MGS 662: Assignment 2

Ordinary Least Squares: In statistics, ordinary least squares(OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function.

Objective Function: OLS as a Quadratic Program

$\hat{\beta}$=$\frac{argmin}{\beta\epsilon R^p}\sum \quad _i = 1^n(y_i - x_i'\beta)^2$

(Ordinary Least Squares)

$\hat{\beta}$=$\frac{argmin}{\beta\epsilon R^p}$ (-2 β'X'y+ β'X'X β)

(OLS as a Quadratic Program)

Constraints: The constraint matrix is the feature matrix with weight vectors for each feature>= the target feature,

i.e. X $\beta \geq y$

Method to divide data-set into train-test pairs: The code for sampling the data into 80% train set and 20% test set is documented in the R code. We first randomly sample 5000 entries from our blog data set and then divide the randomly sampled dataset into train and test set as documented in the R code using 'sample' function.

Building your model on the train data: Selecting the features: Using Lmfit to fit our model using linear regression. We will use the Lmfit as our adhoc solution to compare the fit obtained through the solution obtained by minimizing the OLS using Rmosek.
Performance of LM fit on train data:

MSE train	**MSE test**
568.5049	460.5903
MSE train (scaled)	**MSE test(scaled)**
543.7526	558.6224

Implemented a solution for our problem in Rmosek:
Optimal model for the train set is the variable 't'

Test the performance of the model on train and test data:

MSE Rmosek train	**MSE Rmosek test**
594.0603	468.9966
MSE Rmosek train (scaled)	**MSE Rmosek test (scaled)**
692.7137	533.5293

Report your observation on how the optimal model performance is different from any adhoc solution obtained above:
The adhoc solution performed better than the Rmosek method. This could be because we used Lmfit as our adhoc solution and contained an intercept term. Four our Rmosek we didn't consider a intercept in our model and hence we got a worse solution in both the cases for scaled and unscaled data.

Implementation and Testing: Report on the performance metric have you used. What motivated the choice of this metric?

- The MSE values is better for unscaled data.
- The MSE value is better for lm fit without scaling when compared to the Rmosek method.

MSE: In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.