MTAT.03.227 Machine Learning  
Spring 2012 Exercise session  
Linear Regression  

February 20, 2012  

1. Analyze dataset [https://raw.githubusercontent.com/johnmylesw...](https://raw.githubusercontent.com/johnmylesw.../master/05-Regression/data/top_1000_sites.tsv) by doing the following steps:  
   - load the data and declare types of features (continuous or discrete), draw scatterplots and check what are the possible relationships between response variable PageViews and other features. (0.5 p)  
   - choose any of the predictors (one) and build simple linear regression in R (Hint: lm function). Interpret it. Draw a scatterplot with the line of your model. Is the model good? Try to improve it (do not add additional features) by using polynomial model and transformations (1 p)  
   - now build a multinomial regression for predicting PageViews. Find the optimal model for the prediction. What is the measure (or measures) you follow to find the best model? Is it the good way to check the performance of the model? Give the thorough analysis of the best model. Show the example of dummy variables and interpret the results in terms of dummy variables. Perform diagnostics. (2 p)  

2. Write a function in R that builds linear regression model on a train set and declares the RSS and R squared on the test set, maps into a plot the data and the predefined number of prediction points. An input of function should be: training and test sets and number of prediction points to plot. (1.5 p)  

3. Write a function in R that computes OLS estimation of $\beta$ for multiple linear regression. Use the above dataset with predictors from your best model to find the $\hat{\beta}$ (1 p)  

4. Show, how to derive the OLS estimation of $\beta_1$ and $\beta_0$ in case of simple linear regression (Hint: solve linear equations) (2 p)  

5. Implement function that performs polynomial regression of pre-defined degree. Do not use lm function in R (perform calculations yourself). The resulting output should consist of coefficients, RSS, and predicted values (2 p)