CMDA 4654: Intermed Data Analytics & ML
Homework 3
Christopher J. Mobley
Due: Monday, 18APR16

1. (6.8 – 1) 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. Explain your answers:
    a. Which of the three models with k predictors has the smallest training RSS?
        While computationally less efficient than Forward and Backward Stepwise Selection, **Best Subset Selection** will have the smallest train RSS because its models do not depend on predictor(s) chosen in the previous kth steps. As a result, it searches over the largest possible set of models with k predictors.
    b. Which of the three models with k predictors has the smallest test RSS?
        Without testing this question is impossible to answer. While Best Subset Selection will have the smallest train RSS, this in no guarantees that its performance will exceed Forward or Backward Stepwise Selection on the test dataset.
    c. 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.
            True, in Forward Stepwise Selection going from a k to k+1 variable model involves adding the one predictor that reduces the RSS the most to the k variable model.
        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.
            True, in Backward Stepwise Selection going from a k+1 to a k variable model involves removing the one predictor from the k+1 variable model that leave us with the smallest RSS.
        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.
            False, forward and backward stepwise selection depend on the predictors selected/deleted by pervious iterations. As a result, the predictors depend on the order in which they were selected/deleted. Therefore, while it is certainly possible that they be a subset, this is not inherently true.
        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.
            False, forward and backward stepwise selection depend on the predictors selected/deleted by pervious iterations. As a result, the predictors depend on the order in which they were selected/deleted.

Therefore, while it is certainly possible that they be a subset, this is not inherently true.

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.

False, best subset selection picks the kth model with the smallest RSS from all possible subset not just the previous k -1 variable model. Therefore, while it is certainly possible that they be a subset, this is not inherently true.

2. (6.8 – 3) Suppose we estimate the regression coefficients in a linear regression model by minimizing

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

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

a. As we increase s from 0, the training RSS will:

As we increase s from 0, the training RSS will **steadily decrease** as flexibility of the model increases as the value of the $\beta_j s$ go from zero to their Ordinary Least Squares values.

b. Repeat (a) for test RSS

As we increase s from 0, the test RSS will **decrease initially, and then eventually start increasing in a U shape**. As we increase s from 0, the flexibility of model also increases as the value of the $\beta_j s$ go from zero to their Ordinary Least Squares values. As the β_js approach their Ordinary Least Values they will begin to over fit the training dataset resulting in increased RSS on test dataset. This is also known as the bias-variance tradeoff.

c. Repeat (a) for variance.

As we increase s from 0, the variance will **steadily increase** due to the increased flexibility of our model.

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

i. As we increase s from 0, the squared bias will **steadily decrease** due to the increased flexibility of our model.

e. Repeat (a) for the irreducible error.

i. As we increase s from 0, the irreducible error as its name suggests will **remain constant** as it is independent of the chosen model and therefore s.

3. (6.8 – 5) It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting. Suppose that n = 2, p = 2, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

a. Write out the ridge regression optimization problem in this setting.

$$\text{minimize} \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^2$$

$$\text{minimize}\left(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12}\right)^2 + \left(y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22}\right)^2 + \lambda\hat{\beta}_1^2 + \lambda\hat{\beta}_2^2$$

$$\text{minimize}\left(y_1 - (\hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12})\right)^2 + \left(y_2 - (\hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22})\right)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

Since $x_{11} = x_{12}$, $x_{21} = x_{22}$

$$\text{minimize}\left(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2 + \left(y_2 - (\hat{\beta}_1 + \hat{\beta}_2)x_{21}\right)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

Since $x_{12} + x_{22} = 0$, $x_{11} = -x_{21}$

$$\text{minimize}\left(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2 + \left(y_2 + (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

Since $y_1 + y_2 = 0$, y1 = -y2

$$\left(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2 + \left(-y_1 + (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

In order to minimize the ridge regression, we must take the derivatives with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ and set them equal to 0.

$$\frac{d}{d\hat{\beta}_1}(y_1^2 - 2y_1\hat{\beta}_1 x_{11} - 2y_1\hat{\beta}_2 x_{11} + \hat{\beta}_1^2 x_{11}^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{11}^2 + \hat{\beta}_2^2 x_{11}^2 + y_1^2 - 2y_1\hat{\beta}_1 x_{11} - 2y_1\hat{\beta}_2 x_{11} + \hat{\beta}_1^2 x_{11}^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{11}^2 + \hat{\beta}_2^2 x_{11}^2 + \lambda\hat{\beta}_1^2 + \lambda\hat{\beta}_2^2) = 0$$

$$4(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})(-x_{11}) + 2\lambda\hat{\beta}_1 = 0$$

Solve for $\hat{\beta}_1$ using TI-89

$$\hat{\beta}_1 = \frac{-2x_{11}(\hat{\beta}_2 x_{11} - y_1)}{\lambda + 2x_{11}^2}$$

