Hw1

Runcheng Liu 2018010316

1.1

For SS_{tot} ,

$$\begin{aligned} SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 &= \begin{pmatrix} y_1 - \bar{y} & \cdots & y_n - \bar{y} \end{pmatrix} \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \\ &= \left(y^\intercal - \bar{y}\mathbf{1}^\intercal\right)(y - \bar{y}\mathbf{1}) \\ &= y^\intercal \left(\mathbf{I}_n - \mathbf{1}\mathbf{1}^\intercal/\mathbf{n}\right) y \end{aligned}$$

For SS_{reg} ,

$$\begin{aligned} SS_{reg} = \sum_{i=1}^{n} \left(\hat{y_i} - \bar{y}\right)^2 &= \begin{pmatrix} \hat{y_1} - \bar{y} & \cdots & \hat{y_n} - \bar{y} \end{pmatrix} \begin{pmatrix} \hat{y_1} - \bar{y} \\ \vdots \\ \hat{y_n} - \bar{y} \end{pmatrix} \\ &= \hat{y}^\intercal \left(\mathbf{I}_n - \mathbf{1}\mathbf{1}^\intercal\right) \hat{y} \\ &= (Py)^\intercal \left(\mathbf{I}_n - \mathbf{1}\mathbf{1}^\intercal\right)(Py) \\ &= y^\intercal P^\intercal (\mathbf{I}_n - \mathbf{1}\mathbf{1}^\intercal/\mathbf{n}) Py \end{aligned}$$

For SS_{res} ,

$$\begin{aligned} SS_{res} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 &= \begin{pmatrix} y_1 - \hat{y}_1 & \cdots & y_n - \hat{y}_n \end{pmatrix} \begin{pmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix} \\ &= \left(y^\intercal - \hat{y}^\intercal\right)\left(y - \hat{y}\right) \\ &= y^\intercal \left(\mathbf{I}_n - P\right) y \end{aligned}$$

Use the above three fomulas, thus:

$$\begin{aligned} SS_{reg} + SS_{res} &= y^\intercal P^\intercal (\mathbf{I}_n - \mathbf{1}\mathbf{1}^\intercal/\mathbf{n}) Py + y^\intercal \left(\mathbf{I}_n - P\right) y \\ &= y^\intercal \left(P^\intercal P - P^\intercal \mathbf{1}\mathbf{1}^\intercal/\mathbf{n} P + \mathbf{I}_n - P\right) y \\ &= y^\intercal \left(P - (P\mathbf{1})^\intercal (P\mathbf{1})/n + \mathbf{I}_n - P\right) y \\ &= y^\intercal \left(\mathbf{I}_n - \mathbf{1}\mathbf{1}^\intercal/n\right) y \\ &= SS_{tot} \end{aligned}$$

1.2

We know,

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

TSS measures the total variance in the response Y , and can be squares thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSS − RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X. An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, or the inherent error σ^2 is high, or both.

1.3

First, we know that,

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2\left(\left(X^\intercal X\right)^{-1}\right)_{jj}$$

Now let $r = X^\intercal X$, and without losing generality, we reorder the columns of X to set the first column to be X_j,

$$r^{-1} = \begin{bmatrix} r_{j,j} & r_{j,-j} \\ r_{-j,j} & r_{-j,-j} \end{bmatrix}^{-1}$$
$$r_{j,j} = X_j^T X_j, r_{j,-j} = X_j^T X_{-j}, r_{-j,j} = X_{-j}^T X_j, r_{-j,-j} = X_{-j}^T X_{-j}$$

By using Schur complement, the element in the first row and first column in r^{-1} is,

$$r_{1,1}^{-1} = \left[r_{j,j} - r_{j,-j} r_{-j,-j}^{-1} r_{-j,j}\right]^{-1}$$

Then we have,

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2\left[\left(X^T X\right)^{-1}\right]_{jj} = \sigma^2 r_{1,1}^{-1}$$
$$= \sigma^2\left[X_j^T X_j - X_j^T X_{-j}\left(X_{-j}^T X_{-j}\right)^{-1} X_{-j}^T X_j\right]^{-1}$$

And we have,

$$RSS_j = \sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2 = X_j^\intercal\left(\mathbf{I}_n - P\right)X_j = X_i^\intercal X_j - X_j^\intercal P X_j = X_j^\intercal X_j - X_j^\intercal X_j - X_j^\intercal X_{-j}\left(X_{-j}^\intercal X_{-j}\right)^{-1} X_j^\intercal X_j$$

