STAT 425 — Fall 2018

Homework 4

(Posted on Fri. Oct 12; Due Thu., Oct 25)

Please submit your assignment *on paper*, following the Guidelines for Homework posted at the course website. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. Problem 4 from Homework 3.

2. Using the `seatpos` data, fit the linear regression of `hipcenter` on all of the other variables.

   (a) Produce a summary of the regression results.

   (b) Do any variables appear to be significant based on the individual t-tests for their coefficients? What about based on the overall F-test (for all of the variables together)?

   (c) Compute the variance inflation factors (VIFs) for the variables. Using the threshold of 10 to determine if a VIF indicates a problem of collinearity, which variables have a VIF indicating a possible problem?

   (d) Reduce the model by removing all variables that had VIFs you identified as problematic in the previous part. Produce a summary of the regression results.

   (e) For the model of the previous part, do any variables appear to be significant based on the individual t-tests for their coefficients? What about based on the overall F-test (for all of the variables together)?

   (f) Compute the VIFs for the reduced set of variables. (Have they changed?) Again using the threshold of 10, which variables have a VIF indicating a possible problem?

3. Recall that $H = P_X = X(X^TX)^{-1}X^T$. Assume that X contains the column of 1's corresponding to the intercept. Show the following:

   (a) $H\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the $n \times 1$ vector of 1's. Therefore, show $\sum_{j=1}^{n} h_{ij} = 1$.

   (b) For simple linear regression where X is an $n \times 2$ matrix, show the identity for h_{ii} displayed on page 13 of the diagnostics lecture notes

4. We would like to evaluate the performance of regression diagnostics using synthetic data that we ourselves create. Follow the below steps to generate a synthetic dataset with $n = 100; p = 5$.

   (a) Generate n draws from a multivariate normal distribution with mean 0 and covariance matrix

   $$\Sigma_{5\times 5} = 0.8I + 0.2J,$$

   where I is the identity matrix and J is the matrix of all 1's. Define $Z_{n\times p}$ to be the matrix with these n draws as its rows. Add a column on 1's to the matrix Z to obtain $X_{n\times(p+1)}$.

(b) Generate 90 random errors from $N(0,1)$ and the remaining 10 errors from a t_2 distribution. Randomly shuffle these 100 random errors to define ϵ and obtain $Y = X\beta + \epsilon$, where

$$\beta_{6\times 1} = (0, 1, -1, 1, -1, 0).$$

(c) Now perform a multiple linear regression model of Y on X. Provide your conclusions based on the R output.

(d) Perform all the relevant regression diagnostics to see if you detect any problems.

Repeat steps (a) - (d) another time to see if your answers are consistent for a new dataset.

Some reminders:

- Unless otherwise stated, all data sets are from the `faraway` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.