STA302/1001H1S - Method of Data Analysis

Assignment 2

Due March 28th, 23:59

Instructions: This is individual assignment. It is worth 100 points. Please use Rmarkdown to write your solutions and submit your solutions with relevant R code included as a pdf file via **Crowdmark**.

Question

Census data was collected on the 50 states and Washington, D.C. We are interested in determining whether average lifespan (LIFE) is related to the ratio of males to females in percent (MALE), birth rate per 1,000 people (BIRTH), divorce rate per 1,000 people (DIVO), number of hospital beds per 100,000 people (BEDS), percentage of population 25 years or older having completed 16 years of school (EDUC) and per capita income (INCO). The data stored in the data file Census.txt can be found on the course webside.

Answer the following questions.

Part 1 (20 Marks): In this Part, compute by hand using matrix formulas. DO NOT USE lm() command in this part.

We consider a multiple linear regression model with LIFE (y) as the response variable, and MALE (x_1), BIRTH (x_2), DIVO (x_3), BEDS (x_4), EDUC (x_5), and INCO (x_6), as predictors. Answer the following questions using least square estimates in term of matrix formulas.

(a) Compute and report the least-squares estimates. Write down the least-squares regression equation.

(b) Explain in context what the coefficients corresponding to MALE and BIRTH mean.

(c) Compute the biased and the unbiased estimates of the error variance σ^2.

(d) Using the unbiased estimate of error variance, Compute the standard errors of the estimators of the regression coefficients.

(e) Compute the coefficient of determination. Give a practical interpretation of your result.

Part 2 (60 Marks): In this part, you may use all R commands you need, including lm() function, to answer the following questions.

(a) Fit the MLR model with LIFE (y) as the response variable, and MALE (x_1), BIRTH (x_2), DIVO (x_3), BEDS (x_4), EDUC (x_5), and INCO (x_6), as predictors.

(b) At level $\alpha = 5\%$, conduct the F-test for the overall fit of the regression. Comment on the results.

(c) At level $\alpha = 1\%$, test each of the individual regression coefficients. Do the results indicate that any of the explanatory variables should be removed from the model?

(d) Determine the regression model with the explanatory variable(s) identified in part (c) removed. Write down the estimated regression equation.

(e) Perform a partial F-test at level $\alpha = 1\%$ to determine whether the variables associated with MALE and INCO can be removed from the model

(f) Compute and report the F test statistic for comparing the two models

$$\mathbf{E}(Y_i|x_i) = \beta_0 + \beta_1 x_{i1},$$

$$\mathbf{E}(Y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6},$$

(g) Perform a partial F-test at level $\alpha = 1\%$ for comparing the two models

$$\mathbf{E}(Y_i|x_i) = \beta_0,$$

$$\mathbf{E}(Y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

(h) Compute and report the terms in the decomposition

$$\text{SSreg}(\beta_1, \beta_2, \beta_3|\beta_0) = \text{SSreg}(\beta_3|\beta_0) + \text{SSreg}(\beta_2|\beta_0, \beta_3) + \text{SSreg}(\beta_1|\beta_0, \beta_3, \beta_2)$$

(i) Suppose we are interested in fitting a regression model using LIFE as the response variable and some subset of the variables (MALE, BIRTH, DIVO, and INCO) as predictor.

(i.1) Perform variable selection by finding the subset model that minimizes the AIC criteria. State the 'best model'.

(i.2) Perform variable selection using forward selection. State the 'best model'.

(i.3) Perform variable selection using backward selection. State the 'best model'.

Part 3 (20 Marks): In this part, you may use all R commands you need.

We consider the multiple linear regression with LIFE (y) as the response variable, and MALE, BIRTH, DIVO , BEDS, EDUC, and INCO, as predictors.

(a) Plot the standardized residuals against the fitted values. Are there any notable points. In particular look for points with large residuals or that may be influential.

(b) Compute and plot the leverage of each point. Identify any points that have a leverage larger than 0.5.

(c) Compute the Cook's distance for each point. Identify any points that have a Cook's distance larger than 1. Are these the same observations as those seen in part (b)?

(d) Plot the standardized residuals against the variable BEDS. Specifically mark the point corresponding to Washington, D.C. What can you say about this observation?

(e) Remove the observation corresponding to Washington, D.C. and refit the model. Are there any notable differences with the model fit in part (a)?

(f) Plot the standardized residuals against each of the 6 explanatory variables. Specifically mark the observation corresponding to UT. What is notable about this state?

(g) Remove the observation corresponding to UT and refit the model. Are there any notable differences with the model fit in part (a)? In particular, how does UT's exclusion impact the R^2 value?