Homework 1

Big Data Analytics

October 18, 2018

Due time: November 1, 2018, 11:59 PM

Instructions:
1. Please turn in your homework in a pdf format generated by R Markdown.
2. Discussion are encouraged, but you need to work on the problems by yourself.

1 Problem 1 (25 points)

This exercise involves the *Auto* dataset from the text book. Make sure that the missing values have been removed from the data.

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

(b) What is the range of each quantitative predictor? You can answer this using the *range()* function.

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

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full dataset, 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.

(f) 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.

2 Problem 2 (25 points)

This exercise involves the *Boston* dataset from the text book.

(a) How many rows are in the dataset? How many columns? What do the rows and columns represent?

(b) Make some pairwise scatterplots of the predictors in the dataset. Describe your findings.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

(e) How many of the suburbs in this dataset bound the Charles river?

(f) What is the median pupil-teacher ratio among the towns in this dataset?

(g) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(h) In this dataset, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

3 Problem 3 (25 points)

This exercise involves the *Carseats* dataset from the text book.

(a) Fit a multiple regression model to predict *Sales* using *Prices*, *Urban*, and *US*.

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

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

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

(e) 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.

(f) How well do the models in (a) and (e) fit the data?

(g) Using the model from (e), obtain 95% confidence intervals for the coefficients.

(h) Is there evidence of outliers or high leverage observations in the model from (e)? Note: The leverage statistics h_i of a data x_i in simple linear regression is defined as:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

Read the textbook for reference.

4 Problem 4 (25 points)

In this exercise you will create some simulated data and will fit simple linear regression models to it. You can use *set.seed(1)* prior to starting part (a) to ensure consistent results.

(a) Using the *rnorm()* function, create a vector, x, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

(b) Using the *rnorm()* function, create a vector, eps, containing 100 observations drawn from a $N(0, 0.25)$ distribution, i.e., a normal distribution with mean zero and variance 0.25.

(c) Using x and eps, generate a vector y according to the model

$$Y = -1 + 0.5X + \epsilon.$$

What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1.

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the *legend()* command to create an appropriate legend.

(g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model in (c) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b).

(i) Repeat (a)-(f) after modifying the data generation process in such a way that there is more noise in the data. The model in (c) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in (b).

(j) What are the confidence intervals for β_0 and β_1 based on the original dataset, the noisier dataset and the less noisy dataset. Comment on your results.