The questions are practice for the quiz next week, and are not to be handed in. I would like you to bring in all of the code you used to run this assignment, and the output it generates. The best way to do this is just to run all of your code, and copy the console output to a .txt file and print it out. Clearly, your code should not print things that are not asked for, like an entire data frame. Even though you might want to look at the data frame while writing your code, your final version should only print the answers to the questions. Your console output (with code) should fit on two pages, double-sided, with two columns per side, at most. If your output is longer than 8 single pages (about 400 lines) you could probably tidy things up a bit. Remember, a whole lot of nonsense is not the same thing as concise, tidy code, even if the conclusions are the same.

If you want to experiment with R Markdown files to make it look nicer, be my guest! I might show these a little later in the course.
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 $P(Y_i = 1 | x_{i,1}, \ldots, x_{i,p-1}) = \pi(x_i)$. Show that a linear model for the log odds is equivalent to

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

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, $\beta_0$. It also the same probability $\pi = P(Y = 1)$ for each case.

   (a) Write $\pi$ as a function of $\beta_0$; 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 $\bar{y}$ (the sample proportion of $Y = 1$ cases) is 0.57. What is $\hat{\beta}_0$? Your answer is a number.

      ii. Suppose $\hat{\beta}_0 = -0.79$. What is $\bar{y}$? 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 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 $\beta$ 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 $\beta$ 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 $\beta$ 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 \( \beta \) 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 \( \beta \) 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 \( \beta \) values.

7. Consider again the popsicle stick dataset from the last assignment. We want to model the probability of a glue break using the covariates available, including some new ones we’ll compute. See the posted R code from last week to use these.

(a) First, start with just the glue types and look only at beige, 2-stick, clamped joints with a medium overlap.
   i. Is glue type a significant predictor of the probability of a glue break?
   ii. Were there any problems fitting this model? What can you conclude?

(b) Let’s look at the beige, 2-stick, clamped, medium overlap joints NOT made with Hot glue.
   i. Is glue type a significant predictor of the probability of a glue break?
   ii. The odds of a glue break are \( \text{___} \) times as great for White glue compared to Carpenters glue.
   iii. The odds of a glue break are \( \text{___} \) times as great for White glue compared to Gorilla glue.

(c) Now fit a model to the beige, 2-stick, medium overlap joints NOT made with Hot glue using glue type, clamp, quality and stress as predictors of whether the joint broke with a glue break or not.
   i. Obtain 95% CIs for each odds ratio under this model, and display this table. The CIs should be for the OR, not the log(OR).
   ii. We are 95% confident that the odds of a glue break are between \( \text{___} \) and \( \text{___} \) \( \text{times as great} \) for White glue compared to Carpenter’s glue, controlling for clamping, quality and stress.
   iii. We are 95% confident that the odds of a glue break are between \( \text{___} \) and \( \text{___} \) \( \text{percent greater} \) for White glue compared to Carpenter’s glue, controlling for clamping, quality and stress.
   iv. Is stress a significant predictor of break type, controlling for quality, clamping and glue type? Give the p-value.
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`.

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.

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 bird-keeping 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.


(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 by first finding a CI for the Odds and then converting it to one for probabilities (this is not necessarily how it should be done). 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? Any idea why it doesn’t match?

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

Remember, the computer assignments in this course are not group projects. You are expected to do the work yourself. You may compare numerical answers but do not share R code!