Reihaneh Moghisi
Assignment 4

Question 1

(a) predictor X

A:

```
set.seed(123)

X <- rnorm(100)

error <- rnorm(100)
```

(b) response vector y

A:

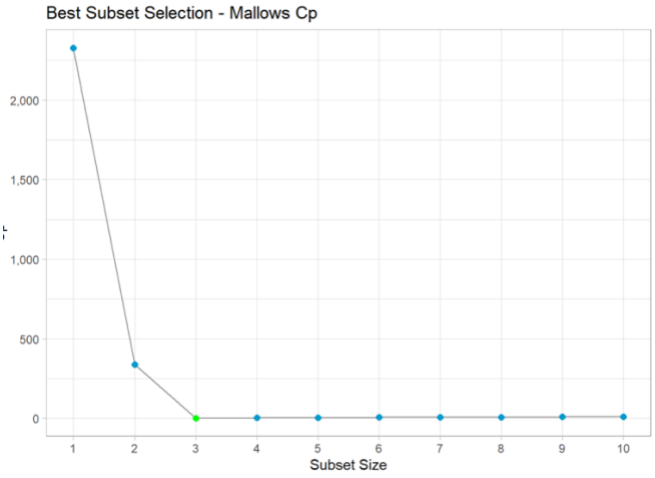
- $\beta_0=8$
- $\beta_1=4$
- $\beta_2=5$
- $\beta_3=10$

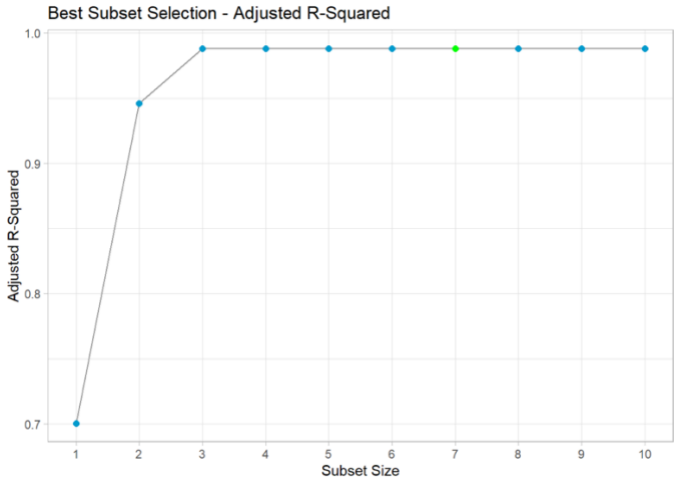
```
Y <- 10 - 5*X + 6*X^2 + 3*X^3 + error
```

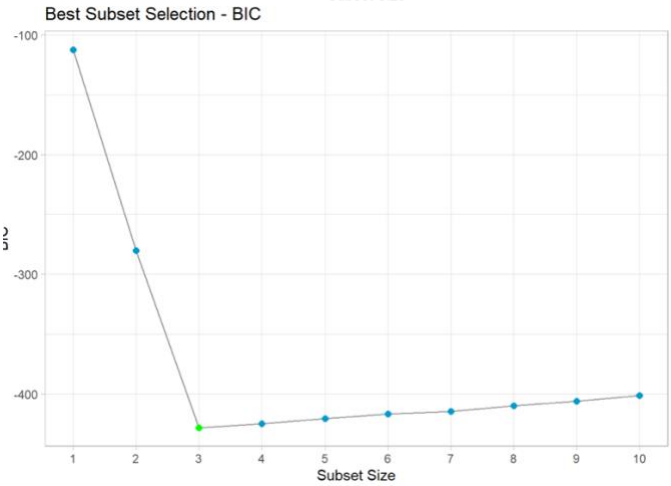
(c) Best Subset Selection

```
X_matrix <- poly(X, 10, raw = T, simple = T)


model <- regsubsets(y = Y,
                    x = X_matrix,
                    nvmax = 10,
                    method = "exhaustive")


model_summary <- summary(model)
```

Reihaneh Moghisi
Assignment 4







When using Mallows C_P & BIC as the selection criterion, the 3-variable model is chosen.