Put this into the last formula, so,

$$\text{Var}\left(\hat{\beta}_j\right) = s^2\left[X_j^T X_j - X_j^T X_{-j}\left(X_{-j}^T X_{-j}\right)^{-1} X_{-j}^T X_j\right]^{-1}$$
$$= \sigma^2 \frac{1}{RSS_j}$$
$$= \frac{\sigma^2}{(n-1)S_{x_j}^2} \cdot \frac{1}{1 - R_j^2}$$

2

Prove $\hat{\beta}_0$ is BLUE is equivalent as proving $\hat{\beta}$ is BLUE. First, prove $\hat{\beta}$ is unbiased, use definition and the law of total expectation, we have,

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}\right] &= \mathrm{E}\left[\left(X^\intercal X\right)^{-1} X^\intercal y\right] \\
&= \mathrm{E}\left[\left(X^\intercal X\right)^{-1} X^\intercal \left(X\beta + \varepsilon\right)\right] \\
&= \beta + \mathrm{E}\left[\left(X^\intercal X\right)^{-1} X^\intercal \varepsilon\right] \\
&= \beta + \mathrm{E}\left[\mathrm{E}\left[\left(X^\intercal X\right)^{-1} X^\intercal \varepsilon | X\right]\right] \\
&= \beta + \mathrm{E}\left[\left(X^\intercal X\right)^{-1} X^\intercal \,\mathrm{E}[\varepsilon | X]\right] \\
&= \beta
\end{aligned}$$

Where $\mathrm{E}[\epsilon | X] = 0$ by the assumption of the model. So $\hat{\beta}$ is unbiased.

Then prove $\hat{\beta}$'s variance is minimal, let $\tilde{\beta} = Cy$ be another linear estimator of β with $C = (X^\intercal X)^{-1} X^\intercal + D$ where D is a $K \times n$ non-zero matrix. As we're restricting to unbiased estimators, minimum mean squared error implies minimum variance. The goal is therefore to show that such an estimator has a variance no smaller than that of $\hat{\beta}$, the OLS estimator. We calculate,

$$\begin{aligned}
\mathrm{E}[\tilde{\beta}] &= \mathrm{E}[Cy] \\
&= \mathrm{E}\left[\left(\left(X'X\right)^{-1} X' + D\right)\left(X\beta + \varepsilon\right)\right] \\
&= \left(\left(X'X\right)^{-1} X' + D\right) X\beta + \left(\left(X'X\right)^{-1} X' + D\right) \mathrm{E}[\varepsilon] \\
&\overset{\mathrm{E}[\varepsilon]=0}{=} \left(\left(X'X\right)^{-1} X' + D\right) X\beta \\
&= \left(X'X\right)^{-1} X'X\beta + DX\beta \\
&= \left(I_K + DX\right)\beta
\end{aligned}$$

Therefore, since β is unobservable, and $\tilde{\beta}$ is unbiased if and only if $DX = 0$. Then,

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}(Cy) \\
&= C\,\mathrm{Var}(y)C^\intercal \\
&= \sigma^2 CC^\intercal \\
&= \sigma^2 \left(\left(X^\intercal X\right)^{-1} X^\intercal + D\right)\left(X\left(X^\intercal X\right)^{-1} + D^\intercal\right) \\
&= \sigma^2 \left(\left(X^\intercal X\right)^{-1} X^\intercal X\left(X^\intercal X\right)^{-1} + \left(X^\intercal X\right)^{-1} X^\intercal D^\intercal + DX\left(X^\intercal X\right)^{-1} + DD^\intercal\right) \\
&= \sigma^2\left(X^\intercal X\right)^{-1} + \sigma^2\left(X^\intercal X\right)^{-1}(DX)^\intercal + \sigma^2 DX\left(X^\intercal X\right)^{-1} + \sigma^2 DD^\intercal \\
&\overset{DX=0}{=} \sigma^2\left(X^\intercal X\right)^{-1} + \sigma^2 DD^\intercal \\
&\overset{\mathrm{Var}(\hat{\beta})=\sigma^2(X^\intercal X)^{-1}}{=} \mathrm{Var}(\hat{\beta}) + \sigma^2 DD^\intercal
\end{aligned}$$

Since $DD^\intercal$ is positive semidefinitive matrix. $\mathrm{Var}(\tilde{\beta})$ is always bigger than $\mathrm{Var}(\hat{\beta})$.

