Spring 2020

Quantitative Analysis
Chung-Ming Kuan

Problem Set 7: Solution

Part One: Hand-Written Exercise

1. For a data set $(y_i, \mathbf{x}_i)_{i=1}^n$, where y_i is a scalar and $\mathbf{x}_i$ a $p \times 1$ column vector. That is, the regression model is $y_i = \sum_{j=1}^p \beta_j x_{ij} + u_i$. Please show that the Ridge Regression estimator $\hat{\beta}_R$ is given by:

$$\hat{\beta}_R = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' + \lambda \mathbf{I} \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i y_i \right),$$

where λ is the tuning parameter and $\mathbf{I}$ the p-dimensional identity matrix.

Let

$$Q := \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2$$

The F.O.Cs. are

$$\begin{cases} \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_1 x_{i1} - ... - \beta_p x_{ip}) x_{i1} + 2\lambda \beta_1 = 0 \\ \frac{\partial Q}{\partial \beta_2} = -2 \sum_{i=1}^n (y_i - \beta_1 x_{i1} - ... - \beta_p x_{ip}) x_{i2} + 2\lambda \beta_2 = 0 \\ \vdots \\ \frac{\partial Q}{\partial \beta_p} = -2 \sum_{i=1}^n (y_i - \beta_1 x_{i1} - ... - \beta_p x_{ip}) x_{ip} + 2\lambda \beta_p = 0 \end{cases}$$

$$\Rightarrow \begin{cases} (\sum_{i=1}^n x_{i1}^2 + \lambda)\beta_1 + \sum_{i=1}^n x_{i1}x_{i2}\beta_2 + \ldots + \sum_{i=1}^n x_{i1}x_{ip}\beta_p = \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}x_{i1}\beta_1 + (\sum_{i=1}^n x_{i2}^2 + \lambda)\beta_2 + \ldots + \sum_{i=1}^n x_{i2}x_{ip}\beta_p = \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}x_{i1}\beta_1 + \sum_{i=1}^n x_{ip}x_{i2}\beta_2 + \ldots + (\sum_{i=1}^n x_{ip}^2 + \lambda)\beta_p = \sum_{i=1}^n x_{ip}y_i \end{cases}$$

$$\Rightarrow \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 + \lambda & \sum_{i=1}^n x_{i1}x_{i2} & \ldots & \sum_{i=1}^n x_{i1}x_{ip} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 + \lambda & \ldots & \sum_{i=1}^n x_{i2}x_{ip} \\ \vdots & \vdots & \ldots & \vdots \\ \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \ldots & \sum_{i=1}^n x_{ip}^2 + \lambda \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{bmatrix}$$

$$\Rightarrow (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = (\sum_{i=1}^{n} \mathbf{x_i}\mathbf{x_i}' + \lambda\mathbf{I})^{-1}(\sum_{i=1}^{n} \mathbf{x_i} y_i)$$

2. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p+1$ models, containing 0, 1, 2, ..., p predictors. For $k = 1, .., p$, please answer the following questions and justify your answers:

   (a) Which of the three models with k predictors has the smallest training RSS?
   Best subset. Because best subset approach is to find the smallest training RSS model among all combinations given k predictors.

   (b) Which of the three models with k predictors has the smallest testing MSE?
   Not sure. It depends on the data testing the model.

   (c) (True or False) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.
   True. For forward stepwise selection, once a predictor is included in the model, it will never be kicked out for the next step or thereafter.

   (d) (True or False) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.
   True. Because it is to remove a predictor in the (k+1)-predictors model to become the k-predictors model.

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\min_{\beta} \text{RSS} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

   for some $s \in \mathbb{R}$. Please answer the following questions and justify your answers:

   (a) As s increases from 0 to ∞, what will happen to the training RSS?
   The constraint decreases when s goes up. Hence the training RSS will decreases.

   (b) As s increases from 0 to ∞, what will happen to the testing MSE?
   The testing MSE forms a U-shape when s goes up, which means it will first decrease amd then increase.

   (c) As s increases from 0 to ∞, what will happen to the variance of our estimated coefficients?

When s increases, the constraint of the range of coefficients value will decreases, which means more disperse and larger variance.

(d) As s increases from 0 to ∞, what will happen to the bias of our estimated coefficients?
Without constraint, the coefficients are unbiased. So when s increases, the bias of the coefficients will decrease.