Similarly

$$\frac{d}{d\hat{\beta}_2}(y_1^2 - 2y_1\hat{\beta}_1 x_{11} - 2y_1\hat{\beta}_2 x_{11} + \hat{\beta}_1^2 x_{11}^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{11}^2 + \hat{\beta}_2^2 x_{11}^2 + y_1^2 - 2y_1\hat{\beta}_1 x_{11} - 2y_1\hat{\beta}_2 x_{11} + \hat{\beta}_1^2 x_{11}^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{11}^2 + \hat{\beta}_2^2 x_{11}^2 + \lambda\hat{\beta}_1^2 + \lambda\hat{\beta}_2^2) = 0$$

$$4(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})(-x_{11}) + 2\lambda\hat{\beta}_2 = 0$$

Solve for $\hat{\beta}_2$ using TI-89

$$\hat{\beta}_2 = \frac{-2x_{11}(\hat{\beta}_1 x_{11} - y_1)}{\lambda + 2{x_{11}}^2}$$

b. Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

$$\frac{-2x_{11}(\hat{\beta}_2 x_{11} - y_1)}{\lambda + 2{x_{11}}^2} = \frac{-2x_{11}(\hat{\beta}_1 x_{11} - y_1)}{\lambda + 2{x_{11}}^2}$$

$$-2x_{11}(\hat{\beta}_2 x_{11} - y_1) = -2x_{11}(\hat{\beta}_1 x_{11} - y_1)$$

$$\hat{\beta}_2 x_{11} - y_1 = \hat{\beta}_1 x_{11} - y_1$$

$$\hat{\beta}_2 = \hat{\beta}_1$$

Due to Symmetry between $\hat{\beta}_1$ and $\hat{\beta}_2$, we can conclude that $\hat{\beta}_1 = \hat{\beta}_2$

c. Write out the lasso optimization problem in this setting.

$$\text{minimize} \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j|$$

$$\text{minimize } (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda|\hat{\beta}_1| + \lambda|\hat{\beta}_2|$$

$$\text{minimize } (y_1 - (\hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12}))^2 + (y_2 - (\hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22}))^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$$

Since $x_{11} = x_{12}$, $x_{21} = x_{22}$

$$\text{minimize } (y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})^2 + (y_2 - (\hat{\beta}_1 + \hat{\beta}_2)x_{21})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$$

Since $x_{12} + x_{22} = 0$, $x_{11} = -x_{21}$

$$\text{minimize } (y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})^2 + (y_2 + (\hat{\beta}_1 + \hat{\beta}_2)x_{11})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$$

Since $y_1 + y_2 = 0$, $y1 = -y2$

$$(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})^2 + (-y_1 + (\hat{\beta}_1 + \hat{\beta}_2)x_{11})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$$

In order to minimize the ridge regression, we must take the derivatives with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ and set them equal to 0.

$$\frac{d}{d\hat{\beta}_1}(y_1{}^2 - 2y_1\hat{\beta}_1x_{11} - 2y_1\hat{\beta}_2x_{11} + \hat{\beta}_1{}^2x_{11}{}^2 + 2\hat{\beta}_1\hat{\beta}_2x_{11}{}^2 + \hat{\beta}_2{}^2x_{11}{}^2 + y_1{}^2 - 2y_1\hat{\beta}_1x_{11}$$
$$- 2y_1\hat{\beta}_2x_{11} + \hat{\beta}_1{}^2x_{11}{}^2 + 2\hat{\beta}_1\hat{\beta}_2x_{11}{}^2 + \hat{\beta}_2{}^2x_{11}{}^2 + \lambda|\hat{\beta}_1| + \lambda|\hat{\beta}_2|) = 0$$

$$4(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})(-x_{11}) + \lambda|\hat{\beta}_1| = 0$$

Solve for $\hat{\beta}_1$ using TI-89

$\hat{\beta}_1 = \frac{-4x_{11}(\hat{\beta}_2x_{11}-y_1)}{\lambda+4x_{11}{}^2}$ and $\hat{\beta}_1 = \frac{x_{11}(\hat{\beta}_2x_{11}-y_1)}{\lambda+4x_{11}{}^2} \leq 0$ or $\hat{\beta}_1 = \frac{4x_{11}(\hat{\beta}_2x_{11}-y_1)}{\lambda-4x_{11}{}^2}$ and $\hat{\beta}_1 = \frac{x_{11}(\hat{\beta}_2x_{11}-y_1)}{\lambda-4x_{11}{}^2} < 0$

Similarly

$$\frac{d}{d\hat{\beta}_2}(y_1{}^2 - 2y_1\hat{\beta}_1x_{11} - 2y_1\hat{\beta}_2x_{11} + \hat{\beta}_1{}^2x_{11}{}^2 + 2\hat{\beta}_1\hat{\beta}_2x_{11}{}^2 + \hat{\beta}_2{}^2x_{11}{}^2 + y_1{}^2 - 2y_1\hat{\beta}_1x_{11}$$
$$- 2y_1\hat{\beta}_2x_{11} + \hat{\beta}_1{}^2x_{11}{}^2 + 2\hat{\beta}_1\hat{\beta}_2x_{11}{}^2 + \hat{\beta}_2{}^2x_{11}{}^2 + \lambda|\hat{\beta}_1| + \lambda|\hat{\beta}_2|) = 0$$

$$4(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})(-x_{11}) + \lambda|\hat{\beta}_2| = 0$$

Solve for $\hat{\beta}_2$ using TI-89

$\hat{\beta}_2 = \frac{-4x_{11}(\hat{\beta}_1x_{11}-y_1)}{\lambda+4x_{11}{}^2}$ and $\hat{\beta}_2 = \frac{x_{11}(\hat{\beta}_1x_{11}-y_1)}{\lambda+4x_{11}{}^2} \leq 0$ or $\hat{\beta}_2 = \frac{4x_{11}(\hat{\beta}_1x_{11}-y_1)}{\lambda-4x_{11}{}^2}$ and $\hat{\beta}_2 = \frac{x_{11}(\hat{\beta}_1x_{11}-y_1)}{\lambda-4x_{11}{}^2} < 0$

d. Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

As one can see from part (c), the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique – in other words, there are many possible solutions to the optimization problem in (c).

