STAT 435: Homework 1

Due 09/23/2022

This homework should be done independently. All work should be done in Rmarkdown. Please submit electronic copies of both ".Rmd" and ".pdf" files.

Q1. This question involves the use of simple linear regression on the `Auto` data set.

a. Use the `lm()` function to perform a simple linear regression with *mpg* as the response and *horsepower* as the predictor. Use the `summary()` function to print the results. Comment on the output. For example: (i) Is there a relationship between the predictor and the response? (ii) How strong is the relationship between the predictor and the response? (iii) Is the relationship between the predictor and the response positive or negative? (iv) What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

b. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

c. Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Q2. In this exercise you will create some simulated data and will fit a linear regression models to it. Make sure to 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, ϵ, 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 ϵ, 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, β_1 in this linear model?

d. Create a scatterplot displaying the relationship between x and y. Fit a least squares linear model to predict y using x. Display the least squares line on the scatterplot. Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.

e. Then fit a separate quadratic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the quadratic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

f. Answer (e) using a test rather than RSS.

g. Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. You can do this by decreasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

Q3. This problem involves the Boston data set, which we saw in class. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

a. For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

b. Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

c. How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.