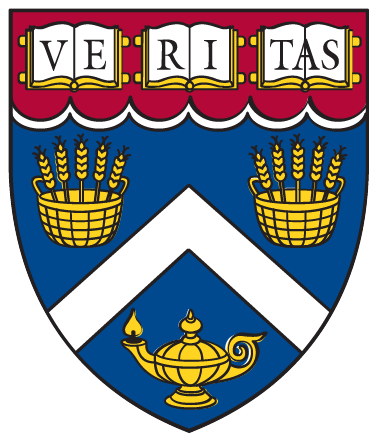
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 2

Due: Monday, 09/30/19 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.

1. The regression model we would like to study is:

and

a-) Write down the likelihood function (5pts)

b-) Find the MLE for and (10pts)

1. Refer to the Grade Point Average (GPA) date set attached below.
2. Obtain the least squares estimates of β0 and β1, and state the estimated regression function. (5pts)
3. Obtain a 99 percent confidence interval for β1.Interpret your confidence interval. (5pts)
4. Test, using the test statistic t\*, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). (5pts)
5. Refer to the Grade Point Average (GPA) date set attached below.
6. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval. (5pts)
7. Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA-using a %95 prediction interval. Interpret your prediction interval. (5pts)
8. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be? (5pts)
9. Calculate %95 percent confidence band for the regression line when Xh = 28. Is your-confidence band wider at this point than the confidence interval in

part (a)? Should it be? (5pts)

1. Repeat question 3, by building the models on the development sample (a random sample of 70% of GPA data), and calculating MSE’s on the hold out sample (remainder 30% of the GPA data).
2. Five observations on Y are to be taken when X = 4, 8, 12, 16, and 20, respectively. The true regression function is E{Y} = 20 + 4X, and the εi are independent N(0, 25).
3. Generate five normal random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at X = 4,8, 12, 16, and 20 and calculate Y1, Y2, Y3, Y4 , and Y5. Obtain the least squares estimates β0 and β1, when fitting a straight line to the five cases. Also calculate when Xh = 10 and obtain a %95 confidence interval for

E{Yh} when Xh = 10. (10 pts)

1. Repeat part (a) 200 times, generating new random numbers each time. (15 pts)
2. Make a frequency distribution of the 200 estimates β1. Calculate the mean and standard deviation of the 200 estimates β1. Are the results consistent with theoretical expectations? (10 pts)
3. What proportion of the 200 confidence intervals for E{Yh} when Xh = 10 include E{Yh}? Is this result consistent with theoretical expectations? (10 pts)