4. (6.8 – 7) We will now derive the Bayesian connection to the lasso and ridge regression discussed in Section 6.2.2.
   a. Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where ϵ_1,..., ϵ_n are independent and identically distributed from a *N(0, σ2)* distribution. Write out the likelihood for the data.

$$L(\theta|\beta) = p(\beta|\theta)$$
$$= p(\beta_1|\theta) * \dots * p(\beta_n|\theta)$$

$$= \prod_{i=1}^{n} p(\beta_i|\theta)$$

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(y_i-\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2}\right)}$$

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2}\right)}$$

b. Assume the following prior for β: $\beta_1,\ldots,\beta_p$ are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b: i.e $p(\beta) = \frac{1}{2b} e^{-\frac{|\beta|}{b}}$ .Write out the posterior for β in this setting.

$$p(\beta|X,Y) \propto f(Y|Y,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2}\right)} \left(\frac{1}{2b} e^{-\frac{\sum_{j=1}^{p}|\beta_j|}{b}}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right) e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2} - \frac{\sum_{j=1}^{p}|\beta_j|}{b}\right)}$$

c. Argue that the lasso estimate is the *mode* for β under this posterior distribution.

In order to find the mode for β under this posterior distribution, we must find the most likely value for β or its MAP value. This is found by maximizing the solution given in (b).

$$\arg\max_{\beta} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right) e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2} - \frac{\sum_{j=1}^{p}|\beta_j|}{b}\right)}$$

$$\arg\max_{\beta} \log\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right) e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2} - \frac{\sum_{j=1}^{p}|\beta_j|}{b}\right)}\right)$$

$$\arg\max_{\beta} \log\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right)\right) - \left(\frac{\sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2} + \frac{\sum_{j=1}^{p}|\beta_j|}{b}\right)$$

The maximum of the following equation occurs at the minimum of the second term

$$\arg\min_{\beta} \frac{\sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2}{2\sigma^2} + \frac{\sum_{j=1}^{p}|\beta_j|}{b}$$

$$\arg\min_{\beta} \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|\right)$$

$\frac{1}{2\sigma^2}$is a constant. Therefore the minimization of the function does not depend on it.

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|$$

If we let $\lambda = \frac{2\sigma^2}{b}$ then

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

$$= \arg\min_{\beta} RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

This is equal to the lasso estimate given by equation 6.7 in Introduction to Statistical Learning. As a result, when the posterior comes from a double-exponential distribution with mean zero and common scale parameter b, the mode for β is given by the Lasso estimate with $\lambda = \frac{2\sigma^2}{b}$.

d. Now assume the following prior for *β: $\beta_1,\ldots,\beta_p$* are independent and identically distributed according to a normal distribution with mean zero and variance c. Write out the posterior for *β* in this setting.

$$p(\beta) = \prod_{i=1}^{p} p(\beta_i)$$

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi c}} e^{\left(-\frac{\beta_i^2}{2c}\right)}$$

$$= \left(\frac{1}{\sqrt{2\pi c}}\right)^{p} e^{\left(-\frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}\right)}$$

$$p(\beta|X,Y) \propto f(Y|Y,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{n} e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma^2}\right)} \left(\frac{1}{\sqrt{2\pi c}}\right)^{p} e^{\left(-\frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}\right)}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{n} \left(\frac{1}{\sqrt{2\pi c}}\right)^{p} e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma^2}-\frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}\right)}$$

e. Argue that the ridge regression estimate is both the *mode* and the *mean* for β under this posterior distribution.

In order to find the mode and mean for β under this posterior distribution, we must find the most likely value for β or its MAP value. This is found by maximizing the solution given in (d).

$$\arg\max_{\beta} \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{n} \left(\frac{1}{\sqrt{2\pi c}}\right)^{p} e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma^2}-\frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}\right)}$$

$$\arg\max_{\beta} \log\left(\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{n} \left(\frac{1}{\sqrt{2\pi c}}\right)^{p} e^{\left(-\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma^2}-\frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}\right)}\right)$$

$$\arg\max_{\beta} \log\left(\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{n} \left(\frac{1}{\sqrt{2\pi c}}\right)^{p}\right) - \left(\frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma^2} + \frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}\right)$$

The maximum of the following equation occurs at the minimum of the second term

$$\arg\min_{\beta} \frac{\sum_{i=1}^{n}(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j)^2}{2\sigma^2} + \frac{\sum_{i=1}^{p}\beta_i{}^2}{2c}$$

$$\arg\min_{\beta} \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \frac{2\sigma^2}{c}\sum_{i=1}^{p}\beta_i{}^2\right)$$

$\frac{1}{2\sigma^2}$is a constant. Therefore the minimization of the function does not depend on it.

