STA 312f10 Assignment 8

Do this assignment in preparation for the quiz on Friday, Nov. 5th. Please bring your *R* printout to the quiz; part or all of it may be handed in. Please do *not* write anything on your printout before the quiz, except possibly your name and student number.[1]

1. If two events have equal probability, the odds ratio equals ____.

2. A logistic regression model with no independent variables has just one parameter, β_0. It also the same probability $p = P(Y = 1)$ for each case.

   (a) Write p as a function of β_0; show your work.

   (b) The *invariance principle* of maximum likelihood estimation says the MLE of a function of the parameter is that function of the MLE. It is very handy. Now, still considering a logistic regression model with no independent variables,

      i. Suppose $\widehat{p}$ (the sample proportion of $Y = 1$ cases) is 0.57. What is $\widehat{\beta}_0$? Your answer is a number.

      ii. Suppose $\widehat{\beta}_0 = -0.79$. What is $\widehat{p}$? Your answer is a number.

3. Consider a logistic regression in which the cases are newly married couples with both people from the same religion, the independent variable is religion (A, B, C and None – let's call "None" a religion), and the dependent variable is whether the marriage lasted 5 years (1=Yes, 0=No).

   (a) Make a table with four rows, showing how you would set up indicator dummy variables for Religion, with None as the reference category.

   (b) Add a column showing the odds of the marriage lasting 5 years. The *symbols* for your dummy variables should not appear in your answer, because they are zeros and ones, and different for each row. But of course your answer contains β values.

   (c) What is the ratio of the odds of a marriage lasting 5 years or more for Religion C to the odds of lasting 5 years or more for No Religion? Answer in terms of the β symbols of your model.

   (d) What is the ratio of the odds of lasting 5 years or more for religion A to the odds of lasting 5 years or more for Religion B? Answer in terms of the β symbols of your model.

   (e) You want to test whether Religion is related to whether the marriage lasts 5 years. State the null hypothesis in terms of one or more β values.

[1] Here are the usual suggestions about the computer work. It would be smart to compose your commands in a text file (Windows users could use Notepad), and drag them to *R* a bit at a time, debugging as you go. If I were you I would put the question numbers (but *not* the answers to the questions, please!) in comment statements. Save the text file. This way if you discover a mistake or omission, it will be easy to fix.

(f) You want to know whether marriages from Religion A are more likely to last 5 years than marriages from Religion C. State the null hypothesis in terms of one or more β values.

(g) You want to test whether marriages between people of No Religion have a 50-50 chance of lasting 5 years. State the null hypothesis in terms of one or more β values.

4. People who raise large numbers of birds inhale potentially dangerous material, especially tiny fragments of feathers. Can this be a risk factor for lung cancer, controlling for other possible risk factors? From the Data Sets link on the course home page, you can find the Bird Lung Cancer data. For a sample of birdkeepers and non-birdkeepers, it has Gender, Socioeconomic Status, Age, How many years they have been smoking (including zero), Cigarettes per day, and whether they got lung cancer.

   (a) First, produce simple but nicely-labelled one-dimensional frequency tables for the binary variables, including percentages. Obtain the means and standard deviations of age, Years smoked and Cigarettes per day.

   (b) There is one primary issue in this study: Controlling for all other variables, is birdkeeping significantly related to the chance of getting lung cancer? Perform a likelihood ratio test to answer the question.

      i. In symbols, what is the null hypothesis?
      ii. What is -2 Log Likelihood for the reduced model? The answer is a number.
      iii. What is -2 Log Likelihood for the full model? The answer is a number.
      iv. What is the value of the test statistic G? The answer is a number.
      v. What are the degrees of freedom for the test? The answer is a number.
      vi. What is the p-value? The answer is a number.
      vii. What do you conclude? Presence of a relationship is not enough. Say what happened.
      viii. For a non-smoking, bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? The answer (a single number) should be based on the full model.
      ix. For a non-smoking, non-bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? The answer (a single number) should be based on the full model.
      x. Naturally, you should be able to interpret all the Z-tests too. Which one is comparable to the main likelihood ratio test you have just done?

(c) Now, make an indicator variable for whether the person is a smoker, and include it in a stepwise logistic regression.

   i. First, start with the null model, and do stepwise variable selection with the `direction="both"` option. This is basically forward selection, but with possible removal of variables that are already in the equation.

   ii. Next start with the *full* model, and do stepwise variable selection with the `direction="both"` option. This is basically backward selection, but with possible inclusion of variables that were dropped from the equation at an earlier stage.

   iii. What is your favourite model? It may or may not be the result of a stepwise selection. Do the `summary` function on it, and be ready to interpret the output, including odds ratios.

(d) Finally, take your favourite model and test it against the full model with a likelihood ratio test, including p-value. In symbols, what is the null hypothesis? In words, what is your conclusion? Do you still have the same favourite model?