STA 312f12 Assignment Seven[1]

Please bring your R printout for the last question to the quiz. The non-computer questions are practice for the quiz on Friday Nov. 2nd, and are not to be handed in.

1. If two events have equal probability, the odds ratio equals ____.
2. For a multiple logistic regression model, if the value of the kth explanatory variable is increased by c units and everything else remains the same, the odds of Y=1 are ____ times as great. Prove your answer.
3. For a multiple logistic regression model, let $P(Y_i = 1|x_{i,1}, \ldots, x_{i,p-1}) = \pi(\mathbf{x}_i)$. Show that a linear model for the log odds is equivalent to

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

4. Write the log likelihood for the last question, and simplify it as much as possible.
5. A logistic regression model with no explanatory variables has just one parameter, β_0. It also the same probability $\pi = P(Y = 1)$ for each case.
   (a) Write π 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 explanatory variables,
      i. Suppose 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 p? Your answer is a number.
6. Consider a logistic regression in which the cases are newly married couples with both people from the same religion, the explanatory variable is religion (A, B, C and None – let's call "None" a religion), and the response 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.
   (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.

[1]

7. 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 Whether they got lung cancer (1=Yes, 0=No), Gender (0=M, 1=F), Socioeconomic Status (0=Low, 1=High), Whether they are birdkeepers (1=Yes, 0=No) Age, How many years they have been smoking (including zero), and Cigarettes per day. `help(colnames)` may be useful.

   First, make tables of the binary variables using `table`. Can you tell whether this study is retrospective, prospective or cross-sectional?

   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.

   (a) In symbols, what is the null hypothesis?
   (b) What is deviance for the reduced model? The answer is a number.
   (c) What is deviance for the full model? The answer is a number.
   (d) What is the value of the test statistic G? The answer is a number.
   (e) What are the degrees of freedom for the test? The answer is a number.
   (f) What is the p-value? The answer is a number.
   (g) What do you conclude? Presence of a relationship is not enough. Say what happened.
   (h) 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.
   (i) 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.
   (j) 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?