HW6 R Markdown

Santanu Mukherjee

11/21/2021

R Markdown

Chapter 5 page 197:

Q1

Using basic statistical properties of the variance, as well as single variable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $Var(\alpha X + (1-\alpha)Y)$.

Answer 1

So, we have

$Var(\alpha X + (1-\alpha)Y)$ = $\alpha^2 \sigma_X^2$+$(1-\alpha)^2\sigma_Y^2$+$2\alpha(1-\alpha)\sigma_X\sigma_Y$.

We now take the **First Derivative** of $Var(\alpha X + (1-\alpha)Y)$ relative to α and we get

$$\frac{\delta}{\delta\alpha}$$

$Var(\alpha X + (1-\alpha)Y)$ = $2\alpha\sigma_X^2$ - $2\sigma_Y^2$ + $2\alpha\sigma_Y^2$ + $2\sigma_{XY}$ - $4\alpha\sigma_{XY}$.

Equating the above expression to 0 gives us teh following equation:

$2\alpha\sigma_X^2$ - $2\sigma_Y^2$ + $2\alpha\sigma_Y^2$ + $2\sigma_{XY}$ - $4\alpha\sigma_{XY}$ = 0,

which implies

α =

$$\frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

.

The above is the **derivation** for (5.6).

To provide that this is the **minimum**, we have to prove that the **Second Derivative** is **greater than** 0.

$$\frac{\delta^2}{\delta\alpha^2}$$

$Var(\alpha X + (1-\alpha)Y)$ = $2\sigma_X^2$ + $2\sigma_Y^2$ - $4\sigma_{XY}$ = $2Var(X-Y)$, which is $>= 0$ as variance is always **positive**.

Q2

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

Q2a

What is the probability that the first bootstrap observation is not the jth observation from the original sample ? Justify your answer.

Answer 2 (a)

$1 - 1/n$

Q2b

What is the probability that the second bootstrap observation is not the jth observation from the original sample ?

Answer 2 (b)

$1 - 1/n$

Q2c

Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - 1/n)^n$

Answer 2 (c)

In bootstrapping, we sample with replacement, and so the probability that the jth observation is **NOT** in the bootstrap sample is the product of the probabilities that each bootstrap observation is **NOT** the jth observation from the original sample, which means

$(1 - 1/n)$ * $(1 - 1/n)$ * ………. * $(1 - 1/n)$ = $(1 - 1/n)^n$, as these probabilities are **independent**.

Q2d

When $n = 5$, what is the probability that the jth observation is in the bootstrap sample ?

Answer 2 (d)

So, $P(jth\ observation\ in\ bootstrap\ sample)$ = $(1 - 1/5)^5$ = 0.672

Q2e

When $n = 100$, what is the probability that the jth observation is in the bootstrap sample ?

Answer 2 (e)

So, $P(jth\ observation\ in\ bootstrap\ sample)$ = $(1 - 1/100)^{100}$ = 0.634

Q2f

When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample ?

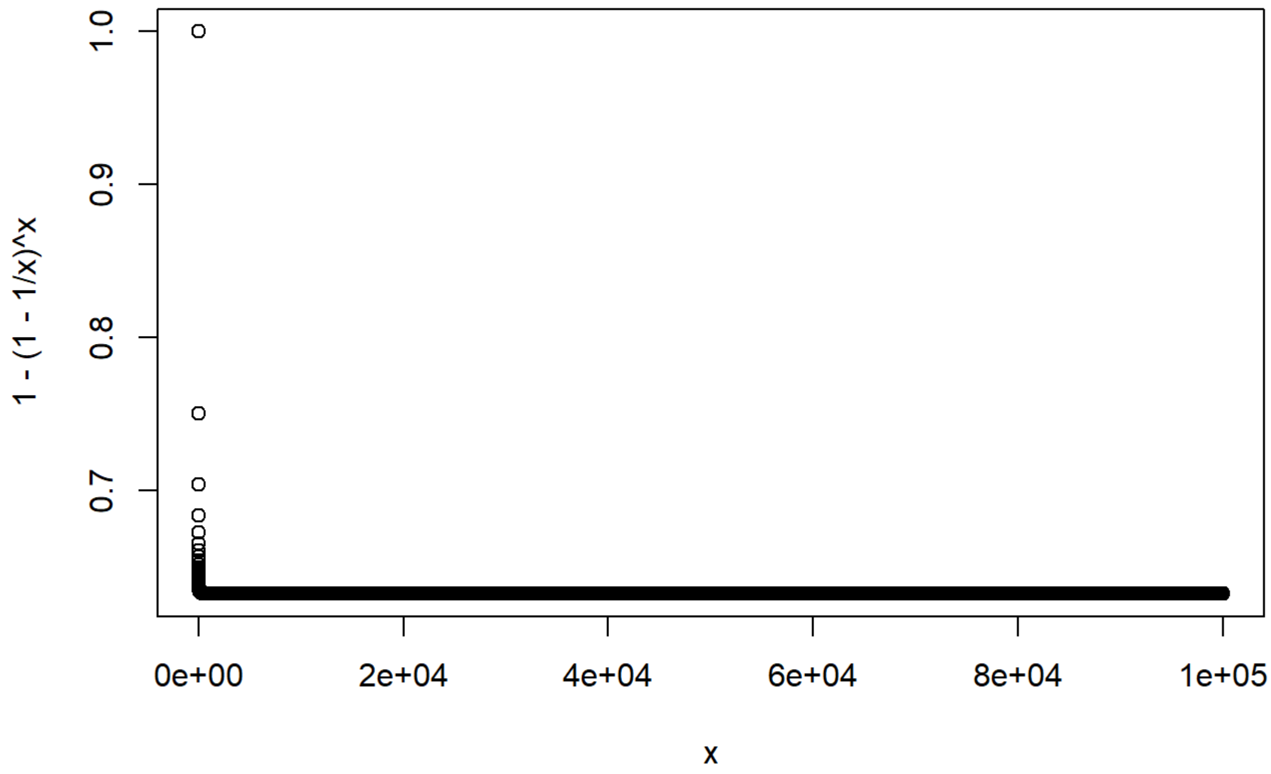
Answer 2 (f)