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \frac{2\sigma^2}{c}\sum_{i=1}^{p}\beta_i{}^2$$

If we let $\lambda = \frac{2\sigma^2}{c}$ then

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{i=1}^{p}\beta_i{}^2$$

$$= \arg\min_{\beta} RSS + \lambda\sum_{i=1}^{p}\beta_i{}^2$$

This is equal to the ridge regression estimate given by equation 6.5 in Introduction to Statistical Learning. As a result, when the posterior comes from a normal distribution with mean zero and variance c, the mode for β is given by the Ridge Regression estimate with $\lambda = \frac{2\sigma^2}{c}$. Since the posterior is given by a Gaussian distribution, we know it is also the posterior mean.

5. (6.8 – 8) In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.
   a. Use the rnorm() function to generate a predictor X of length n = 100, as well as a noise vector ϵ of length n = 100.

   b. Generate a response vector Y of length n = 100 according to the model
   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$
   where β_0, β_1, β_2, and β_3 are constants of your choice.

   c. Use the regsubsets() function to perform best subset selection in order to choose the best model containing the predictors X, X^2,...,X^{10}. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the data.frame() function to create a single data set containing both X and Y.

      Figure 1 shows that the best model obtain by Cp, BIC, and adjusted R^2 is a three variable model. Figure 2 shows each β_j is all less than 1% away from its true values.

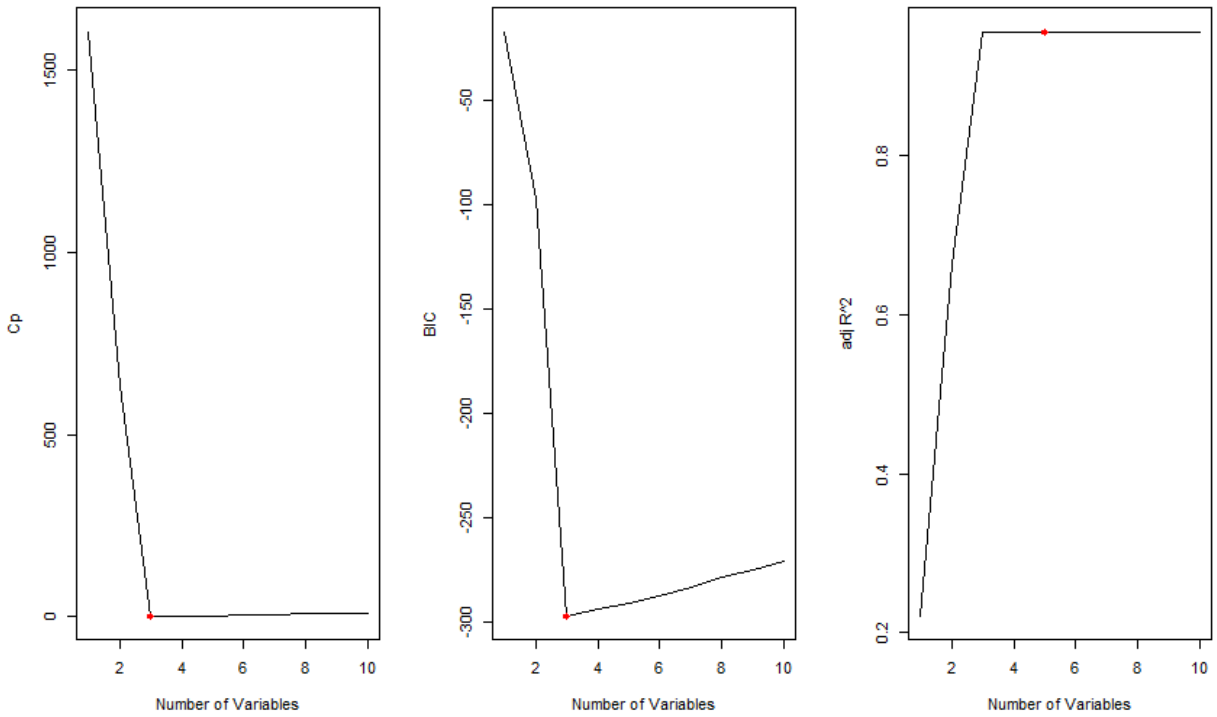

Figure 1. Summary of Cp, BIC, and adjusted R2 for each Best Subset Selection Model

(Intercept)	X	X2	X3
0.4966017	0.7530186	0.2470644	-0.2449610

Figure 2. Summary of β_j for the best Best Subset Selection model

d. Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

Figure 3 and 5 shows that the best model obtain by Cp, BIC, and adjusted R^2 is a three variable model just like in part (c). Figure 4 and 6 shows each β_j is less than 1% away from its true values as in part (c).

