STA 4853
Homework #1
Due on 2/5/2024 (Monday)
(submit by uploading pdf to Canvas)

Before doing the homework, you should read `06_homework_guidelines.pdf` which is posted in mordor. This describes the homework policies for this course and various rules for how your homework should be written and formatted.

Don't forget to put all your code at the end of your assignment!

The file `hw1dat.txt` contains simulated data with 103 observations and four variables, a response variable Y and three covariates X1, X2, X3, given in this order. (That is, the data consists of 103 rows and 4 columns containing the variables Y, X1, X2, X3.) Do the following.

1. Regress Y on X1, X2, X3. Include the table of Parameter Estimates with your homework.

2. (a) Using the data and the parameter estimates in the output, compute (using a calculator) the predicted values and residuals for the first two observations. Show your work. (That is, for the first two observations, write the formula for the predicted values with all the variables replaced by their numerical values.)

   (b) Compute (using a calculator) approximate 95% confidence intervals for β_1 and β_2, the regression coefficients for X1 and X2. Show your work.

   (c) Assuming the usual regression assumptions are at least approximately true, give a value A such that you would expect about 68% of the residuals to be between $-A$ and $+A$.

3. Three unusual observations (i.e., rows) have been planted in this data set. Two of these observations are easily visible in the the pairwise scatterplots of Y, X1, X2, X3 (there are 6 such plots). Create these plots and locate these two observations in those plots in which they are easily visible (i.e., they stick out from the other points). Determine which observations these are. Circle the observations and write their observation numbers next to the circled points.

   [Note: the scatterplots may be created one by one using PROC SGPLOT as illustrated in class, or a matrix of scatterplots may be created using PROC CORR or PROC SGSCATTER.]

   [Hint: It is easier to first do problem 4, and then, after determining which are the three unusual observations, go back and do problem 3.]

4. (a) Now examine the case diagnostic plots involving RStudent (the Studentized residuals), Leverage, and Cook's D, and locate all three of the unusual observations in these plots; circle them in all the plots where they are easily visible, and (after doing part (b) below) write their observation numbers next to the circled points.

   (b) Create a data set (named stuff, or whatever you want) containing the values of these diagnostics. Print this data set, and use this along with the diagnostic plots to determine which observations these unusual points are. Underline or mark the three unusual observations in the printed data and then go back to the plots in 4(a) and write the observation numbers next to the points you circled.

5. Use only the case diagnostics to answer the following questions.

   (a) Two of the three points have unusual covariate values. Which are they and how do you know?

   (b) Two of the points have unusual response values. Which are they and how do you know?

   (c) One of the points has a much greater effect on the regression model (such as on the the estimated parameters and predicted values) than the other two. Which is it and how do you know?

6. Delete the three unusual observations from the data (by deleting the three corresponding rows from the data file) and then run the regression of Y on X1, X2, X3 again. Include the table of Parameter Estimates with your homework. Examine the plot of the residuals versus predicted values, the plot of the residuals versus the normal quantiles, and the plots of the residuals versus the covariates X1, X2, X3. (Include these plots with your assignment.) Do the regression assumptions appear to be satisfied? Support your answer by comments about the appearance of the plots.