STA429/1007 Assignment 4

Do Questions 1 through 3 in preparation for Quiz 4 on Thursday Oct. 21st. They are not to be handed in. For question 4, please bring both your log file and your list file to the quiz. It is okay to write the answer to 4f on your printout. In fact, it is is recommended

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	D1	D2	D3
NORTHEAST			
NORTH CENTRAL			
SOUTH			
WEST			

b. Representing infection risk by Y, age by the variable x, and your three dummy variables by D1, D2 and D3, write a regression equation with an intercept and 4 independent variables. Complete the equation below.

$$E[Y] =$$

c. Give E[Y] for each region. The symbols "D1," D2" and "D3" should *not* appear in your answer.

Region	E[Y]
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 D1, D2 and D3 (or equivalently, β_2, β_3, and β_4), and the test was not significant. 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] for the full model.

ii) Give E[Y] 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:

$$E[Y] =$$

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 = N. 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] for the full model.

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

4. Please refer to the TV data from the last assignment. Create a new variable representing total number of people in a household. Also, make indicator dummy variables for Location (that City versus Town versus Rural variable). The question we want to answer is this: Controlling for number of people in the household, location, value of home and number of TV sets, is total number of TV hours watched last week related to Price willing to pay for cable TV? As usual, we will assume that the 0.05 level is set in stone.

a. Give the value of the test statistic. The answer is a number. More than one right answer is possible, but they are equivalent.

b. Give the p−value. The answer is a number.

c. Are the results statistically significant at the 0.05 level? Answer Yes or No.

d. Controlling for the variables in the reduced model, is the sample relationship between TV hours watched and Price willing to pay positive, or negative? How can you tell? You can answer this question even if the results are not significant, because you are being asked about the *sample* relationship.

e. What proportion of the variation in the dependent variable is explained by the full model? The answer is a number.

f. After allowing for the variables in the reduced model what proportion of the *remaining* variation is explained by total TV hours watched? The answer is a number. You will need a calculator for this question. It would be a good idea to write the answer on your printout in advance.

g. In the simplest language possible, what do you conclude from this analysis. Please start your answer with "When we allow for ..."