3.1

First, the ordinary least squares model posits that the conditioning distribution of the response y is,

$$y \mid X, \beta \sim \mathcal{N}\left(X\beta, \sigma^2 I\right)$$

Also we know the prior distribution of β,

$$\beta \sim \mathcal{N}\left(0, \sigma^2 I/\lambda\right)$$

Thus,

$$\begin{aligned}
\hat{\beta}^{MAP} &= \arg\max_{\beta} \prod_{i=1}^{n} P(Y_i|X_i,\beta)P(\beta) \\
&= \arg\max_{\beta} \log\left(\prod_{i=1}^{n} P\left(Y_i|X_i,\beta\right)P\left(\beta\right)\right) \\
&= \arg\max_{\beta} \left(\sum_{i=1}^{n} \log P\left(Y_i|X_i,\beta\right) + \log P\left(\beta\right)\right) \\
&= \arg\max_{\beta} \left(\sum_{i=1}^{n} -\frac{\left(Y_i - X_i^{\intercal}\beta\right)^2}{2\sigma^2\mathbf{I}} - \frac{\lambda\|\beta\|_2^2,}{2\mathbf{I}\sigma^2}\right) \\
&= \arg\min_{\beta} \left(Y_i - X_i^{\intercal}\beta\right)^2 + \lambda\|\beta\|_2^2 \\
&= \arg\min_{\beta} (X\beta - Y)^T(X\beta - Y) + \lambda\|\beta\|_2^2
\end{aligned}$$

3.2

As $\lambda \to 0$, the prior distribution $P(\beta)$ is very wide, which means that β has huge variance, thus is not constrained. However, as $\lambda \to \infty$, the prior distribution $P(\beta)$ is very narrow, which means that β has small variance, thus is constrained.

4

```
#Load data
prostate <- read.csv("prostate.csv")
train = subset(prostate, train==TRUE)
test = subset(prostate, train==FALSE)


#Split data into training data and test data
train_x=model.matrix(lpsa~. ,train[,c("lcavol","lweight","age","lbph","svi","lcp","gleas
on","pgg45","lpsa")])[,-1]
train_y=train$lpsa
test_x=model.matrix(lpsa~. ,test[,c("lcavol","lweight","age","lbph","svi","lcp","gleaso
n","pgg45","lpsa")])[,-1]
test_y=test$lpsa
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.0-2
```