So, $P(jth\ observation\ in\ bootstrap\ sample)$ = $(1 - 1/10,000)^{10,000}$ = 0.632

Q2g

Create a plot that displays, for each integer value of n from $1\ to\ 100,000$, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

Answer 2 (g)



It might be seen that the plot quickly reaches an asymptotic value of about 0.632.

Q2h

We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the **fourth** observation is contained in the bootstrap sample.

```
store <- rep(NA, 10000)
for (i in 1:10000) {
    store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
}
mean(store)
```

```
## [1] 0.6339
```

Comment on the results obtained.

Answer 2 (h)

From calculus, we know that $\lim_{x\to\infty}(1+x/n)^n$ = e^x.
Now if we apply the above to our situation here, we get the probability that a bootstrap sample of size n contains the jth observation converges to $1 - 1/e$ = 0.632 as $n- > \infty$.

Q3

We now review $k - fold$ cross-validation.

Q3a

Explain how $k - fold$ cross-validation is implemented.

Answer 3 (a)

The $k - fold$ cross validation is implemented by taking the n observations and randomly splitting it into k non-overlapping groups of length of (approximately) n/k. These groups acts as a validation set, and the remainder of length $(n - n/k)$ acts as a training set. The test error is then estimated by averaging the k resulting MSE estimates.

Q3b

What are the advantages and disadvantages of $k - fold$ cross-validation relative to:

Q3b i

i. The validation set approach ?

Answer 3 (b i)

The validation set approach has two main drawbacks compared to $k - fold$ cross-validation. Firstly, the validation estimate of the test error rate can be highly variable. This is because depending on which observations are included in the training set and which observations are included in the validation set, this is vary significantly. Secondly, only a subset of the observations are used to fit the model. Since statistical methods tend to perform worse when trained on fewer data observations, this suggests that the validation set (contains fewer observations as this is a subset of the entire data set) error rate may tend to overestimate the test error rate for the model fit on the entire data set.

Q3b ii

ii. **LOOCV ?**

Answer 3 (b ii)

So, the $LOOCV$ cross-validation approach is a special case of $k - fold$ cross-validation in which $k = n$. This approach has two drawbacks compared to $k - fold$ cross-validation. Firstly, it requires fitting the potentially computationally expensive model n times compared to $k - fold$ cross-validation which requires the model to be fitted only k times. Secondly, the $LOOCV$ cross-validation approach may give approximately unbiased estimates of the test error, since each training set contains $n - 1$ observations. However, this approach has higher variance than $k - fold$ cross-validation (since we are averaging the outputs of n fitted models trained on an almost identical set of observations, these outputs are highly correlated, and the mean of highly correlated quantities has higher variance than less correlated ones). So, there is a bias-variance trade-off associated with the choice of k in $k - fold$ cross-validation. Typically using $k = 5$ or $k = 10$ yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

Q5

In Chapter 4, we used logistic regression to predict the probability of *default* using *income* and *balance* on the **Default** data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

Q5a

Fit a logistic regression model that uses *income* and *balance* to predict *default*.

