3AS/3AS4: Applied Statistics (Autumn Term)

ASSIGNMENT 1 Date: October 25, 2019

To be submitted by 5pm, November 8, 2019

1. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   (c) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

2. Consider the `Auto` data set available in the Course CANVAS page (The filename is `Auto.data`). Download and read the data int `R`. 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) Suppose that we wish to predict gas mileage (`mpg`)on the basis of the other variables. Can you make some plots to suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.

3. To begin, load in the `Boston` data set. The `Boston` data set is part of the `MASS` library in `R`.

   ```
   > library (MASS)
   ```

   Now the data set is contained in the object `Boston`.

   ```
   > Boston
   ```

   Read about the data set:

```
> ?Boston
```

(a) How many rows are there in this data set? How many columns are there? What do the rows and columns represent?

(b) Are any of the predictors associated with per capita crime rate? If so, explain the relationship. Use scatter plots only to justify your answer.

(c) 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.

(d) How many of the suburbs in this data set bound the Charles river?

(e) What is the median pupil-teacher ratio among the towns in this data set?

(f) 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.

(g) In this data set, 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.

4. I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic 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.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic 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.

(d) Answer (c) using test rather than training RSS.