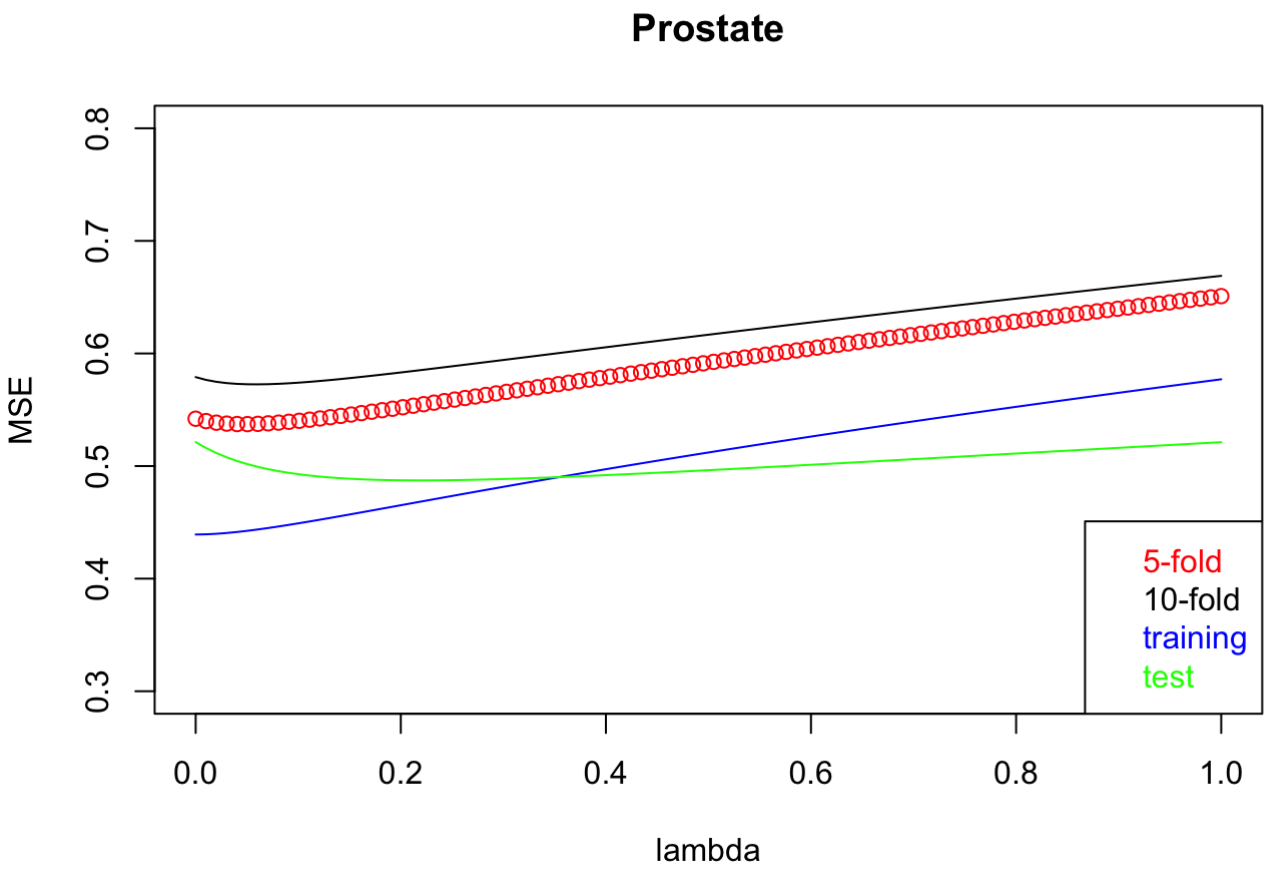
```
#Ridge regression
grid=seq(0,1, length =100)
ridge.mod=glmnet (train_x,train_y,alpha=0, lambda=grid, thresh =1e-12)


#5-fold Cross-validation
set.seed(1)
cv.out=cv.glmnet(train_x,train_y,alpha=0, nfolds = 5, lambda=grid, type.measure = "mse")
cvm_5 = rev(cv.out$cvm)

#Minimum MSE of lambda for 5-fold
bestlam_5 =cv.out$lambda.min
bestlam_5
```

```
## [1] 0.05050505
```

```
#10-fold Cross-validation
set.seed(1)
cv.out=cv.glmnet(train_x,train_y,alpha=0, nfolds = 10, lambda=grid, type.measure = "mse"
)
cvm_10 = rev(cv.out$cvm)

#Minimum MSE of lambda for 10-fold
bestlam_10 =cv.out$lambda.min
bestlam_10
```

```
## [1] 0.06060606
```

```
#Training error
train_error = c()
for (p in grid){
  ridge.pred=predict (ridge.mod ,s=p ,newx=train_x)
  train_error = c(mean((ridge.pred -train_y)^2), train_error)
}
train_error = rev(train_error)


#Test error
test_error = c()
for (p in grid){
  ridge.pred=predict (ridge.mod ,s=p ,newx=test_x)
  test_error = c(mean((ridge.pred -test_y)^2), test_error)
}
test_error = rev(test_error)

# PLot
plot(grid, cvm_5, col="red", ylim = c(0.3, 0.8), main = "Prostate", xlab = "lambda", yla
b="MSE")
lines(grid, cvm_10, col="black")
lines(grid,train_error,col="blue")
lines(grid,test_error,col="green")
legend("bottomright",legend=c("5-fold","10-fold","training","test"), text.col = c("red",
"black","blue","green"))
```



From the above results, we can see that the training error, test error, 5-fold cross-validation error, 10-fold cross-validation for training, test, 5-fold cross-validation, 10-fold cross-validation respectively with respect to λ. I find that 5-fold cross-validation error, 10-fold cross-validation error are always larger than training error, test error. I think the reason for this is that the size of test and training data are larger than size of cross-validation data, thus the error is lower. I also find out that sometimes 5-fold cross-validation error is bigger than 10-fold cross-validation error, sometimes is reverse, unless you fix the random seed. For this random seed(which controls the splits of k-fold), the optimal λ for 5-fold and 10-fold cross-validation are 0.05 and 0.06 respectively.