Answer 5 (a)

```
## 
## Call:
## glm(formula = default ~ income + balance, family = "binomial", 
##     data = Default)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
## 
## Number of Fisher Scoring iterations: 8
```

Q5b i

Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

i. Split the sample set into a training set and a validation set.

Answer 5 (b i)

Randomly selected **50%** of the observations for the training set and **50%** for the test set.

Q5b ii

ii. Fit a multiple logistic regression model using only the training observations.

Answer 5 (b ii)

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5830  -0.1428  -0.0573  -0.0213   3.3395
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.194e+01  6.178e-01 -19.333  < 2e-16 ***
## income       3.262e-05  7.024e-06   4.644 3.41e-06 ***
## balance      5.689e-03  3.158e-04  18.014  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1523.8  on 4999  degrees of freedom
## Residual deviance:  803.3  on 4997  degrees of freedom
## AIC: 809.3
##
## Number of Fisher Scoring iterations: 8
```

Q5b iii

iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the **default** category if the posterior probability is greater than 0.5.

Answer 5 (b iii)

```
##  [1] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
```

Showing the first 10 entries of the predition set.

Q5b iv

iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

Answer 5 (b iv)

```
## [1] "The test error percent with the validation set approach is 2.54"
```

Q5c

Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

Answer 5 (c)

```
## [1] 0.0274
```

```
## [1] 0.0244
```

```
## [1] 0.0244
```

Our observation is that the validation estimate of the test error rate can vary, depending on the training set data and the validation set data.

Q6

We continue to consider the use of a logistic regression model to predict the probability of **default** using **income** and **balance** on the **Default** data set. In particular, we will now compute estimates for the standard errors of the **income** and **balance** logistic regression coefficients in two different ways : (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the $glm()$ function. Do not forget to set a random seed before beginning your analysis.

Q6a

Using the $summary()$ and $glm()$ functions, determine the estimated standard errors for the coefficients associated with **income** and **balance** in a multiple logistic regression model that uses both predictors.

Answer 6 (a)

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

The $glm()$ estimated standard errors for the coefficients associated with **income** and **balance** are $4.985e-06$ and $2.274e-04$ respectively. The intercept estimate is $4.348e-01$.

Q6b

Write a function, $boot.fn()$, that takes as input the **Default** data set as well as an index of the observations, and that outputs the coefficient estimates for **income** and **balance** in the multiple logistic regression model.

Answer 6 (b)

Q6c

Use the $boot()$ function together with your $boot.fn()$ function to estimate the standard errors of the logistic regression coefficients for **income** and **balance**.

Answer 6 (c)

