Spring 2017	**CSC 566:** Advanced Data Mining	Alexander Dekhtyar

Linear Regression

Due date: Wednesday, May 3, in-class

Assignment

The assignment is fairly straightforward.

1. Search UCI Machine Learning Repository, Kaggle, or any other dataset repository for datasets you may like.
2. From the datasets you identify, select **four scenarios** (see below) to conduct linear regression on.

   Scenarios:

   (a) Univariate Linear Regression: one independent numeric variable, one dependent variable.

   (b) Bivariate Linear Regression: two independent numeric variables, one dependent variable.

   (c) Multiple Linear Regression: more than two independent numeric variables, one dependent variable.

   (d) Multiple Linear Regression with categorical attributes: more than two independent variables, one or more of which are categorical, one dependent variable.
3. Perform the appropriate linear regression analysis for each of the scenarios. In order to be acceptable, the results you receive must show an actual non-trivial linear relationship between the independent variable(s) and the dependent variable. You can determine the strength of relationship by looking at the value of the regression coefficients. At least some of them should be reasonably large.

4. Visualize your results for Univariate and Bivariate Linear Regression scenarios. Present your results in a proper, readable form for the other two scenarios.

5. [Optional] If your dataset allows for it, perform backward or forward stepwise regression to discover if there are better combinations of independent variables (for multiple linear regression cases).

Use of Jupyter Notebooks. In general, there is no restriction on the use of programming language/envioronment, etc. However, it may be easiest for you to meet the requirements of the assignment by creating and demonstrating Jupyter Notebooks.

Use of existing tools. You must perform the computations that constitute solving the linear regression equations yourselves. You can use standard linear regression solvers from existing machine learning and statistical packages as the means of testing your code. You can use utility functionality from existing machine learning and statistical packages. You can use numpy for matrix manipulation.

Deliverables

Have your code in a demonstratable form starting May 3, 2017. You will be asked to demo your work in class. You should be able to give a short (5 mins or so) demo describing the independent and dependent variables you used, and showing the resulting regression line/plane and the regression equations.

You may be asked to submit your code via handin at some later time (probably together with your Assignment 2 deliverables).