Reihaneh Moghisi
Assignment 4

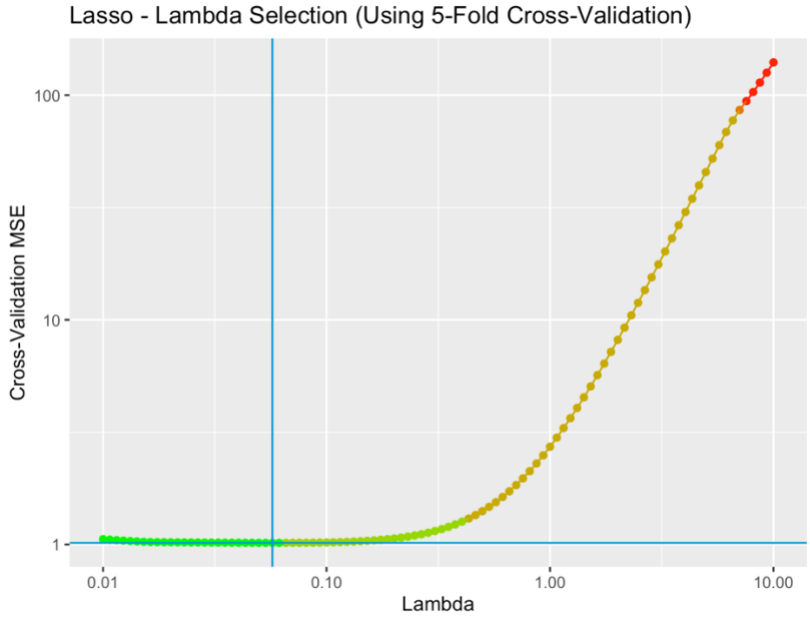
		1	2	3	4	5	6	7	8	9	10
1	(1)	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "
2	(1)	" "	" "	"*"	"*"	" "	" "	" "	" "	" "	" "	" "	" "
3	(1)	"*"	"*"	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "
4	(1)	"*"	"*"	"*"	" "	" "	" "	" "	"*"	" "	" "	" "	" "
5	(1)	"*"	"*"	"*"	" "	" "	" "	" "	" "	"*"	" "	" "	" "
6	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	" "	" "	"*"	" "	" "
7	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	" "	" "	"*"	" "	" "
8	(1)	"*"	"*"	"*"	"*"	"*"	"*"	" "	" "	" "	"*"	" "	" "
9	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	" "	" "	"*"
10	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

Using Adjusted R^2, the best-subset algorithm selects the 7-variable model, with the following coefficients:

(Intercept)	1	2	3	4	6	
7.7931750	3.8907484	7.2758848	10.0530514	-4.0243410	2.2138377	
8	10					
-0.4870719	0.0373538					

(d) Lasso selection

Here is a graph showing the selection. Both the x and y axis are on a $\log_{10}$ scale, as both λ and the cross-validation MSE are over several orders of magnitude.



We can see the coefficients of X to the power of 5 to 10 is exactly zero. The coefficient for X4 is close to zero. The coefficients of X 1,2 and 3 are substantially larger due to the fact we chose the order of 3 polynomial.

Reihaneh Moghisi
Assignment 4

```
11 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 8.11840765
1           3.83641384
2           4.66762787
3           9.99811122
4           0.03330552
5           .
6           .
7           .
8           .
9           .
10          .
```

Question 2:

a)Split data

```
set.seed(3)
train_index <- sample(1:nrow(College), round(nrow(College) * 0.7))

train <- College[train_index, ]
nrow(train) / nrow(College)
## [1] 0.7001287
```

The remaining observations are allocated to the `test` dataset:

```
test <- College[-train_index, ]
nrow(test) / nrow(College)
```

