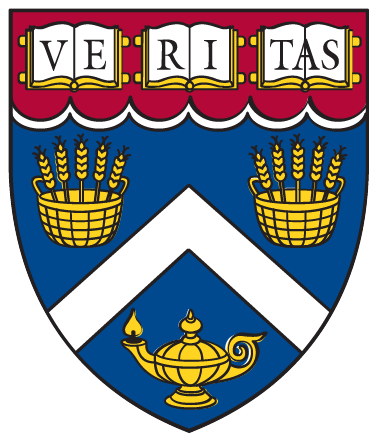
CSCI E-106: Data Modeling



Fall 2020

Dr.Hakan Gogtas

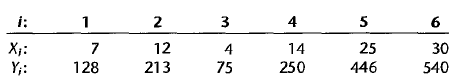
Assignment 1

Due: September 14, 2020 at 7:20 pm 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, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

1. Refer to the Grade point average Data. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). (30 points)
2. Obtain the least squares estimates of β0 and β1, and state the estimated regression function. (5pts)
3. Plot the estimated regression function and the data. "Does the estimated regression function appear to fit the data well? (5pts)
4. Obtain a point estimate of the mean freshman GPA for students with ACT test score X = 30. (5pts)
5. What is the point estimate of the change in the mean response when the entrance test score increases by one point? (5pts)
6. Obtain the residuals . Do they sum to zero? (5pts)
7. Estimate and . In what units is expressed? (5pts)
8. Typographical errors shown below are the number of galleys for a manuscript (X) and the dollar cost of correcting typographical errors (Y) in a random sample of recent orders handled by a firm specializing in technical manuscripts. Assume that the regression model Yi = β1X1 + is appropriate, with normally distributed independent error terms whose variance is a = 16. (20 pts)



1. Evaluate the likelihood function for β1 = 1,2, 3,…,100. For which of β1 values is

the likelihood function largest? (10pts)

1. The maximum likelihood estimator is . Find the maximum likelihood estimate. Are your results in part (a) consistent with this estimate? (10 pts)
2. 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. (30 points)

1. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions. (10 points)
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? (10 points)
3. Calculate MSE for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line? Which variable would you use the estimate Y and why? (10 points)
4. Refer to the CDI data set. Use the number of active physicians as Y and total personal income as X. Select 1,000 random samples of 400 observations, fit the regression model and record β0 and β1 for each selected sample. Calculate the mean and variance of β0 and β1 based on the 1,000 different regression line and compare against the regression model in question 3 part a. (20 points)