```
##   (Intercept)        income       balance
## -1.154047e+01  2.080898e-05  5.647103e-03
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Default, statistic = boot.fn, R = 100)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1* -1.154047e+01  8.556378e-03 4.122015e-01
## t2*  2.080898e-05 -3.993598e-07 4.186088e-06
## t3*  5.647103e-03 -4.116657e-06 2.226242e-04
```

Q6d

Comment on the estimated standard errors obtained using the $glm()$ function and using your bootstrap function.

Answer 6 (d)

Based on the output, the estimated standard errors obtained by the two methods are pretty close to each other.

Chapter 6 page 259:

Q2

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

Q2a

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.

Answer 2 (a)

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

$Reason$: Lasso is less flexible (for λ > 0), giving increased prediction accuracy provided that the he increase in bias is outweighed by the decrease in variance.

This is because lasso selects the $\hat{\beta}$ that minimizes RSS + $\lambda\sum_{i=1}^{p}|\beta i|$, and not just the RSS in least squares.

Since the shrinkage penalty $\lambda\sum_{i=1}^{p}|\beta i|$ is very small for {\beta_{1}},β_2……..β_p close to zero, this tends to shrink the estimates towards zero (because for a given λ > 0, the optimal $LASSO$ $\hat{\beta}$ will be closer to zero than the least squares $\hat{\beta}$). For a larger λ, the shrinkage terms importance is higher relative to the RSS, so the shrinkage increases.

This shrinkage is what reduces the variance of the predictions, at the cost of a small increase in bias.

Q2b

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

Answer 2 (b)

$iii.$ it is for the same reasons as part (a). Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

The only real difference here is that the ridge objective function to be minimized RSS + $\lambda\sum_{i=1}^{p}\beta i^2$, where the shrinkage term for ridge regression is a bit different to that of the $LASSO$.

The meaning of the previous statement is that ridge regression won't shrink coefficients of less-useful variables to exactly $Zero$ (the $LASSO$ can do this), but the rest of the arguments (shrinkage reducing the variance, thus increasing the bias) still applies.

Q2c

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

Answer 2 (c)

$ii.$ - More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

When we are using a non-linear method and have the relationship $Y = f(X)$+ ϵ, we are going to have a more flexible method as we are making less assumptions about the what the functional form of f (by not assuming linearity) is . In this case, where f is better approximated using non-linear relationships between the predictors and the response, this will lead to a decrease in bias that outweighs any increase in variance, and so we will have a higher prediction accuracy.

Q3

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

$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta j x_{ij})^2$ 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.

Q3a

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.

Answer 3 (a)

$iv.$ - Steadily decrease.

Minimizing the RSS (subject to the constraint $\sum_{j=1}^{p}|\beta j| \leq s$) means $LASSO$, this is the way parameters are selected.

The least squares solution will satisfy the given constraint, once s is sufficiently large. For this particular situation, the β that minimizes $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta j x_{ij})^2$ and also satisfies the given constraint will always be the least squares solution. Now until that point, the training RSS will decrease monotonically.

Q3b

Repeat (a) for test RSS.

Answer 3 (b)

$ii.$ - decrease initially, and then eventually start increasing in a U shape.

When s = 0, the only $\hat{\beta}$ that will satisfy $\sum_{j=1}^{p}|\beta j| \leq s$ will be a vector of zeros, so here we will simply have the null model ($\hat{y}$ = $\overline{y}$). Now, as s increases and the given constraint is loosened, the model's flexibility will increase. So, test RSS will therefore decrease, up to the point where it will start to overfit (and at that point, the test RSS will start increasing again).

Q3c

Repeat (a) for variance.

Answer 3 (c)

iii.
- Steadily increase.

The reason is that the given constraint region increasing in size (s increasing from zero) corresponds to λ decreasing (the shrinkage reduction), so model's flexibility is increasing and so an increase in variance will occur. If s is sufficiently large so that $\hat{\beta}$ falls within the given constraint region, the variance going forward will no longer increase, because the $\hat{\beta}$ chosen will always be the least squares estimate.

Q3d

Repeat (a) for (squared) bias.

Answer 3 (d)

$iv.$ - Steadily decrease.

The reasoning is the same as part (c) above - increasing the model's flexibility will decrease the bias. Again, this will stop reducing if the least squares solution falls within the given constraint region.

Q3e

Repeat (a) for the irreducible error.

Answer 3 (e)

$v.$ - Remain constant.

The irreducible error is the error introduced by inherent uncertainty/noise in the system being approximated. It remains constant regardless of model’s flexibility, because there may be unmeasured variables not in X that would be required to explain it, or unmeasurable variation in Y that cannot be predicted with the variables in X, regardless of how well-specified the model is (so basically, it is completely independent of s).

Q9

In this exercise, we will predict the number of applications received using the other variables in the **College** data set.

Q9a

Split the data set into a training set and a test set.

Answer 9 (a)

Randomly selected **50%** of the observations for the training set and **50%** for the test set.

```
## [1] "The training data set percent is  50"
```

```
## [1] "The test data set percent is  50"
```

Q9b

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

Answer 9 (b)

```
##
## Call:
## lm(formula = Apps ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5741.2  -479.5    15.3   359.6  7258.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.902e+02  6.381e+02  -1.238 0.216410
## PrivateYes  -3.070e+02  2.006e+02  -1.531 0.126736
## Accept       1.779e+00  5.420e-02  32.830  < 2e-16 ***
## Enroll      -1.470e+00  3.115e-01  -4.720 3.35e-06 ***
## Top10perc    6.673e+01  8.310e+00   8.030 1.31e-14 ***
## Top25perc   -2.231e+01  6.533e+00  -3.415 0.000708 ***
## F.Undergrad  9.269e-02  5.529e-02   1.676 0.094538 .
## P.Undergrad  9.397e-03  5.493e-02   0.171 0.864275
## Outstate    -1.084e-01  2.700e-02  -4.014 7.22e-05 ***
## Room.Board   2.115e-01  7.224e-02   2.928 0.003622 **
## Books        2.912e-01  3.985e-01   0.731 0.465399
## Personal     6.133e-03  8.803e-02   0.070 0.944497
## PhD         -1.548e+01  6.681e+00  -2.316 0.021082 *
## Terminal     6.415e+00  7.290e+00   0.880 0.379470
## S.F.Ratio    2.283e+01  2.047e+01   1.115 0.265526
## perc.alumni  1.134e+00  6.083e+00   0.186 0.852274
## Expend       4.857e-02  1.619e-02   2.999 0.002890 **
## Grad.Rate    7.490e+00  4.397e+00   1.703 0.089324 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1083 on 370 degrees of freedom
## Multiple R-squared:  0.9389, Adjusted R-squared:  0.9361
## F-statistic: 334.3 on 17 and 370 DF,  p-value: < 2.2e-16
```