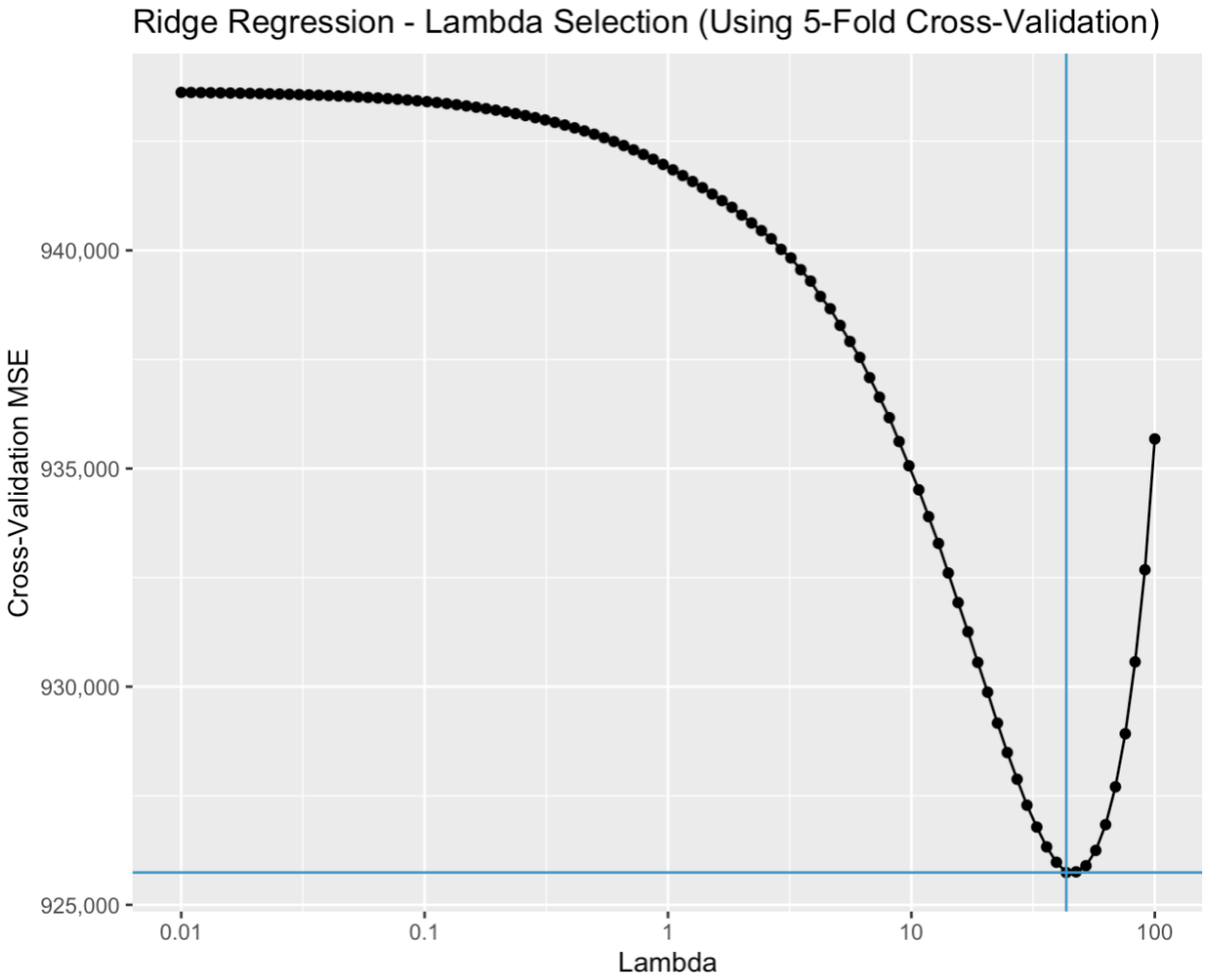
(b) OLS regression

Reihaneh Moghisi
Assignment 4

```
Call:
lm(formula = Apps ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3149.1  -366.6   -42.3   296.6  5657.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -625.31793  418.96146  -1.493 0.136156
PrivateYes  -531.00812  149.13545  -3.561 0.000404 ***
Accept         1.31888    0.05618  23.476  < 2e-16 ***
Enroll        -0.43546    0.19874  -2.191 0.028885 *
Top10perc     50.93958    5.74902   8.861  < 2e-16 ***
Top25perc    -12.40222    4.52484  -2.741 0.006335 **
F.Undergrad    0.07675    0.03237   2.371 0.018120 *
P.Undergrad    0.03967    0.03172   1.251 0.211625
Outstate      -0.04319    0.01936  -2.231 0.026096 *
Room.Board     0.21327    0.04889   4.362 1.55e-05 ***
Books          0.14681    0.23792   0.617 0.537454
Personal       0.02492    0.07000   0.356 0.721965
PhD           -8.50811    4.95017  -1.719 0.086248 .
Terminal      -1.80323    5.56078  -0.324 0.745858
S.F.Ratio     11.20908   13.00766   0.862 0.389229
perc.alumni   -6.99057    4.34316  -1.610 0.108094
Expend         0.04041    0.01221   3.310 0.000998 ***
