CS 5565, HW5(Linear Model Selection and Regularization) 100pts.

Name ______________________________________

1. (15 points total) We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots p$ predictors. Explain your answers:

   (a) (5 points) Which of the three models with k predictors has the smallest *training* RSS?

   (b) (5 points) Which of the three models with k predictors has the smallest *test* RSS?

   (c) (5 points) True or False

      i. 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.

      ii. 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.

      iii. 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 forward stepwise selection.

      iv. 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 backward stepwise selection.

      v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

2. (15 points total) For parts (a) through (c), indicate which of i. through iv. is correct. Justify your answer.

   (a) (5 points) The lasso, relative to least squares, is:

      i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
      ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
      iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
      iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   (b) (5 points) Repeat (a) for ridge regression relative to least squares.

   (c) (5 points) Repeat (a) for non-linear methods relative to least squares.

3. (25 points total)Constraint vs. weight

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

$$\sum_{i-1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

   for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) (5 points)As we increase λ from 0, the training RSS will:

      i. Increase initially, and then eventually start decreasing in an inverted U shape.
      ii. Decrease initially, and then eventually start increasing in a U shape.
      iii. Steadily increase.
      iv. Steadily decrease.
      v. Remain constant.

   (b) (5 points)Repeat (a) for test RSS.

   (c) (5 points)Repeat (a) for variance.

   (d) (5 points) Repeat (a) for (squared) bias.

   (e) (5 points) Repeat (a) for the irreducible error.

4. (25 points total)Constraint vs. weight

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

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

   for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) (5 points) As we increase s from 0, the training RSS will:

      i. Increase initially, and then eventually start decreasing in an inverted U shape.
      ii. Decrease initially, and then eventually start increasing in a U shape.
      iii. Steadily increase.
      iv. Steadily decrease.
      v. Remain constant.

   (b) (5 points) Repeat (a) for test RSS.

   (c) (5 points) Repeat (a) for variance.

   (d) (5 points) Repeat (a) for (squared) bias.

   (e) (5 points)Repeat (a) for the irreducible error.

5. (10 points) Ridge regression vs. Lasso
Consider a simple special case with $n = p$, and $\boldsymbol{X}$ a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements. To simplify the problem further, assume also that we are performing regression without an intercept.

   With these assumptions, the least squares problem simplifies to finding $\beta_1, \ldots, \beta_p$ that minimize.

$$\sum_{j=1}^{p} (y_j - \beta_j)^2$$

   In this case the least squares solution can be found by taking the derivative with respect to β and setting to 0. We should get

$$\hat{\beta}_j = y_j$$

   (a) (5 points) For ridge regression use the same technique to find $\beta_1, \ldots, \beta_p$ that minimize.

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

   (b) (5 points) For lasso regression use the same technique to find $\beta_1, \ldots, \beta_p$ that minimize.

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

   (note: there will be 3 parts to this answer)

6. (10 points) Model selection Suppose you have found the best subsets of size 2,4,6,8, and 10 predictors for a data set of $n = 20$ and you need to choose the best model. Using AIC and BIC, determine which models would be best to use.

   Note: R uses the following expressions for AIC and BIC.

   AIC $= n \ln(RSS/n) + 2(p + 1)$
   BIC $= n \ln(RSS/n) + \ln(p + 1)$

p	2	4	6	8	10
RSS	220	200	190	187	186.8
AIC					
BIC					