STA 6923 Homework 4, Fall 2021

Due date: Tuesday, October 5, 11:59 pm

Instruction: Answer your questions item by item. Write your conclusions clearly and concisely. You need to provide two files as follows.

- Your .Rmd code file, with comments on your code
- Your full homework report, written by R markdown pdf (or html) file.

Not from the text book

A data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are given in Table 1. An R data (CDI_data.rda) is included. You can put it into your homework file directory and use **readRDS("CDI_data.rda")** to load the data.

Do the following problems. Note that each question may ask different regression model.

1. The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4, X_5) that will be defined next.
   (a) Fit a first-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise. Comments on your results.
   (b) Use a level of significance $\alpha = 0.10$ to test whether any geographic effects are present. State the alternatives, decision rule, and conclusion. What is the p-value of the test?
2. A public safety official wishes to predict the rate of serious crimes (Y, total number of serious crimes per 100,000 population, i.e., you need to recalculate the response here). The pool of potential predictor variables includes all other variables in the data set except total population, total serious crimes, county, state, and region. It is believed that a model with predictor variables in first-order terms with no interaction terms will be appropriate. Consider the even-numbered cases to constitute the model-building data set to be used for the following analyses.
   (a) Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. Is there evidence of strong linear pairwise associations among the predictor variables here?
   (b) Use several model selection techniques to pick up models and comment on them. If there are multiple candidates, pick one you deem attractive.
   (c) Now use the odd-numbered cases as test (validate) data. Fit the regression model from part (b) above.
      i. Compare the estimated regression coefficients and their estimated standard deviations with the obtained in part (b). In addition, compare the error mean squares and coefficients of multiple determination. Does the model fitted to the test data set yield similar estimates as the model fitted to the model-building data set?

ii. Use the model-building data fitting to calculate the MSE of the predictions to the test data and compare it to MSE of the predictions to training (model-building) data set. Is there evidence of a substantial bias problem in MSE here?

iii. Fit the selected regression model to the combined training and test data sets. Are the estimated regression coefficients and their estimated standard deviations appreciably different from those for the model fitted to the training data set? Should you expect any differences in the estimates? Explain.

3. Now consider the regression model, with total crime (per 100,000) as response and the variables 6, 8, 9, 13, 14, and 15 as predictors, in first-order terms is to be evaluated in detail based on the training data set (even-numbered cases).

   (a) Obtain the residuals and plot them separately against $\hat{Y}$, each predictor variable in the model,. On the basis of these plots, should any modifications in the model be made?

   (b) Do a normal probability plot of the residuals and test for normality at a significance level of 0.10.

   (c) Obtain the scatter plot matrix, the correlation matrix of the X variables, and the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

   (d) Obtain the studentized deleted residuals and prepare a dot plot of these residuals. Are any outliers present? Use the Bonferroni outlier test (in R, use outlierTest in the *car* package) procedure with $\alpha = 0.05$. State the decision rule and conclusion.

   (e) Obtain the diagonal elements of the hat matrix. Using the rule of thumbs described in the notes, identify any outlying X observations.

   (f) Cases 2, 8, 48, 128, 206, and 404 are outlying with respect to their X values, and Cases 2 and 6 are reasonably far outlying with respect to their Y values. Obtain $DFFITS$, $DFBETAS$, and Cook's distance values for these cases to assess their influence. What do you conclude?

Table 1: Variable names for CDI data

Variable Number	Variable Name	Description
1	Identification number	1-440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18-34	Percent of 1990 CDI population aged 18-34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active non-federal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: $1 = NE$, $2 = NC$, $3 = S$, $4 = W$