STA312f10 Assignment 7

Please do the following questions in preparation for the quiz on Friday Oct. 29th. Questions are preparation for the quiz; they are not to be handed in.

1. In a study attempting to predict starting salary from linguistic background, the first language of new employees is classified as English, French or Other.

a. Write a regression equation with dummy variables for predicting starting salary just from linguistic background. You need not say how the dummy variables are defined; you will do that in the next part. Just complete this:

$$E[Y|\mathbf{x}] \ =$$

b. In the table below, define your indicator dummy variables for First Language. Make English the reference category. Also, write $\hat{Y}$ for each category.

	Dummy Variables	
		$E[Y|\mathbf{x}]$
English		
French		
Other		

c. To test whether employees whose first language is French have a different average starting salary than employees whose first language is English, you would test H_0:

d. To test whether employees whose first language is Other have a different average starting salary than employees whose first language is English, you would test H_0:

e. To test whether employees whose first language is Other have a different average starting salary than employees whose first language is French, you would test H_0:

2. In a study of how people may get sick by staying in hospital, the cases are hospitals, and the dependent variable is "Infection risk," the (estimated) probability of getting sick in hospital. Two variables of interest are Age (average age of patient in the hospital) and Geographic Region in the U. S..

a. Representing infection risk by Y, age by the variable x, and three dummy variables by D1, D2 and D3, write a regression equation with an intercept and 4 independent variables. Complete the following equation: E[Y|x,**D**] =

b. In the table below, show how indicator dummy variables for geographic region would be set up so that **Northeast** is the reference category. Write E[Y|x,**D**] for each region. Of course the symbols "D1," D2" and "D3" should *not* appear in your expression for E[Y|x,**D**], because they equal zero or one.

Region	D1	D2	D3	E[Y|x,**D**]
NORTHEAST				
N. CENTRAL				
SOUTH				
WEST				

c. For the West region, when average patient age is increased by one year, expected infection risk increases by _______.

d. Controlling for average patient age, the difference between expected infection risk in the Northeast and South regions is ____.

e. Suppose we simultaneously tested the regression coefficients for D1, D2 and D3, and the test was significant at the 0.05 level. What would you conclude? Use plain, non-technical language.

3. For a multiple logistic regression model, let $P(Y_i=1|\, x_{i,1}, \ldots x_{i,p-1}) = p(\mathbf{x}_i)$. Show that a linear model for the log odds is equivalent to

$$p(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

4. Write the log of the likelihood function for Question 3, and simplify it as much as possible.

5. For a multiple logistic regression model, if the value of the kth independent variable is increased by c units (c>0) and everything else remains the same, the odds of Y=1 are _____ times as great. Prove this and show your work.

6. In a market research study, the dependent variable is whether or not a consumer has done any shopping in the United States during the past 30 days. One of the independent variables is place of birth, coded as 1=Canada, 2=United States, 3=Asia, 4=Europe, 5=Central or South America (including Mexico), and 6=other. Make up a table of dummy variables, chosen so that for a consumer born in location k+1, the odds of shopping in the U.S. are e^{β_k} times as great as the odds of a Canadian-born consumer shopping in the US. You don't need to prove anything; just write down the table of dummy variables.

7. Consider a logistic regression in which the cases are newly married couples with both people from the same religion, the independent variable is religion (*A*, *B*, *C* and *None* -- let's call "None" a religion), and the dependent 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 years. The symbols for your dummy variables should not appear in your answer, because they are zeros and ones, and different for each row.

C. What is the ratio of the odds of 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.