STA 2101/442 Assignment 6[1]

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

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

4. 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.

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 $\overline{y}$ (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.

6. Consider a logistic regression in which the cases are newly married couples with both people from the same religion. The explanatory variables are total family income and religion. Religion is coded 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) Write a linear model for the log odds of a successful[2] marriage. You do not have to say how the dummy variables are defined. You will do that in the next part.

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

   (c) 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. Denote income by x.

   (d) For a constant value of income, 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.

   (e) Holding income constant, 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.

[1] This assignment was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/appliedf18

[2] I agree, this may be a modest definition of success.

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

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

(h) You want to test whether marriages between people of No Religion with an average income have a 50-50 chance of lasting 5 years. State the null hypothesis in symbols. To hold income to an "average" value, just set $x = \overline{x}$.

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? Which of those other possible risk factors are important? Here are the variables in the file `http://www.utstat.utoronto.ca/~brunner/data/illegal/birdlung.data.txt`. These data are from a textbook called the *Statistical Sleuth* by Ramsey and Schafer, and are used without permission.

Variable	**Values**
Lung Cancer	1=Yes, 0=No
Gender	1=Female, 0=Male
Socioeconomic Status	1=High, 0=Low
Birdkeeping	1=Yes, 0=No
Age	
Years smoked	
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? Carry out 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) Obtain a 95% confidence interval for that last probability. Your answer is a pair of numbers. There is an easy way and a hard way. Do it the easy way.

(h) Your answer to the last question made you uncomfortable. Why? Another approach is to start with a confidence interval for the log odds, and then use the fact that the function $p(x) = \frac{e^x}{1+e^x}$ is strictly increasing in x. Get the confidence interval this way. Again, your answer is a pair of numbers. Which confidence interval do you like more?

(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) Controlling for all other variables, are the chances of cancer different for men and women?

(k) All other things being equal, when a person smokes 10 more cigarettes a day (ten, not one), the estimated odds of cancer are _____ times as great.

(l) 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 output with a calculator, but use R. Get the p-value, too.

(m) Now please do the same as the last item, but with a Wald test. Of course you should display the value of W_n, the degrees of freedom and the p-value.

(n) Finally and just for practice, fit a simple logistic regression model in which the single explanatory variable is number of cigarettes per day.

i. 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.

ii. Using the `vcov` function and the delta method, give an estimate of the asymptotic variance of r. Your answer is a number.

8. Awards received by students at a particular high school are thought to occur according to a Poisson process. That is, the numbers of awards received by students in one year are independent Poisson random variables, with mean λ that may depend on characteristics of the student. We will adopt a Poisson regression model with a linear model for the natural log of λ_i. Data are given in the file
`http://www.utstat.toronto.edu/~brunner/data/legal/awards.data.txt`.

The variables are Student identification code, Number of awards, Program (1=General, 2=Academic, 3=Vocational), and Score on a test of general academic knowledge. If you use `labels = c("General", "Academic", "Vocational")` in your `factor` statement, you will get nicer output.

(a) Using `table`, make frequency table of number of awards. Does it look roughly normal?

(b) Consider a Poisson regression model, without actually fitting it yet. Your model has no product terms.

i. Make a table with 3 rows, one for each academic program. Make columns showing how R will define the dummy variables for the variable academic program. If you're not sure, you can check your answer with `contrasts`.

ii. Add another column to your table, showing the expected number of awards given score on the academic knowledge test, for each academic program.

iii. The expected number of awards for a student in the Vocational program is ________ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test. Give your answer in terms of model parameters (β quantities).

iv. The expected number of awards for a student in the Academic program is ________ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test. Give your answer in terms of model parameters (β quantities).

v. The expected number of awards for a student in the Academic program is ________ times as great as the expected number of awards for a student in the Vocational program with the same score on the general knowledge test. Give your answer in terms of model parameters (β quantities).

vi. Now fit your Poisson regression model to the awards data. Some of the questions below ask for estimation, while others ask for hypothesis tests. For the estimation questions, give numbers. For the hypothesis test questions, state the null hypothesis, give the value of the test statistic (Z or χ^2), the p-value, and be able to state the conclusion in plain language. Give a *directional* conclusion if possible, even though the test is non-directional.

A. Controlling for academic program, is score on the test of general knowledge related to the expected number of awards?

B. Controlling for score on the test of general knowledge, do students in the Academic program get more awards on average than students in the General program?

C. Controlling for score on the test of general knowledge, do students in the Vocational program get more awards on average than students in the General program?

D. Do any of the explanatory variables matter? You could do this with a calculator from the default output if necessary, but do it with R and get the p-value.

E. Controlling for score on the test of general knowledge, do students in the Vocational program get the same number of awards on average as students in the Academic program? I can't get this from the `summary` output.

F. The expected number of awards for a student in the Vocational program is estimated to be ________ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test.

G. The expected number of awards for a student in the Academic program is estimated to be ________ times as great as the expected number of awards for a student in the General program with the same score on the general knowledge test.

H. The expected number of awards for a student in the Academic program is estimated to be ________ times as great as the expected number of awards for a student in the Vocational program with the same score on the general knowledge test.

I. Give an estimate and an approximate (large-sample) 95% confidence interval for the expected number of awards won by students in the Academic programme with a score of 80 on the knowledge test. Please do *not* use the `predict` function to get the standard error, though you can use it to check your work. Your answer is a set of three numbers.