STA429/1007 Assignment 3

First, read pages 61-76 in Chapter 3 of the online text. Then do the following questions in preparation for Quiz 3 on Thursday Oct. 11th.. Numbers one through three are not to be handed in. The log and list files from Question 4 may be handed in. Please bring them to the quiz.

1. 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. In the table below, set up indicator dummy variables for geographic region so that SOUTH is the reference category.

Region	D_1	D_2	D_3
NORTHEAST			
NORTH CENTRAL			
SOUTH			
WEST			

b. Representing infection risk by Y, age by the variable x, and your three dummy variables by D_1, D_2 and D_3, write a regression equation with an intercept and 4 independent variables. Complete the equation below (put x before the dummy variables).

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

c. Give E[Y|X] for each region. The symbols "D_1," D_2" and "D_3" should not appear in your answer.

Region	$E[Y|\mathbf{X}]$
NORTHEAST	
NORTH CENTRAL	
SOUTH	
WEST	

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

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

f. For any region, when average patient age is increased by one year, expected infection risk increases by _______.

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

h. Controlling for average patient age, the difference between expected infection risk in the North Central and South regions is ____.

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

j. What does β_0 mean?

k. Suppose we simultaneously tested D_1, D_2 and D_3 (or equivalently, β_2, β_3, and β_4), and the test was not significant. If you were in an exploratory mode and allowing yourself to accept the null hypothesis, what would you conclude?

l. Is this study experimental, observational, or both? Why?

m. Suppose the results in (k) were statistically significant. Could you conclude that the difference in infection risk was caused by differences in how hospitals are run in the different regions? Why or why not?

2. In a government study of Canadian business, companies were classified as either heavy manufacturing (type=1), light manufacturing (type=2), retail (type=3) or service (type=4). The size of each company was also recorded, as well as the profit after taxes.

a. Make a table showing how you would set up indicator dummy variables for type of business.

b. You want to know whether average profit after taxes is related to type of business, once you control for size of company.

i) Give $E[Y|\mathbf{X}]$ for the full model.

ii) Give $E[Y|\mathbf{X}]$ for the reduced model.

3. In a study of the fuel efficiency of automobiles, investigators selected independent random samples of automobiles located in Canada and manufactured in either (1) North America, (2) Japan, (3) Europe, or (4) Other location. The dependent variable Y is kilometers per litre.

a. Write this as a multiple regression model with r-1 dummy variables; call them x_1, x_2 and x_3. You do not need to define how the dummy variables are coded; you are asked to do that in the next part of this question. In fact, all you need to do is complete this:

E[Y|**X**] =

b. In the table below, define dummy variables for location of car's manufacture. Make North American the reference category.

Country of Origin	X_1	X_2	X_3
1 = North America			
2 = Japan			
3 = Europe			
4 = Other			

c. The difference in average fuel efficiency between Japanese and European cars is ____.

d. You want to know whether average fuel efficiency is related to location.

i) Give E[Y|**X**] for the full model.

ii) Give E[Y|**X**] for the reduced model.

4. For the trees data, first run proc means to get means and standard deviations of all the variables. Then fit a multiple regression model predicting volume from diameter and height.

a. Part of the default output is a simultaneous test for height and diameter. Give the numerical value of the test statistic and the corresponding p-value. Your answer to this question is a pair of numbers.

b. What is your interpretation of the finding in part (a)? The words “height,” “diameter” and “volume” should be prominent in your answer.

c. What proportion of the variation in volume is explained by height and diameter together? The answer to this question is a singe number. My answer is 0.948.

d. If you control for diameter, is height still a significant predictor of volume? Answer Yes or No, and give 2 numbers: the numerical value of the test statistic, and the p-value.