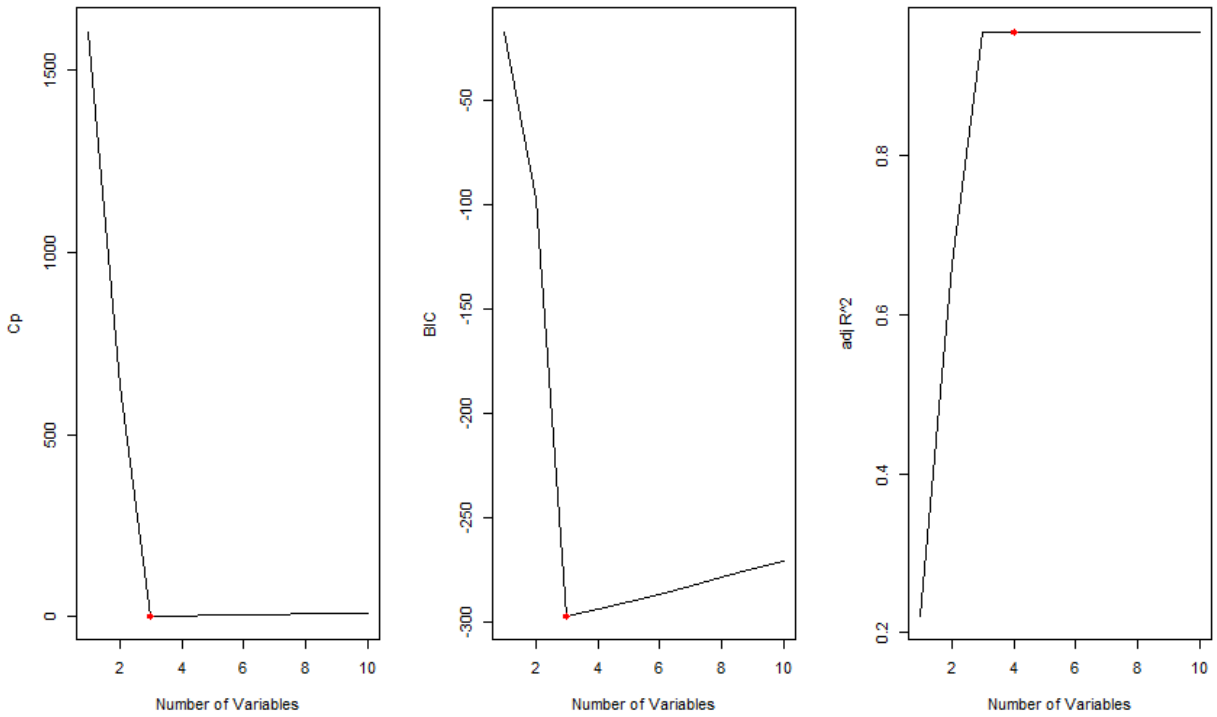

Figure 3. Summary of Cp, BIC, and adjusted R2 for each Forward Stepwise Selection Model

(Intercept)	X	X2	X3
0.4966017	0.7530186	0.2470644	-0.2449610

Figure 4. Summary of β_j for the best Forward Subset Selection model

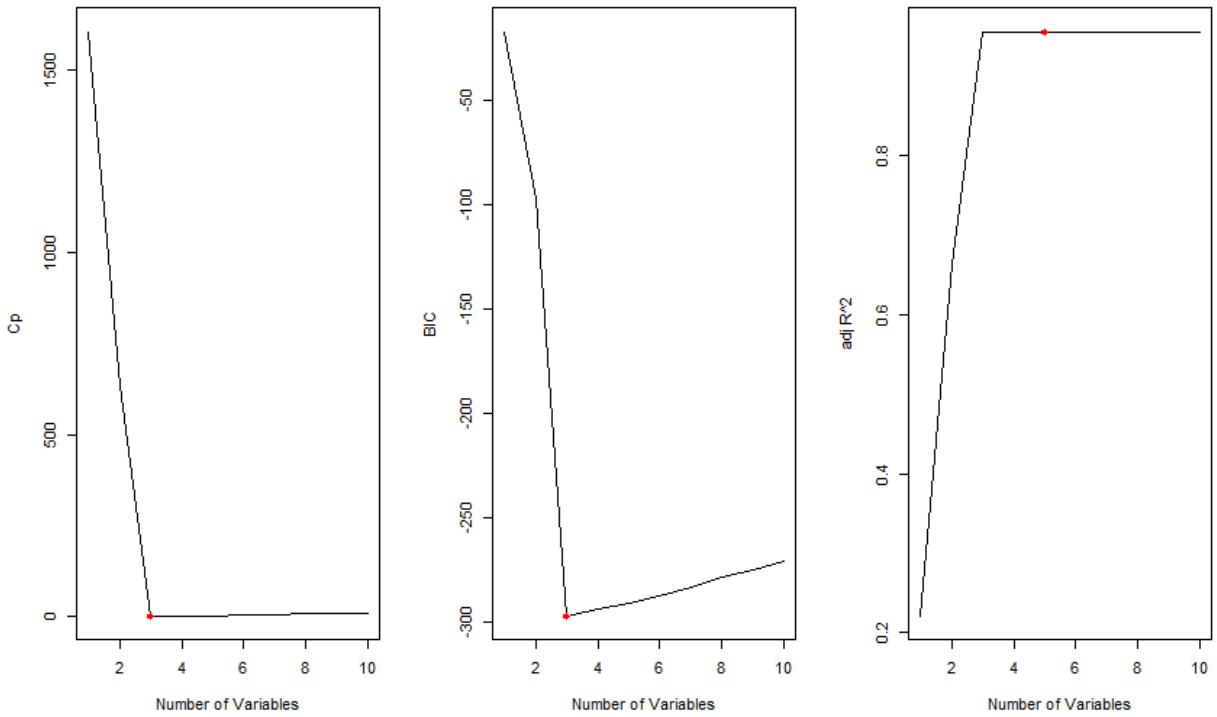

Figure 5. Summary of Cp, BIC, and adjusted R2 for each Backward Stepwise Selection Model

