STA 2101/442 Assignment Nine[1]

1. Returning to the Chick Weights study, we seek to compare the weights of chickens fed `linseed`, `meatmeal` and `soybean`. This time we'll do it by discarding the other data.
   (a) Carry out an ordinary F-test.
   (b) Do the same thing with a randomization test, obtaining a randomization p-value. Is it close to what you got from the F-test?
   (c) Of course your randomization p-value is merely an *estimate* of the permutation p-value. Give an approximate 99% confidence interval for the permutation p-value.

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

3. 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.

4. 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

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

5. Write the log likelihood for a general logistic regression model, and simplify it as much as possible. Of course use the result of the last question.

6. 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 $\overline{y}$ (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 $\overline{y}$? Your answer is a number.

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

7. 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.

8. 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?

   The data are available in the file `birdlung.data`. There is a link from the course home page in case the one in this document does not work. In this question, you will analyze the data with R.

   For a sample of birdkeepers and non-birdkeepers, the data file 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. If you look at `help(colnames)`, you can see how to add variable names to a data frame. It's a good idea, because if you can't remember which variables are which during the quiz, you're out of luck.

   First, make tables of the binary variables using `table`, Use `prop.table` to find out the percentages. What proportion of the sample had cancer. Any comments?

   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 the value of the likelihood ratio test statistic G^2? The answer is a number.

(c) What are the degrees of freedom for the test? The answer is a number.

(d) What is the p-value? The answer is a number.

(e) What do you conclude? Presence of a relationship is not enough. Say what happened.

(f) 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.

(g) 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.

(h) Obtain a 95% confidence interval for that last probability. Do it the easiest way you can. Your answer is a pair of numbers.

(i) 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?

(j) Also, are *any* of the explanatory variables related to getting lung cancer? Carry out a single likelihood ratio test. You could do it from the default outut with a calculator, but use R. Get the p-value, too.

(k) Now please do the same as the last item, but with a Wald test.

9. Finally and just for practice, fit a simple logistic regression model in which the single explanatory variable is number of cigarettes per day. When a person from this population smokes ten more cigarettes per day, the odds of lung cancer are multiplied by r (odds ratio). Give a point estimate of r. Your answer is a number.
