STA 312f12 Assignment Nine[1]

Please bring your R printouts to the quiz. The non-computer questions are practice for the quiz on Friday Nov. 16th, and are not to be handed in. **Bring a calculator to the quiz**.

1. The Data Sets page on our course website has a link to 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

   (a) Just to make sure you know what you're doing, please start with a small example. Fit a model with just Family History and outcome. The output from `summary` contains a test of whether the explanatory variable and response variable are related.

      i. Write the regression equations (linear predictors) for this little model, in symbols. What is the null hypothesis? This should tell you whether your degrees of freedom are correct.

      ii. To see whether you got the right numerical value of the test statistic, carry out a likelihood ratio test of independence for these data. Do it the easiest way you can. What is your value of G^2?

      iii. In plain, non-statistical language, what do you conclude from this test?

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

(b) We're almost ready to fit a big full model. But 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. Now 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. In plain language, what do you conclude?

ii. 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? For the degrees of freedom, consider subtracting the lengths of the vectors of parameter estimates.

(c) 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?

(d) 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?

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

2. In lecture, log-linear models for multinomial data were introduced for two-way and higher contingency tables. But they can also be applied to a multinomial model for a single categorical variable (like job status after graduation).

   Consider a survey in which respondents indicate their current marital status: Married, Single, Widowed or Divorced. We don't need the effect coding scheme that is so helpful for contingency tables, so we'll use indicator dummy variables.

   (a) Write a linear model for $\log \mu_j$, where mu_j is a multinomial expected value. Just give the linear predictor, which is basically a regression equation. You need not say how the dummy variables are defined; you'll do that in the next part of the question.

   (b) Now make a table with 4 rows showing how your dummy variables are defined. There will be a column for each dummy variable. Make Single the reference category. Now add a final column showing the expected number of people in each category (not the log expected value).

   (c) In terms of the β values in your model,

      i. The expected number of Married people is _____ times the expected number of Single people.

      ii. The expected number of Married people is _____ times the expected number of Divorced people.

      iii. If we randomly select one person from this population, what is the probability that the person will be married?

      iv. Suppose we want to test whether all four probabilities are equal. What is the null hypothesis in terms of the β values from your model? Are the degrees of freedom right?

3. In a sample of people walking dogs, we record (that is, we guess) the sex of the person and the sex of the dog. People walking more than one dog are excluded. These data will be analyzed as a 2×2 contingency table, and this time we will use effect coding in the log-linear regression model.

   (a) Write a linear regression model for $\log \mu_{ij}$, including the interaction of Sex of Dog by Sex of Owner.

   (b) Make a table with 4 rows showing how your dummy variables are defined. There will be a column for each dummy variable. Use effect coding. Make as many extra columns as necessary to show the product term(s) corresponding to the interaction(s). Add a last column showing $\log \mu_{ij}$ in terms of the β parameters.

   (c) Now make a 2×2 table. In the cells of the table, write the $\log \mu_{ij}$ in terms of your β parameters. Display the marginal means and the grand mean as well.

   (d) Make another 2×2 table showing the same thing in terms of the book's λ notation. How many of the λ_{ij}^{XY} interaction terms are redundant?

(e) Make a third 2×2 table showing the expected frequencies (not the log expected frequencies this time) in terms of the β parameters.

(f) True or False: When you calculate odds ratios, it does not matter whether you use probabilities or expected values.

(g) Using the last table you made, calculate the odds ratio (cross-product ratio) θ, and simplify.

(h) Show $\beta_3 = 0 \Leftrightarrow \theta = 1$

(i) To test for independence, what is the null hypothesis in terms of β values?

(j) To test for independence, what is the null hypothesis in terms of λ values?

(k) In terms of β values, what are the log odds of observing a female dog?

(l) In terms of λ values, what are the log odds of observing a female dog?