(Intercept)	X	X2	X3
0.4966017	0.7530186	0.2470644	-0.2449610

Figure 6. Summary of β_j for the best Backwards Subset Selection model

e. Now fit a lasso model to the simulated data, again using X, X^2, ...,X^{10} as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

The optimal value of λ is approximately 0.002. Figure 7 shows a plot of the cross-validation error as a function of λ and Figure 8 shows the resulting coefficients. While the lasso estimate for the intercept is within 1% error the remaining values are off and X4 and X5 were picked despite being negligible in the actual model.

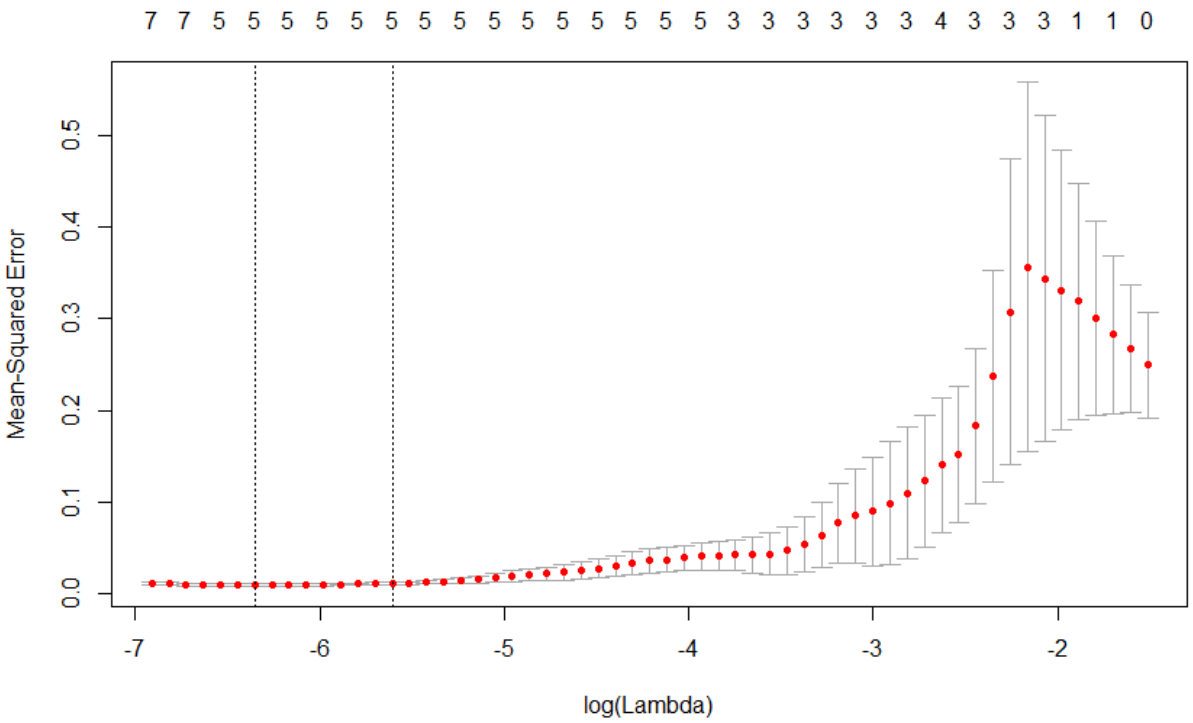

Figure 7. Cross-validation error as a function of λ for Lasso

```
12 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  0.504089885
(Intercept)  .
X            0.591597419
X2           0.225296348
X3          -0.110318072
X4           0.004114621
X5          -0.018347713
X6           .
X7           .
X8           .
X9           .
X10          .
```

Figure 8. Resulting coefficients for Lasso

f. Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \epsilon$$

and perform best subset selection and the lasso. Discuss the results obtained.

Unlike the previous Y from which we obtain better results using best subset selection than lasso, Figure 9 through 12 show that lasso obtain better results for the 1 variable model above.

