Homework 9

Due Friday, 11/13/20

1. Consider the data "Steam.txt" where you are trying to predict steam usage using fat, glycerine, wind, freezday and temp. Perform forward and backward selection with $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$, respectively. Which covariates did you include in the full model? How does this compare when you perform best subset regression using AIC to rank models? Does the model chosen with best subset regression change when you use BIC to rank models? Example R code can be found on Canvas to help you get started.

2. (Training vs. test error) Suppose $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ for some unknown function f, where $\boldsymbol{x}_i$ is non-random, $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}\left(\epsilon_i^2\right) = \sigma^2$ for $i = 1, \dots, n$. Suppose you use the training set $\mathcal{T} = \{y_1, \dots, y_n\}$ to obtain $\hat{f}$, an estimate for f. Define the in-sample **test error** and **training error** to be

$$\text{Err}_{\text{in}} = \mathbb{E}\left[n^{-1}\sum_{i=1}^{n}\left\{\tilde{y}_i - \hat{f}(\boldsymbol{x}_i)\right\}^2 \mid \mathcal{T}\right], \quad \overline{\text{err}} = n^{-1}\sum_{i=1}^{n}\left\{y_i - \hat{f}(\boldsymbol{x}_i)\right\}^2$$

where $\tilde{y}_i$ is an independent copy of y_i (i.e. $\tilde{y}_i$ and y_i are independent and have the same distribution). Err_{in} determines how the estimator performs in the prediction of new unobserved data. Note that Err_{in} and $\overline{\text{err}}$ are random quantities, where the randomness is due to the randomness in the training data $\mathcal{T}$.

   (a) Show that

$$\mathbb{E}(\text{Err}_{\text{in}}) = n^{-1}\sum_{i=1}^{n}\left(\left[\text{Bias}\left\{\hat{f}(\boldsymbol{x}_i)\right\}\right]^2 + \text{Var}\left\{\hat{f}(\boldsymbol{x}_i)\right\}\right) + \sigma^2,$$

   where the expectation is taken over the training set $\mathcal{T}$. Why does this imply that minimizing $\mathbb{E}(\text{Err}_{\text{in}})$ with respect to the choice of estimator $\hat{f}$ is equivalent to minimizing the average MSE $n^{-1}\sum_{i=1}^{n} MSE_i$, where $MSE_i = \mathbb{E}[\{\hat{f}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\}^2]$?

   (b) Show that the expected **optimism** in the training error, $\omega = \text{Err}_{\text{in}} - \overline{\text{err}}$, is

$$\mathbb{E}(\omega) = \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\left\{y_i, \hat{f}(\boldsymbol{x}_i)\right\},$$

   where the expectation is taken over the training set $\mathcal{T}$. Why does this show that the training error usually **underestimates** the test (i.e. prediction) error?

3. (Ridge regression) Suppose $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^T \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ and $\lambda \geq 0$.

   (a) Show that $\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p$ is invertible for all $\lambda > 0$, regardless of whether or not $\boldsymbol{X}$ is full rank.

(b) Suppose $\text{Cov}\left(y_i, y_j\right) = \sigma^2 1\{i = j\}$ and let $\hat{\boldsymbol{Y}}_\lambda^{(ridge)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(ridge)}(\lambda)$, where $\hat{\boldsymbol{\beta}}^{(ridge)}(\lambda)$ is ridge regression's estimate for $\boldsymbol{\beta}$ using penalty parameter λ.

(i) Find an $n \times n$ matrix $\boldsymbol{H}_\lambda$ that only depends on $\boldsymbol{X}$ and λ such that $\hat{\boldsymbol{Y}}_\lambda^{(ridge)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(ridge)}(\lambda) = \boldsymbol{H}_\lambda \boldsymbol{Y}$. Let df_λ be the number of degrees of freedom (see midterm) in the ridge regression estimator with pentalty parameter λ. Conclude that $df_\lambda = \text{Tr}\left(\boldsymbol{H}_\lambda\right)$.

(ii) Show that for $0 < \lambda_1 < \lambda_2$, $df_{\lambda_2} < df_{\lambda_1} < \text{rank}(\boldsymbol{X})$.

(c) One way to choose λ is with leave-one-out cross validation (i.e. PRESS):

$$PRESS(\lambda) = \sum_{i=1}^{n} \left\{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(-i)}^{(ridge)}(\lambda)\right\}^2.$$

Show that

$$PRESS(\lambda) = \sum_{i=1}^{n} \left\{\frac{y_i - \hat{y}_i^{(\lambda)}}{1 - h_{ii}^{(\lambda)}}\right\}^2,$$

where $\hat{y}_i^{(\lambda)}$ is the ith element of $\hat{\boldsymbol{Y}}_\lambda^{(ridge)}$ and $h_{ii}^{(\lambda)}$ is the ith diagonal element of $\boldsymbol{H}_\lambda$. (**Hint**: your proof should be similar to that from HW8.)

(d) Using part (c), define the corresponding generalized cross validation estimator for λ.

4. Consider the data Fat.txt. Remove every tenth observation from the data for use as a test sample. Use the remaining data to fit (i.e. train) the following models where % body fat, `siri`, is the response and all other variables are predictors:

(i) A simple linear model.

(ii) Ridge regression, where the tuning parameter λ is chosen with generalized cross validation. See RidgeExample.R for an example of how to do this in R. Plot the generalized cross validation value as a function of λ so that the minimum value is clearly visible.

(a) Which model has the smaller training error (see problem 2)? Why is training error a poor judge of how well the model will predict future data?

(b) Now use the models you fit in (i) and (ii) to predict the held out data. Which model performs better? Clearly indicate the metric (i.e. loss function) you used to judge model performance. (**Hint**: since you are using squared loss to choose λ, you might want to use squared loss to judge prediction...)