Homework 1

Due Monday 9/13

Turn in R code and output – try to use `R markdown`.

Norah Alkhnefr, Madhu Balachandran, Sydney Bornstein and Anlan Chen will lead HW discussion.

This homework uses the wage data available as `Wage.csv` on Blackboard. A full description of the data can be found in the ISLR textbook, pages 1–2. We are interested in explaining/predicting `wage` as a function of 9 possible covariates.

1. (45 points) *Exploratory Data Analysis*: We begin all data analyses with exploratory data analysis (EDA) to understand basic descriptive statistics and relationships in the data.

   For all EDA plots, provide complete labels and legends as appropriate.

   (a) Read the .csv file into `R`. How many columns/variables and how many rows/observations are in the data set?

   (b) Use the `summary` command to obtain descriptive statistics of all variables. Make sure all variables are assigned the correct type (e.g. numeric, factor, etc.). Comment on any interesting or unusual findings.

   (c) Read chapter 5.7.1-5.7.2 and 5.7.4-5.7.5 in *R Book*. Use the `hist` command to describe the variable of interest `wage`. Overlay the empirical distribution using the `density` command. Comment on any interesting or unusual findings.

   (d) Read chapter 5.8.1 in *The R Book*. Use the `pairs` and `cor` (correlation) commands to depict two-way relationships between all continuous variables including `wage`. Comment on any interesting or unusual findings.

   (e) Read chapter 5.6.0 in *The R Book*. Use the `boxplot` command to depict two-way relationships between each factor variable and `wage`. Comment on any interesting or unusual findings.

   (f) Carry out further EDA using descriptive statistics and/or graphical displays for any interesting relationships you discovered.

   (g) Based on your EDA, which variables do you think are strongly related to `wage`? Why?

2. (15 points) *Linear Regression MLEs from Scratch**: Recall the linear regression model in matrix notation is $E(\mathbf{y}) = X^T\boldsymbol{\beta}$ where $\mathbf{y}$ is an n-dimensional vector of responses, X is an $n \times (p+1)$ matrix of p covariates – with a column of ones for the intercept, and $\boldsymbol{\beta}$ is a $(p+1)$-dimensional vector of regression coefficients – including the intercept. The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1} X^T\mathbf{y}$.

   Consider a linear regression of `wage` as a function of the continuous covariates `year` and `age`.

   Calculate $\hat{\boldsymbol{\beta}}$ for the above linear regression model using matrix algebra as described in Exercise 3 of the "Intro R Workshop Slides" from Lecture 1. Clearly indicate your solutions $\hat{\beta}_j$ for $j =$ intercept, year, age.

3. (40 points) *Bootstrap Standard Errors and Confidence Intervals for Linear Regression from Scratch**: The bootstrap is a statistical procedure for empirically determining the distribution of statistical quantities. In this exercise you will use the bootstrap to find the standard error and confidence interval for the regression coefficient $\hat{\beta}_{age}$ in the linear regression of `wage` as a function of the continuous covariates `year` and `age`.

   (a) Use the `sample` command to generate a sample of the observations (sampling with replacement) from the wage data that has the same number of observations as the original data set. This generates one "bootstrap sample." Use the `summary` command to obtain descriptive statistics for the bootstrap sample and briefly comment on how they compare to summary statistics of the original data found in Q1(b). [Note: you will want to `set.seed` so that the random draw generating your bootstrap sample can be reproduced.]

   (b) Use your code from Q2 to estimate the regression coefficient for `age` using the bootstrap sample from part (a). Report the estimate.

   (c) Repeat steps (a) and (b) 1000 times using a for loop to obtain 1000 estimates of $\hat{\beta}_{age}$ from 1000 bootstrap samples. Use the `summary` command to get descriptive statistics for the 1000 bootstrapped $\hat{\beta}_{age}$. Compare the mean of the bootstrap estimates to the $\hat{\beta}_{age}$ estimated from the original sample in Q2.

   (d) Provide a histogram of the 1000 bootstrapped $\hat{\beta}_{age}$. Overlay the empirical distribution as in Q1(c). Comment on the shape of the empirical distribution.

   (e) The bootstrap estimate of the standard error for $\hat{\beta}_{age}$ is the empirical standard deviation of the set of bootstrapped $\hat{\beta}_{age}$. Find the standard error using the `sqrt` and `var` functions.

   (f) The bootstrap estimate of the 95% confidence interval for $\hat{\beta}_{age}$ is defined by the empirical quantiles of the bootstrapped $\hat{\beta}_{age}$. That is, the endpoints are the .025 and .975 percentiles of the set of bootstrapped $\hat{\beta}_{age}$. Find the 95% confidence interval for $\hat{\beta}_{age}$ using the `quantile` function.

   [See chapter 8.2 and 8.12 in *The R Book* for a further description and example of the bootstrap procedure.]

* These HW problems are all about programming common statistical routines using base `R` commands. You may check your work with the relevant `R` packaged commands (e.g. `lm` and `boot`) but your solutions must use base `R` commands such as matrix operations and loops.