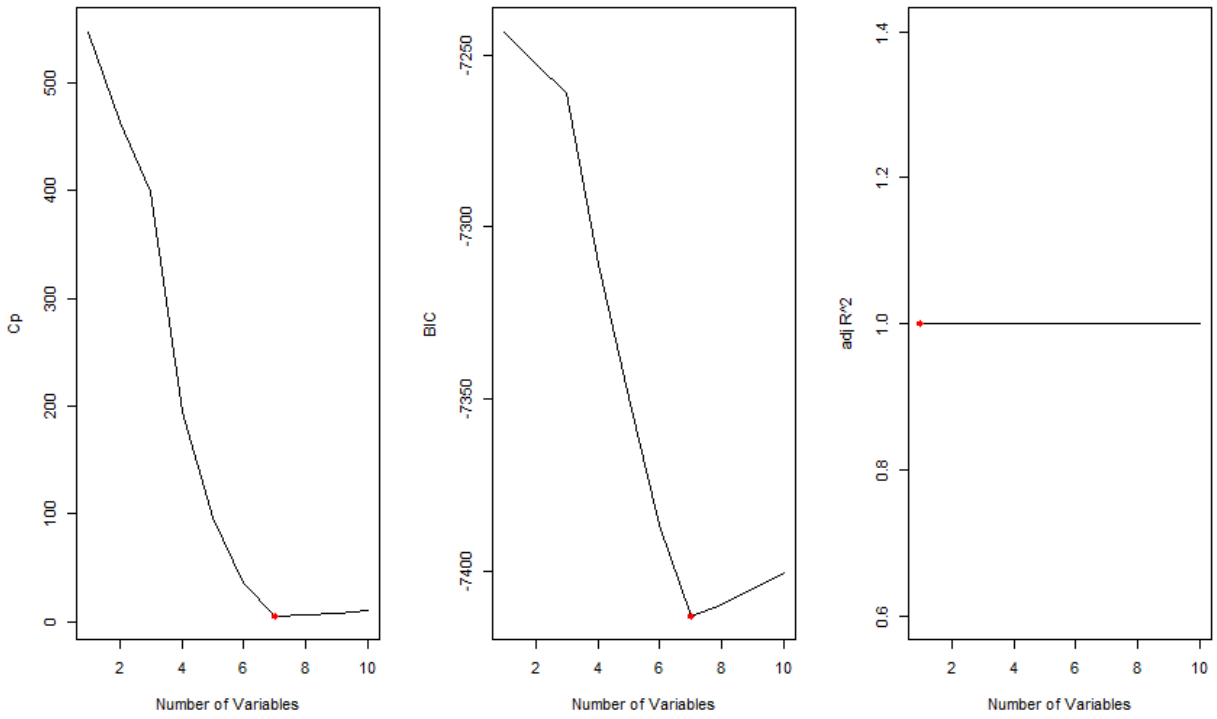

Figure 9. Summary of Cp, BIC, and adjusted R2 for each Best Subset Selection model

(Intercept)	Y	X2	X4	X6	X8	X9	X10
4.440892e-16	1.000000e+00	-1.421085e-14	2.664535e-14	-2.009504e-14	5.911938e-15	5.551115e-17	-5.551115e-16

