# STA 302 Assignment Eleven[1]

1. If the odds of an event are 4 to 1, what is the probability of the event?

2. If the odds of an event are 0.25 to 1, what is the probability of the event?

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

4. For a multiple logistic regression model, if the value of the $j$th 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.

5. 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

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_{i,1}+\ldots+\beta_k x_{i,k}}}{1 + e^{\beta_0+\beta_1 x_{i,1}+\ldots+\beta_k x_{i,k}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

6. A logistic regression model with no independent variables has just one parameter, $\beta_0$. It also the same probability $\pi = P(Y = 1)$ for each case.

   (a) If $\pi = 0.57$, what is $\beta_0$? Your answer is a number.

   (b) Suppose $\beta_0 = -0.79$. What is $\pi$? Your answer is a number.

7. Let $Y = 1$ if a high school student finishes university, and zero otherwise. Let $x = 1$ if at least one of the student's parents finished university, and zero otherwise. Consider a logistic regression for this situation.

   (a) In terms of probabilities, what would $\beta_0 = 0$ mean?

   (b) In terms of probabilities, what would $\beta_1 = 0$ mean?

   (c) In terms of the $\beta$ parameters of your model, what is the conditional probability that a student will finish university given that at lest one of his parents finished university?

   (d) The odds of finishing university are __ times as great if at least one parent has finished university.

   (e) Would you expect $\beta_1$ to be positive, negative or zero? Why?

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

8. Consider a logistic regression in which the cases are newly married couples, with both people from the same religion. The dependent variable is whether the marriage lasted 5 years (1=Yes, 0=No). The independent variables are annual household income in thousands of dollars (call it $x$) and Religion (A, B, C and None – let's call "None" a religion). There are no interactions in this model.

   (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 $\beta$ values. It also contains the symbol $x$ for household income.

   (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 $\beta$ symbols of your model. Is $x$ part of the answer?

   (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 $\beta$ symbols of your model. Is $x$ part of the answer?

   (e) You want to test whether Religion is related to whether the marriage lasts 5 years, controlling for income. State the null hypothesis in terms of one or more $\beta$ values.

   (f) You want to know whether marriages from Religion A are more likely to last 5 years than marriages from Religion C, controlling for income. State the null hypothesis in terms of one or more $\beta$ values. Is $x$ part of the answer?

   (g) You want to test whether marriages between people of No Religion with a household income of \$50,000 have a 50-50 chance of lasting 5 years. State the null hypothesis in terms of one or more $\beta$ values.

9. 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. Does this seem odd?

   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. That's full versus reduced model.

(a) In symbols, what is the null hypothesis?

(b) What is the value of the likelihood ratio chi-square test statistic? 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) Naturally, you should be able to interpret all the $Z$-tests. Which one is comparable to the main likelihood ratio test you have just done?

(i) Also, are *any* of the independent 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.

(j) Controlling for all other variables, when a person from this population smokes ten more cigarettes per day, the odds of lung cancer are multiplied by ___ (odds ratio). Give a point estimate of the odds ratio. Your answer is a number.