 (β 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 (β quantities).

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

      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 β values in the full model.

         C. What is the reduced model?

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

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

(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 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 `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 prediction. Your answer is a set of three numbers. You may need to think about this one a bit.

2. Arsenic is a powerful poison, which is why it has been used on farms for many years to kill insects. Even in very small amounts, arsenic can cause cancer in humans, and recently it has been found that rice and foods made from rice tend to be very high in arsenic. Brown rice is worse, by the way.

   In a controlled experiment, pots of rice were prepared by either washing the rice first or not, and by cooking the rice in either a low, a medium or a high amount of water. The response variable is amount of arsenic in the cooked rice.

   (a) Use a regression model with *cell means coding*. That's the model with no intercept, and one indicator dummy variable for each treatment combination. You don't have to say how the dummy variables are defined. That will become clear in the next part. Just give the regression equation.

   (b) Write the expected amounts of arsenic in the table below, in terms of the β parameters of your model.

	Amount of Water		
	Low	Medium	High
Washed			
Unwashed			

   (c) If you wanted to test whether the effect of washing the rice depended on how much water you cook it in, what is the null hypothesis? Give your answer in terms of the β values in your model.

   (d) If you wanted to test whether washing the rice before cooking has any effect if the rice is cooked in a lot of water, what is the null hypothesis? Give your answer in terms of β values.

   (e) Suppose you want to test whether the amount of water used to cook the rice makes any difference if the rice has been washed. What is the null hypothesis? Give your answer in terms of β values.

   (f) Averaging across different amounts of water used to cook the rice, does pre-washing affect the amount of arsenic in the rice. What null hypothesis would you test to answer this question? Give your answer in terms of β values.

   (g) If you wanted to test whether the effect of the amount of water used to cook the rice depends on whether you wash it first, what is the null hypothesis? Give your answer in terms of β values.

3. Consider a two-factor analysis of variance in which each factor has two levels. Use this regression model for the problem:

$$Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 d_{i,1} d_{i,2} + \epsilon_i,$$

   where $d_{i,1}$ and $d_{i,2}$ are dummy variables.

   (a) Make a two-by-two table showing the four treatment means in terms of β values. Use *effect coding*. 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 a two-by-two table showing the four treatment means in terms of β values. Use *indicator dummy variables* (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?

4. In a study of math education in elementary school, equal numbers of boys and girls were randomly assigned to one of three training programmes designed to improve spatial reasoning. After five school days of training, the students were given a standardized test of spatial reasoning. Score on the spatial reasoning test is the response variable. You will define a regression model for this factorial analysis of variance. Don't write the model yet.

(a) In the table below, show how your dummy variables are defined. *Use effect coding.* That's the scheme with an intercept and minus ones. Write the name of each dummy variable at the head of its column.

Girls, Programme 1	
Girls, Programme 2	
Girls, Programme 3	
Boys, Programme 1	
Boys, Programme 2	
Boys, Programme 3	

(b) Give $E[Y_i|\mathbf{X}_i = \mathbf{x}_i]$ for the full model. Include the interaction terms. Notice you are *not* being asked to write expected values in the table. They are too messy.

(c) Suppose you want to test whether, averaging across training programmes, there is a difference between girls and boys in their average performance on the spatial reasoning test. State the null hypothesis in terms of the β values in your model.

(d) Suppose you want to test whether, averaging across boys and girls, there is a difference between training programmes in average performance on the spatial reasoning test. State the null hypothesis in terms of the β values in your model.

(e) Suppose you want to test whether the sex difference in average performance depends on which training programme the children are in. State the null hypothesis in terms of the β values in your model.