Figure 10. Summary of β_j for the best Best Subset Selection model

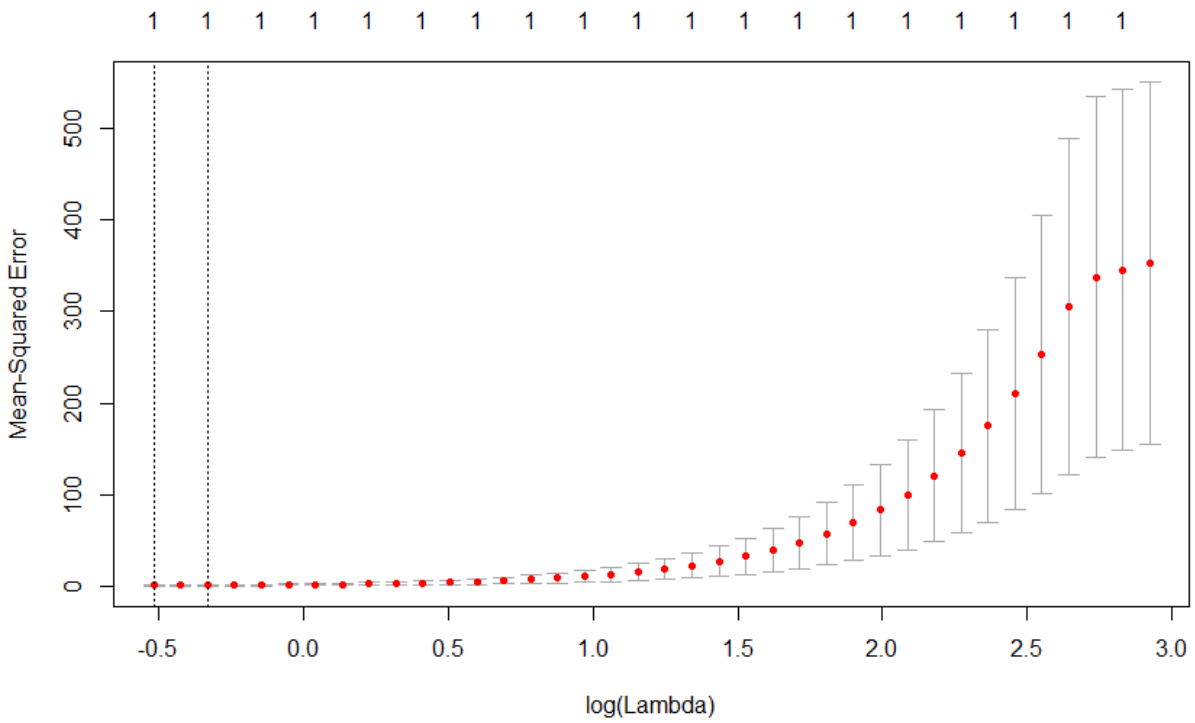

Figure 11. Cross-validation error as a function of λ for Lasso

```
12 x 1 sparse Matrix of class "dgCMatrix"
                            1
(Intercept)  0.435530606
(Intercept)  .
X            .
X2           .
X3           .
X4           .
X5          -0.000200665
X6           .
X7          -0.241818770
X8           .
X9           .
X10          .
```

Figure 12. Resulting coefficients for Lasso

6. (6.8 – 9) In this exercise, we will predict the number of applications received using the other variables in the College data set.
    a. Split the data set into a training set and a test set.

    b. Fit a linear model using least squares on the training set, and report the test error obtained.

        The test MSE is 1615967.

    c. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

With a chosen λ of 351.9759, the test MSE is 2580252.

d. Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

With a chosen λ of 17.51792, the test MSE is 1729118. Figure 13 shows the value of all non-zero coefficient estimates.

```
19 x 1 sparse Matrix of class "dgCMatrix"
                          1
(Intercept) -1.065356e+03
(Intercept)  .
PrivateYes  -3.634484e+02
Accept       1.255288e+00
Enroll       .
Top10perc    2.181212e+01
Top25perc    .
F.Undergrad  3.014998e-02
P.Undergrad  .
Outstate    -3.805344e-02
Room.Board   1.086076e-01
Books        6.892697e-02
Personal    -1.201697e-02
PhD         -6.743534e+00
Terminal     .
S.F.Ratio    1.302591e+01
perc.alumni -6.889357e+00
Expend       9.336182e-02
Grad.Rate    7.898015e+00
```

Figure 13. Resulting coefficients for Lasso

e. Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

With a chosen M of 17, the MSE is 1615967.

f. Fit a PLS model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

With a chosen M of 10, the MSE is 1601426.

g. Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

Part (b) through (f) show that the MSE error all model is relatively close. However, ridge regression does preform significantly worse than the other model. In addition the MSE for OLS and PCR is the same suggesting we used the entire model for PCR. Figure 14 shows a comparison of the

model's R^2 value of our models. The values are all close to 0.9, suggesting that they are all fairly accurate in predicting the number of college applications. There isn't much difference in test error amount the five approaches. However, as mentioned before ridge regression does perform slightly worse.

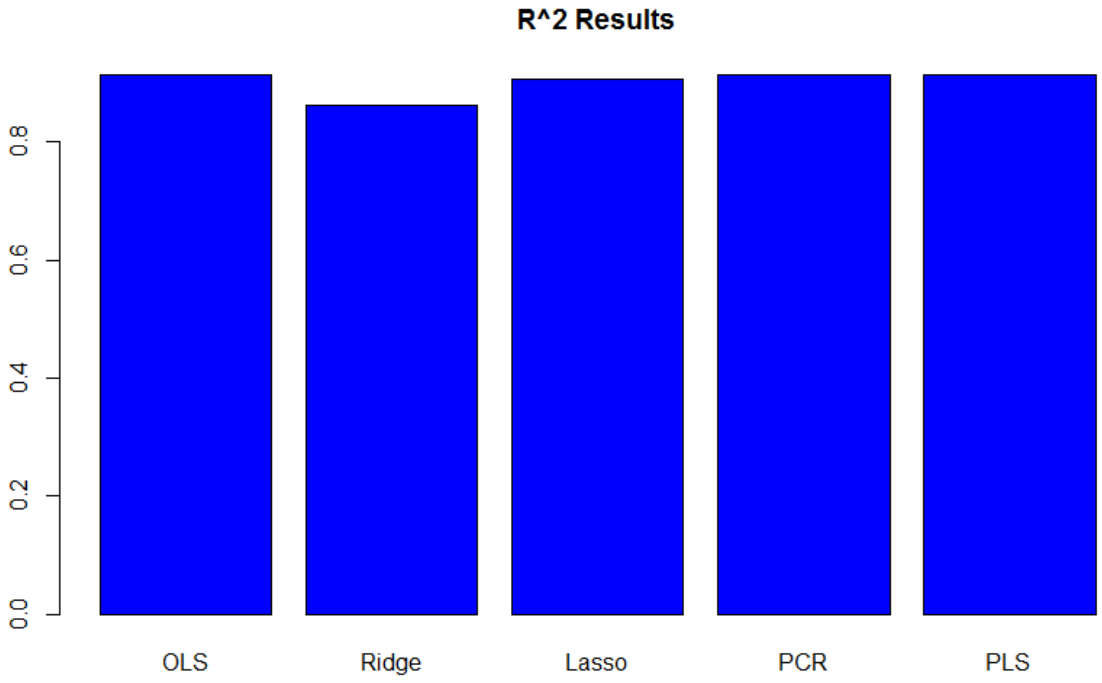

Figure 14. R^2 Comparison