Assignment 1 Part 1 (40%)

The Traffic Data which can be downloaded from Blackboard, contain measurements on Volumes and Occupancies from 7 measurement locations in an urban network; these data were used in the 2013 Traffic Forecasting Competition.

- For each measurement location, apply 4-fold (each fold should be correspond to one week of data) cross-validation to examine the predictive power of polynomials of occupancies with orders 2 ,4, 6, 8 on volumes and log-transformed volumes. Estimated models using conventional least squares (use function lm) and robust regression (use rlm). Convert your results for log-transformed volumes to correspond to the original scale. Report Mean Absolute Error, Median Absolute Error and Symmetric Median Absolute Percentage Error $(sMdAPE) = median(200\frac{|Y_t - F_t|}{Y_t + F_t})$ with Y_t observed and F_t predicted volumes in 3 matrices. Each column should correspond to a measurement location, each row to one of the (16) predictive models; report averages of performance metrics across folds.
- Discuss your findings; use a few plots to support your arguments.

Assignment 1 Part 2 (20%)

- For each measurement location, compute bootstrap confidence intervals for the least squares coefficients of the best performing polynomial (according to MAE).
- Save coefficient estimates and confidence intervals in a csv file; each row should correspond to a measurement location.
- Compare bootstrap confidence intervals to the ones derived by adopting the normality assumption and discuss you findings.

Assignment 1 Part 3 (40%)

- Apply logistic regression to predict the probability of default using income and balance on the Default data set (this dataset is included in the ISLR library). Set a random seed before beginning your analysis.
- Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
  - Split the sample set into a training set and a validation set.
  - Fit a multiple logistic regression model using only the training observations.
  - Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
- Repeat the process in above three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.
- Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.
- Obtain bootstrap estimates of the standard error of the 'Pearson' and 'Spearman' correlation between income and balance.

Assignment 1

- E-mail your responses in a **single** pdf file by Thursday, September 17, noon.
- Use the following file name: LASTNAME_FIRSTNAME_ASUID_ASSIGNMENTNUMBER
- Prepare your pdfs carefully; each week some of you will present their work.
- Include a script with the `R` commands you used.