Homework 2 (Oct. 5th)

Deadline: Wednesday, October 19th, at 11:59pm.

Submission: You need to submit one file through Quercus with your answers to Questions 1, 2 and 3, as well as R code and R outputs requested for Questions 2 and 3. It should be a PDF file titled `hw2_writeup.pdf`. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Neatness Point: You will be deducted one point if we have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the total possible marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

- **Problem 1 (8 pts)**
  I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. by hypothesizing the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

  1. (**2 pts**) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

  2. (**2 pts**) Answer (a) using test error rather than training RSS.

  3. (**2 pts**) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

  4. (**2 pts**) Answer (c) using test error rather than training RSS.

- **Problem 2 (17 pts)**
  We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

  1. (**2 pts**) Generate a data set with $p = 20$ features, $n = 1000$ observations, and an associated quantitative response vector generated according to the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}$ satisfies $\beta_1 = \beta_2 = \beta_3 = 2, \beta_4 = \beta_5 = 0.5$ and $\beta_6 = \cdots = \beta_{20} = 0$. The design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ has entries generated as i.i.d. realizations of $N(0, 1)$. The error $\boldsymbol{\epsilon} \in \mathbb{R}^n$ also contains entries generated as i.i.d. realizations of $N(0, 1)$.

(**Note:** for reproducibility, you need to specify `set.seed(0)` at the beginning before generating the data.)

2. (**1 pts**) Randomly split your dataset into a training set containing 100 observations and a test set containing 900 observations.

3. (**2 pts**) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.

4. (**2 pts**) Plot the test set MSE associated with the best model of each size.

5. (**2 pts**) For which model size do the training set MSE and test set MSE take on their minimum value? Comment on your results.

6. (**2 pts**) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values

7. (**2 pts**) Create a plot displaying

$$\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^{(k)})^2}$$

for a range of values of k, where $\hat{\beta}_j^{(k)}$ is the jth coefficient estimate for the best model containing k coefficients. Comment on what you observe. How does this compare to the test MSE plot from part 4?

8. (**2 pts**) Repeat steps 3 - 6 for forward stepwise selection.

9. (**2 pts**) Repeat steps 3 - 6 for backward stepwise selection.

- **Problem 3 (12 pts)**

In this problem you will compare the performance of lasso and ridge regression in different linear models. Consider $p = 50$ and $n = 1100$.

(a) Set the random seed by using `set.seed(0)` and generate the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and the error $\boldsymbol{\epsilon} \in \mathbb{R}^n$ as part 1 of Problem 2.

(b) Generate the response $\mathbf{y} \in \mathbb{R}^n$ based on the model

$$y_i = \sum_{k=1}^{p} \beta_k X_{ik} + \epsilon_i, \qquad \forall i = 1, \ldots, n,$$

with $\beta_1 = \cdots = \beta_5 = 2$, and $\beta_j = 0$ for $j \geq 6$.

(c) Randomly split the data into a training set with 100 observations and a test set containing 1000 observations.

(d) Set the grid of λ by using the R command

```
grid = 10^seq(10,-2,length = 100)
```

1. (**3 pts**) Fit both the ridge regression and the lasso with λ selected by cross validation on the `grid` generated as above. Which method leads to a smaller test set MSE?

2. (**3 pts**) Repeat steps (a)–(d) for generating the data by using different seeds

   `set.seed(2),...,set.seed(50)`

   and also repeat part 1 for each seed. Save the test error for both, lasso and ridge for all seeds. Together with the results from part 1, this should give you 50 test MSEs for ridge and lasso. Make boxplots of the test errors for these two procedures and comment on the results.

3. (**6 pts**) Redo parts 1 and 2 by using $\beta_j = 0.5$ for all $j = 1, \ldots, 50$, in step (b).

- **Problem 1 (8 pts)**
  I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. by hypothesizing the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

  1. (**2 pts**) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

  2. (**2 pts**) Answer (a) using test error rather than training RSS.

  3. (**2 pts**) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

  4. (**2 pts**) Answer (c) using test error rather than training RSS.

11. Since we can say the model hypothesizing $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ has more variable than the true relationship, then the RSS for the hypothesizing model will be less than the true model, because the hypothesizing model is more flexible and fit the data because of more variable.

2). In the case of RSS, the hypothesizing model will have more test error because of the model fits training data set so much, and the actual one is just a linear relationship. Since we know from the question one that the hypothesizing model is likely to be overfitting on the training data, then the error for hypothesizing model is larger than actual one.

3). In this case, we can tell that since we do not know the relationship between X and Y, but we can ensure that the hypothesizing model which is a polynomial function can be more flexiable and fit the model well, and will reduce the training RSS.

4). However, for the test error, we could not judge the situation, there is not enough information for us to judge the true relationship. If it is linear as question 1 and 2, then the test error for linear is less than polynomial. Also, if it is polynomial with higher degress ($n \geq 2$) then the polynomial will have less test error than linear because it fits the relationship more.