Home Work 1
(Due Date – 18th Feb, 2014 by Mid-night)

1) Build linear regression model for Anscombe Quartet

Build a linear regress for the four pairs of data in the Anscombe Quartet (you can get the data here https://github.com/hptdss/DS-Intro/blob/master/data/Anscombe_quartet.txt). The data set has x1, y1, x2, y2, x3, y3, x4, y4. Build four models using

a) x1, y1 (predict y1 using x1)

b) x2, y2

c) x3, y3

d) x4, y4

Discuss the results from the above models and how they are different. Do they violate any assumptions of linear modeling? If so, which assumptions are violated for each model.

2) Create a regression model using several independent variables

You can pick your data set, but please either include it with your submission or provide a URL where we can download it. There are several sets built into R/R packages (such as iris or mtcars), but it might be more interesting for you to find a dataset online that is of personal interest to you.

Perform data analysis to understand the relation between the features and the variable you would like to predict. Build a multiple regression model with the best feature set. Check if the residuals follow normal distribution or if they correlate with any features.

Also, create a simplified version of this regression on just one explanatory variable (pick one that had highest predictive significance) but make a polynomial fit out of it. Try adding several polynomial terms to demonstrate that you can "improve" the R^2 of your regression by overfitting.