STAT 425 — Fall 2018

Homework 4

(Posted on Tue. October 30; Due Tue, November 13)

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. Using the `sat` data, fit a model with total as the response and takers, ratio, and salary as predictors using the following methods:

   (a) Ordinary least squares
   (b) Huber's robust regression

   Compare the results. In each case, comment on the significance of the predictors.

2. We would like to compare Least squares and generalized least squares. Follow the below steps to generate a synthetic dataset with $n = 100; p = 5$ and

$$\beta_{5\times 1} = (1, -1, 1, -1, 1).$$

   (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 $X_{5\times 5}$ to be the matrix with these n draws as its rows.

   (b) Generate n random errors $\epsilon_{n\times 1}$ from $N(0, A)$, where A is a diagonal matrix with diagonal elements $(\sqrt{1}, \sqrt{2}, \cdots, \sqrt{n})$. Define $Y = 1 + X\beta + \epsilon$.

   (c) Now obtain the Least squares, and the generalized least squares estimators. Compute the Average Mean Squared Errors (MSE) for each estimator as

$$MSE(\hat{\beta}) = \sum_{i=1}^{p} (\hat{\beta}_i - \beta_i)^2.$$

   Comment on the performance of the different methods.

   Repeat steps (a) - (c) to obtain 10 synthetic datasets and average the values of the MSEs.

   Now repeat the above procedure if the errors $\epsilon_{n\times 1}$ are from $N(0, B)$, with $B = A + 0.1J$. Provide your conclusive remarks about the performance of least squares and generalized least squares.

3. We investigate a simple variable selection problem via the mean square error (MSE). In general if we estimate a parameter θ using a statistic T, then the mean square error of T is given by $MSE(T) = E[(T - \theta)^2] = \{E(T) - \theta\}^2 + var(T)$. Consider the following two regression models:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, \ldots, n; \tag{1}$$

and

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2}$$

We assume that model (1) is the correct one: it holds with uncorrelated errors having mean zero and constant variance σ^2. Model (2) is considered to be oversimplified because it leaves out the variable X_2. Denote the LS estimators for the two models as follows:

$$\begin{aligned} \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 : &\quad \text{Least squares estimates for Model (1)} \\ \tilde{\beta}_0, \tilde{\beta}_1 : &\quad \text{Least squares estimates for Model (2).} \end{aligned}$$

In addition, assume that $\bar{X}_1 = 0$, $\bar{X}_2 = 0$. Also, the following notation will be useful:

$$SX_1X_1 = \sum_{i=1}^{n} x_{1i}^2, \quad SX_2X_2 = \sum_{i=1}^{n} x_{2i}^2,$$

$$SX_1X_2 = \sum_{i=1}^{n} x_{1i}x_{2i} \quad \text{and} \quad r_{12} = \frac{SX_1X_2}{\sqrt{SX_1X_1}\sqrt{SX_2X_2}}.$$

We will compare the mean square errors of $\hat{\beta}_1$ and $\tilde{\beta}_1$ as estimators of β_1 in Model (1).

(a) Show that

$$MSE(\hat{\beta}_1) = \frac{1}{1 - r_{12}^2} \frac{\sigma^2}{SX_1X_1}.$$

(b) Assuming Model (1) is correct show that

$$E(\tilde{\beta}_1) = \beta_1 + \frac{SX_1X_2}{SX_1X_1}\beta_2 = \beta_1 + r_{12}\sqrt{\frac{SX_2X_2}{SX_1X_1}}\beta_2.$$

(c) Assuming Model (1) is correct, show that

$$var(\tilde{\beta}_1) = \frac{\sigma^2}{SX_1X_1}.$$

(d) Assuming Model (1) is correct show that

$$MSE(\tilde{\beta}_1) < MSE(\hat{\beta}_1)$$

whenever

$$\beta_2^2 < var(\hat{\beta}_2) = \frac{1}{1 - r_{12}^2} \frac{\sigma^2}{SX_2X_2}.$$

4. Using the `happy` data set, fit a model with `happy` as the response and all of the other four variables as predictors.

(a) Use transformations to find a good model for predicting happy.

The `cornnit` data for understanding the relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application were studied in Wisconsin in 1994.

(b) Use transformations to find a good model for predicting yield from nitrogen.

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.