Homework 1

Stat 151A, Spring 2014

Due Thursday, February 6

1. (55 points) Problem 5.2 in the textbook
2. (10 points, continuation of 5.2(a-b)) Write down the equations for $B_{Y|X}$ and $B_{X|Y}$ when Y and X do not have the same mean and standard deviation. Intuition might suggest that $B_{Y|X} = 1/B_{X|Y}$. Is this true?
3. (15 points) Problem 5.3 in the textbook
4. (15 points) Show the statement given in the book (p. 84) that
$$\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0,$$
proving that $TSS = RSS + RegSS$.
5. Suppose $y_1, \ldots, y_n$ are i.i.d random variables with mean μ and variance σ^2. Let $s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$, where $\bar{y} = \frac{1}{n}\sum_i y_i$.
   (a) (10 points) Show that $\sum_i (y_i - \bar{y})$ can be expressed as $\boldsymbol{y}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\boldsymbol{y}$ where $\mathbf{I}_n$ is the identity matrix of order n, and $\mathbf{1}_n$ is a vector of length n consisting of all ones.
   (b) (10 points) Use the matrix form of s^2 to show that $E(s^2) = \sigma^2$ (i.e. do not manipulate the sum directly, but only use the matrix definition)
6. Auto-mpg analysis: The data file `auto-mpg.data.txt` comes from the CMU StatLib site, and includes gas mileage info for a variety of cars from the 1980s, along with other features that may be useful in predicting the gas mileage. Assume that 'mpg' is the response variable of interest. The attributes are (see read me file as well):

1. mpg:	continuous
2. cylinders:	multi-valued discrete
3. displacement:	continuous
4. horsepower:	continuous
5. weight:	continuous
6. acceleration:	continuous
7. model year:	multi-valued discrete
8. origin:	multi-valued discrete
9. car name:	string (unique for each instance)

(a) (75 points) Exploratory Data Analysis (EDA): Create each of the following plots using appropriate variables in the data. Do not include variables for which the plot is not meaningful or descriptive of the data.

Comment on each of the plots: Does anything catch your attention? Are there any outliers or surprises? Give a reason for your answers.

- Histogram / Density Plot
- Boxplot
- Pairs plot with smoothing lines
- Co-plot
- Scatterplot with a smoothing line (lowess) (choose the pair of variables that you find most interesting or most want to look at more closely based on the pairs plot, and explain why you chose them in your comment).

(b) (10 points) Carry out an Ordinary Least Squares analysis relating the gas mileage to the other features. You may use R to do the analysis.

(c) (15 points) Create residual plots. Does anything catch your attention? Are there any outliers? Give a reason for your answers.