**Statistical Learning 3**

1. in linear model is to describe the random component of linear relationship of X and Y, but in the logistic regression we don’t have this “error” term. Why?

2. In linear regression, do we treat X as random variables or not? Why?

3. In classification problems, we usually use 0 and 1 to label classes (when we have only two classes, as Purchase variable in case 1 of lecture 3), but if we label them as -1 and 1, how to handle this change? Will the result be the same? Why?

4. Please read section 3.3.3 (page 92 of textbook) about potential problems in linear regression, and answer the following questions:

(i) When we fit a linear model we assume that the predictors X and the response Y have a straight-line relationship, but what if there are actually non-linear associations in the data. How to handle this non-linearity?

(ii) Please give an example that there are correlations among the error terms in linear model, and specify how this problem will affect the accuracy of fitted linear model.

(iii) How to identify heteroscedasticity in the data and how to deal with it?

(iii) What’s the difference between outlier and high leverage point? How to detect them?

5. This problem focuses on the *collinearity* problem.

(a) Perform the following commands in R:

*> set .seed (1)*

*> x1=runif (100)*

*> x2 =0.5\* x1+rnorm (100) /10*

*> y=2+2\* x1 +0.3\* x2+rnorm (100)*

The last line corresponds to creating a linear model in which *y* is a function of *x1* and *x2*. Write out the form of the linear model. What are the regression coefficients?

(b) What is the correlation between *x1* and *x2*? Create a scatterplot displaying the relationship between the variables.

(c) Using this data, fit a least squares regression to predict *y* using *x1* and *x2*. Describe the results obtained. What are 0, 1, and 2? How do these relate to the true β0, β1, and β2? Can you reject the null hypothesis H0 : β1 = 0? How about the null hypothesis H0 : β2 = 0?

(d) Now fit a least squares regression to predict *y* using only *x1*. Comment on your results. Can you reject the null hypothesis H0 : β1 = 0?

(e) Now fit a least squares regression to predict *y* using only *x2*. Comment on your results. Can you reject the null hypothesis H0 : β1 = 0?

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

(g) Now suppose we obtain one additional observation, which was unfortunately mis- measured.

*> x1=c(x1 , 0.1)*

*> x2=c(x2 , 0.8)*

*> y=c(y,6)*

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.