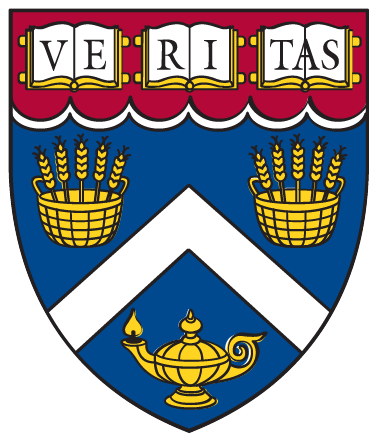
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 1

Due: Monday, 09/23/19 at 7pm EST

**Instructions:** Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Question 1 – 10pts

Question 2 – 10pts

Question 3 – 30pts

Question 4 – 30pts

Question 5 – 20pts

1. For the regression model Yi=β0+εi, derive the least square estimation for β0?
2. The dataset teengamb (see below for the instructions in r) concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

library(faraway) # download the library

data.help("teengamb") # see the description of the data

1. Refer to the CDI data set. The number of active physicians in a CDI (Y) is

expected to be related to total population, number of hospital beds, and total personal income.

* 1. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.
  2. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?
  3. Calculate MSE for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?



1. Repeat question 3, by building the models on the development sample (a random sample of 70% of CDI data), and calculating MSE’s on the hold out sample (remainder 30% of the CDI data).
2. The dataset teengamb concerns a study of teenage gambling in Britain.
   1. Regress the expenditure on gambling (Y) on income (X). State the estimated regression function. Compute the mean and median of the residuals.
   2. Which observation has the largest (positive) residual? Give the case number.