STA 2101 Assignment 5[1]

The questions on this assignment are not to be handed in. They are practice for Quiz Five on Friday October 25th. Please bring your printout for Question 7 to the quiz.

Your printouts should have *only* R input and output. It is okay to have the questions in comment statements, but *no answers*.

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.

[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/2101f19

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.

   (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 $x = \overline{x}$.

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

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) 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.

(l) 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.

(m) 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.

Please bring your R printout for this question to the quiz. Also, this question requires some paper and pencil work, and it would be fair to ask for something like that on the quiz too.

8. In the usual multiple regression model, the X matrix is an $n \times p$ matrix of known constants. But in practice, the explanatory variables are often random and not fixed. Clearly, if the model holds *conditionally* upon the values of the explanatory variables, then all the usual results hold, again conditionally upon the particular values of the explanatory variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathcal{X} = X$. Here, the $n \times p$ matrix $\mathcal{X}$ is used to denote the matrix containing the random explanatory variables. It does not have to be *all* random. For example the first column might contain only ones if the model has an intercept.

(a) Show that the least-squares estimator $(X^\top X)^{-1}X^\top \mathbf{y}$ is conditionally unbiased. You've done this before.

(b) Show that $\widehat{\boldsymbol{\beta}} = (\mathcal{X}^\top \mathcal{X})\mathcal{X}^\top \mathbf{y}$ is also unbiased unconditionally. If it helps, you may assume that the explanatory variables are discrete, so you can write a multiple sum. Or, you could do it using just expected values.

(c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an F-test comparing full and reduced models), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the explanatory variable values. That is, $P(F > f_c|\mathcal{X} = X) = \alpha$. Find the *unconditional* probability of a Type I error. Again, you may assume that the explanatory variables are discrete if you wish. I used the so-called Law of Total Probability.

9. Ordinary linear regression is often applied to data sets where the explanatory variables are best modeled as random variables: write $y_i = \boldsymbol{\mathcal{X}}_i^\top \boldsymbol{\beta} + \epsilon_i$. In what way does the usual conditional linear regression model with normal errors imply that random explanatory variables must be independent of the error term? Hint: assume the random vector $\boldsymbol{\mathcal{X}}_i$ and the scalar random variable ϵ_i are both continuous. What is the conditional distribution of ϵ_i given $\boldsymbol{\mathcal{X}}_i = X_i$?

10. For a model with just one (random) explanatory variable, show that $E(\epsilon_i|X_i = x_i) = 0$ for all x_i implies $Cov(X_i, \epsilon_i) = 0$, so that a standard regression model without the normality assumption still implies zero covariance (though not necessarily independence) between the error term and explanatory variables.