CS306 Data Analysis and Visualization

Lab. 4 Airline Data (Big) Regression Analysis

1. In this lab practical we will work with Big data of Airline. Download 2 files from the lecture folder i.e. 1.2008.csv.bz2 2. Airline.desc . Unzip the files and you will get .csv file for further experiments.

1. Compute the correlation coefficients by taking two variables from the csv file. Take variable X as Distance and Y as Airtime. Next compute the simple regression line equation is $Y = \beta_0 + \beta_1 X$. Find intercept β_0 and Coefficient (slope) β_1 . Find RMSE between the original y's and predicted $^\wedge y$'s using the derived β_0 and β_1 .
2. Compute 95% confidence for the value of slope and the mean value of y_0 when x_0 is 1200.
3. Using bi-weighted robust least square method to compute more reliable intercept β_0 and slope $\beta_{1,}$ which should be more robust than the previous values. Find RMSE using newly computed parameters. In bi-weighted robust least square each data point is weighted by a weight $w_{i,}$ where $w_i = (1-u_i^2)^2$ when $u_i<=1$ otherwise $w_i=0$. Here $u_i = d_i/3s;$ where s is the interquatile range of d_i and $d_i = (y_i - {}^\wedge y_i)$.

You may use Python/R for this exercise