```
## [1] "The MSE as the test error metric is  1135758"
```

Q9c

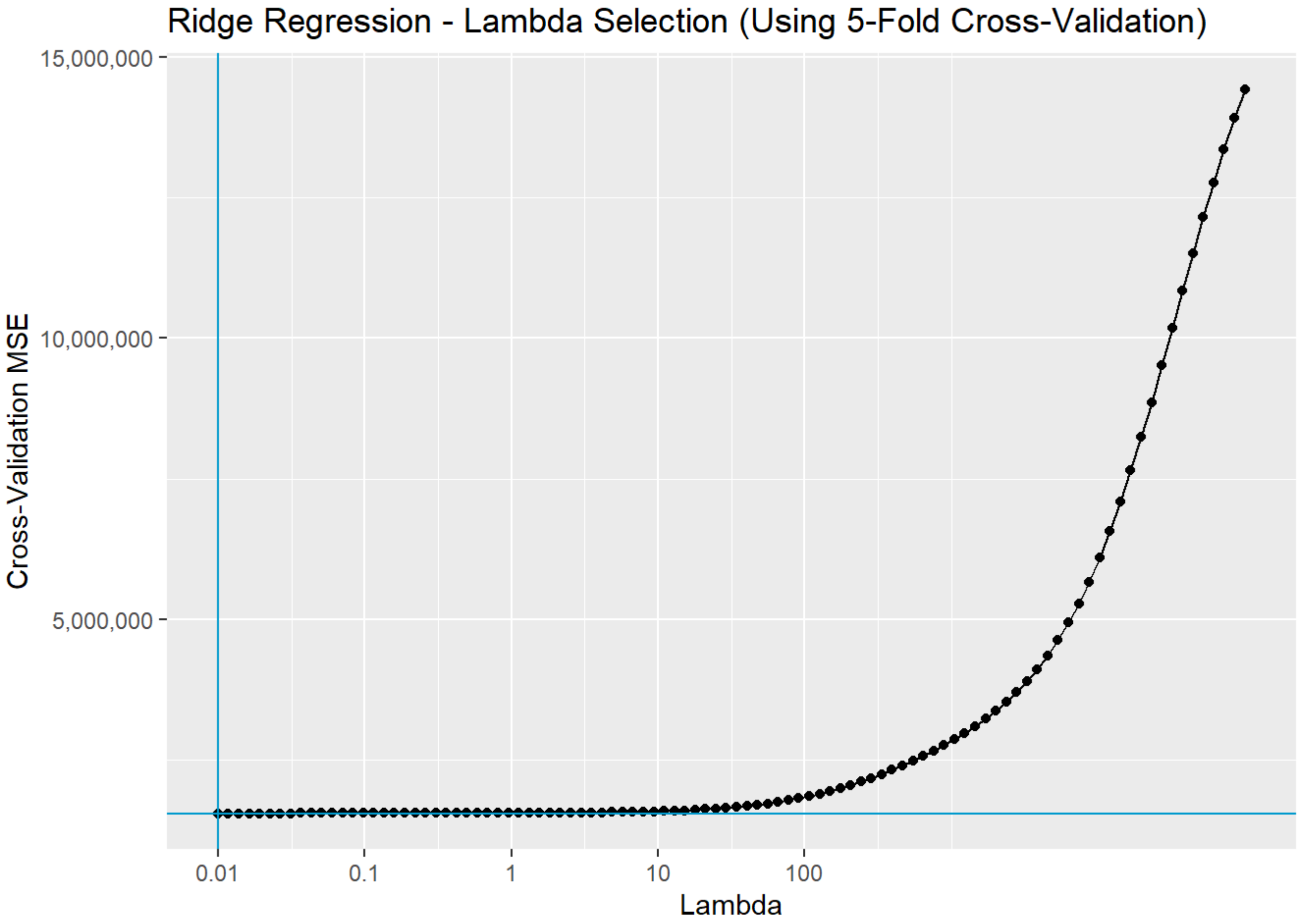
Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

Answer 9 (c)

I first create *train.matrix* and *test.matrix*, which are train & test datasets.

```
## [1] 0.01
```

I am here testing varying values of λ (from 0.01 to 100) using $5-fold$ cross-validation.



```
## [1] "The Ridge test MSE is  1135617"
```