INFO 6105
Data Science Engineering Methods and Tools

Northeastern University, Fall 2019

Problem Set 1, Due: Oct 12, 2019

Problem Set Rules:

1. Each student should hand in an individual problem set at the beginning of class.
2. Discussing problem sets with other students is permitted. Copying from another person or solution set is *not* permitted.
3. Late assignments will *not* be accepted. No exceptions.

1. (Total: 50 points)

   In this question, you should use the Carseats data set to predict the sales in a new store with Price=$120, Advertising=$10000, ShelveLoc = Good, 'Urban=Yes, US=Yes.

   (a) (4 points) Fit a multiple regression model to predict Sales using Price, Advertising Urban, and US. Write out the model in equation form, being careful to handle the qualitative variables properly.

(b) (4 points) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!

(c) (8 points) Using the model from (a), predict sales in the new store and calculate 68% and 95% confidence intervals.

(d) (8 points) Using the model from (a), what is the probability that sales will be greater than 12000 units in the new store?

(e) (8 points) Using the model from (a), what is the probability that sales will be between 6000 and 10000 units in the new store?

(f) (5 points) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(g) (8 points) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. Using this model, predict sales in the new store and calculate 68% and 95% confidence intervals.

(h) (5 points) How well do the models in (a) and (g) fit the data?

2. (Total: 50 points) This problem involves the sales data set for Toyota Corolla, which can be found in the file ToyotaCorolla.csv. The data set contains 1436 observations on the following 10 variables.

   Price (in Dollars)

   Age (in months)

   Mileage

   FuelType Fuel Type (diesel, petrol, CNG)

   MetColor Metallic color (1=yes, 0=no)

   Automatic Automatic transmission (1=yes, 0=no)

   Displacement Engine displacement (in cu. inches)

   Doors Number of doors

   Weight (in pounds)

   Horsepower Engine horsepower

   (a) (3 points) Which of the predictors are quantitative, and which are qualitative?

   (b) (3 points) What is the range (i.e., min and max) of each quantitative predictor?

(c) (3 points) What is the mean and standard deviation of each quantitative predictor?

(d) (5 points) Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(e) (4 points) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables.

Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

(f) (8 points) Fit a simple linear regression with Price as the response and Age as the predictor.

(i) Is there a relationship between the predictor and the response?

(ii) How strong is the relationship between the predictor and the response?

(iii) What is the predicted price associated for a car with an age of 48 months? What are the associated 95% confidence intervals?

(g) (12 points) Fit a multiple linear regression with Price as the response and all other variables the predictors.

(i) Is there a relationship between the predictors and the response?

(ii) How strong is the relationship between the predictors and the response?

(iii) Which predictors appear to have a statistically significant relationship to the response?

(iv) What does the coefficient for the age variable suggest? How accurate can you estimate the effect of age on price?

(v) What is the predicted price associated for a car with a mileage of 45000 miles, 48 months, diesel,

automatic transmission, 4 doors, 2568 pounds, a displacement of 122 cu. inches, a horsepower of 90, and non-metallic color? What are the associated 95% confidence intervals?

(h) (12 points) Which predictors predictors matter most for predicting the price for a car? (Find the first and the second most important variables)