Problem Set 4

Due Oct 31, before class

1. Equivalence of F and $\bar{R}^2$

Consider testing $\beta_2 = 0$ for the two nested Normal linear models:

$$
\begin{aligned}
Y &= 1_n\beta_0 + X_1\beta_1 + \epsilon \\
Y &= 1_n\beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon
\end{aligned}
$$

We can use the standard F statistic. We can also compare the adjusted R^2's from these two models: $\bar{R}_1^2$ and $\bar{R}_2^2$. Show that

$$F > 1 \Longleftrightarrow \bar{R}_1^2 < \bar{R}_2^2.$$

2. Best subset selection in *lalonde* data

Produce the figure similar to the best subset selection example in class using the *lalonde* data in the *Matching* package. Report the AIC, BIC, PRESS, and GCV of the selected model.

3. Derivative of the MSE of the ridge regression

Show that

$$\left.\frac{\partial \text{MSE}(\lambda)}{\partial \lambda}\right|_{\lambda=0} < 0.$$

Remark: This result ensures that the the ridge estimator must have smaller MSE than OLS in a neighborhood of $\lambda = 0$.

4. Ridge as OLS with augmented data

Show that $\hat{\beta}^{\text{ridge}}(\lambda)$ equals the OLS coefficient of $\tilde{Y}$ on $\tilde{X}$ with augmented data

$$\tilde{Y} = \begin{pmatrix} Y \\ 0_p \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix},$$

where $\tilde{Y}$ is an $n+p$ dimensional vector and $\tilde{X}$ is an $(n+p) \times p$ matrix.

5. Leave-one-out formulas for ridge

Define $\hat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y$ as the ridge coefficient (dropping the superscript "ridge" for simplicity), $\hat{\epsilon}(\lambda) = Y - X\hat{\beta}(\lambda)$ as the residual vector using the full data, and $h_{ii}(\lambda) = x_i (X^T X + \lambda I_p)^{-1} x_i^T$ as the (i,i)-th diagonal element of $H(\lambda) = X(X^T X + \lambda I_p)^{-1} X^T$. Define $\hat{\beta}_{[-i]}(\lambda)$ as the ridge coefficient without observation i, and $\hat{\epsilon}_{[-i]}(\lambda) = y_i - x_i^T \hat{\beta}_{[-i]}(\lambda)$ as the predicted residual. Prove the following leave-one-out formulas for ridge regression

$$\begin{aligned} \hat{\beta}_{[-i]}(\lambda) &= \hat{\beta}(\lambda) - \{1 - h_{ii}(\lambda)\}^{-1} (X^T X + \lambda I_p)^{-1} x_i \hat{\epsilon}_i(\lambda), \\ \hat{\epsilon}_{[-i]}(\lambda) &= \frac{\hat{\epsilon}_i(\lambda)}{1 - h_{ii}(\lambda)}. \end{aligned}$$

Hint: You can use the result in the last problem and apply the leave-one-out formulas for OLS.

6. An equivalent form of ridge coefficient

Show that the ridge coefficient has two equivalent forms: for $\lambda > 0$

$$(X^T X + \lambda I_p)^{-1} X^T Y = X^T (X X^T + \lambda I_n)^{-1} Y.$$

Hint: Use the Woodbury formula $(A+BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1}+CA^{-1}B)^{-1}CA^{-1}$ by setting $A = \lambda I_p$, $B = X^T$, $D = I_n$, and $C = X$.

Remark: This formula is more useful when $p > n$; If $p > n$ and XX^T is invertible, then we can let λ go to zero on the right-hand side, yielding

$$\hat{\beta}^{\text{ridge}}(0) \;=\; X^T(XX^T)^{-1}Y.$$

7. Penalized OLS with an orthogonal design matrix

Consider the special case with standardized and orthogonal design matrix:

$$X^T 1_n \;=\; 0, \quad X^T X \;=\; I_p.$$

For the fixed $\lambda \geq 0$, find the explicit formulas of the j-th coordinate of the following estimators in terms of the corresponding j-th coordinate of the OLS estimator $\hat{\beta}_j$ and λ ($j = 1, \ldots, p$):

$$\begin{aligned}
\hat{\beta}^{\text{ridge}}(\lambda) &= \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ ||Y - Xb||_2^2 + \lambda ||b||_2^2 \right\}, \\
\hat{\beta}^{\text{lasso}}(\lambda) &= \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ ||Y - Xb||_2^2 + \lambda ||b||_1 \right\}, \\
\hat{\beta}^{\text{enet}}(\lambda) &= \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ ||Y - Xb||_2^2 + \lambda(\alpha ||b||_2^2 + (1-\alpha)||b||_1 \right\}, \\
\hat{\beta}^{\text{subset}}(\lambda) &= \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ ||Y - Xb||_2^2 + \lambda ||b||_0 \right\},
\end{aligned}$$

where

$$||b||_2^2 = ||b||^2 = \sum_{j=1}^p b_j^2, \quad ||b||_1 = \sum_{j=1}^p |b_j|, \quad ||b||_0 = \sum_{j=1}^p I(b_j \neq 0)$$

with $I(\cdot)$ being the indicator function.

8. Coordinate descent for the elastic net

Give the detailed coordinate descent algorithm for the elastic net.

9. More noise in the Boston housing data

The Boston housing data have $n = 506$ observations. Add $p = n$ columns of covariates of random noise and randomly split the data into the training set and testing set with ratio $8 : 2$. Compare ridge and lasso in terms of the estimated coefficients and the mean squared predicted error in the testing dataset.