Grad.Rate      5.98245    3.01090   1.987 0.047448 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 899.6 on 526 degrees of freedom
Multiple R-squared:  0.9278,    Adjusted R-squared:  0.9255
F-statistic: 397.9 on 17 and 526 DF,  p-value: < 2.2e-16
```

Reihaneh Moghisi
Assignment 4

(c) Ridge regression

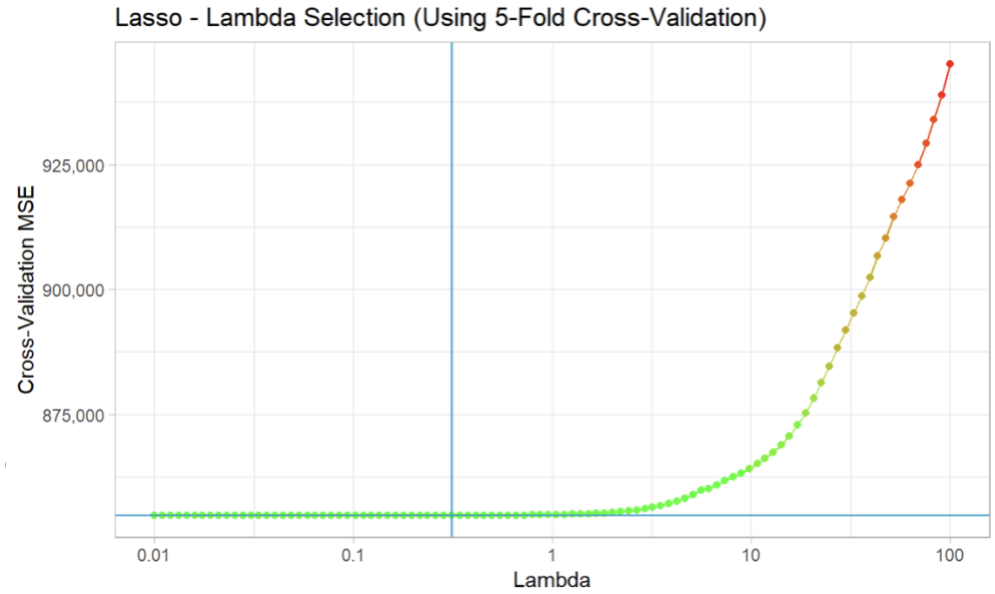


The selected value of λ is **43.2876** when chosen to generate predictions for this lambda the resulting MSE is :

```
## [1] 2297697
```

(d) Lasso

Reihaneh Moghisi
Assignment 4



The selected value of λ is **0.3126**.
Fitting the full model, and evaluating the `test` MSE:

```
## [1] 2025346
```

The coefficients:

Code

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) -1162.486
## Private.No    532.620
## Private.Yes     .
## Accept          1.314
## Enroll         -0.402
## Top10perc      50.654
## Top25perc     -12.190
## F.Undergrad     0.073
## P.Undergrad     0.039
## Outstate       -0.043
## Room.Board      0.214
## Books           0.145
## Personal        0.024
```

```
## PhD            -8.501
## Terminal       -1.760
## S.F.Ratio      11.042
## perc.alumni    -7.051
## Expend          0.040
## Grad.Rate       5.934
```