CSCI E-106:Assignment 3

Due Date: September 28, 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.

Problem 1

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 ϵ_i are independent N(O, 25). (40 points, 10 points each)

a-) 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 Y_1, Y_2, Y_3, Y_4 , and Y_5. Obtain the least squares estimates b_0 and b_1, when fitting a straight line to the five cases. Also calculate Y_h when $X_h = 10$ and obtain a 95 percent confidence interval for E(Y_h) when $X_h = 10$.

b-) Repeat part (a) 200 times, generating new random numbers each time.

c-) Make a frequency distribution of the 200 estimates b_1. Calculate the mean and standard deviation of the 200 estimates b_1. Are the results consistent with theoretical expectations?

d-) What proportion of the 200 confidence intervals for E(Y_h) when $X_h = 10$ include E(Y_h)? Is this result consistent with theoretical expectations?

Problem 2

Refer to the CDI data set (used in homework 1). 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. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians? (20 Point)

Problem 3

Refer to the CDI data set (use in previous homework). For each geographic region, regress per capita income in a CDI (Y) against the percentage of individuals in a county having at least a bachelor's degree (X). Obtain a separate interval estimate of β_1, for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes? (20 points)

Problem 4

In a small-scale regression study, five observatiol)s on Y were obtained corresponding to X = 1, 4,10, ll, and 14. Assume that $\sigma = .6$, $\beta_0 = 5$, and $\beta_1, = 3$. (20 points, 10 points each)

a-) What are the expected values MSR and MSE?

b-) For determining whether or not a regression relation exists, would it have been better or worse to have made the five observations at X = 6,7, 8, 9, and 1O? Why? Would the same answer apply if the principal purpose were to estimate the mean response for X = 8? Discuss.