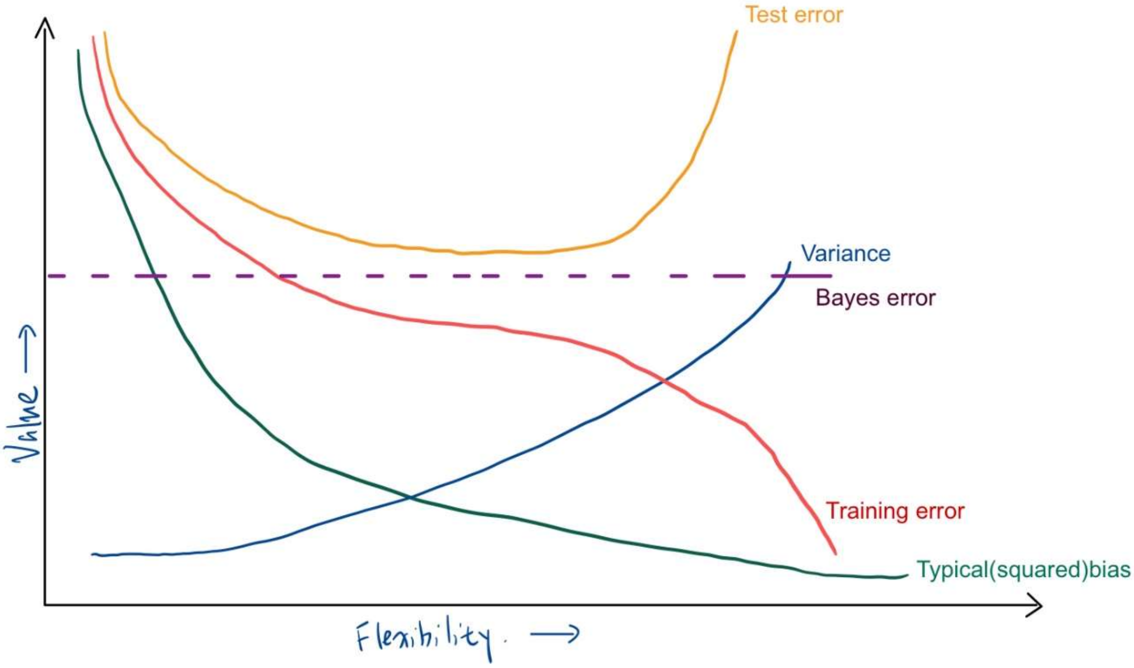
ECE 625 Assignment_1

Q1

(a).



(b)

Training error: The training error decreases monotonically as the flexibility of the model increases and the function over-fits the data, because as the flexibility increases, the curve fits the observed data more closely, and in the end the error can be treated as a constant.

Typical bias: The bias represents the difference between the predicted and true value, and as the flexibility increases creating a more complex function that better reflects the true problem, the difference between the predicted and true values is smaller.

Variance: The variance reflects the amount of change in the value of the function caused by the use of different training data sets, with the increase in the degree of fit, any point of change may lead to a large change in the value, so the variance will be monotonically increasing.

Test error: The test error gradually decreases as the flexibility increases and at some point will equalize, however, as the fit to the data increases and the model is over-fitted, the test error will begin to increase.

Bayes error: The bayes error is calculated from the training data and is a constant that does not change due to the flexibility of the model.

Q2

For a very flexible approach, the advantages of it is the bias is small, and it can fit the data better for a non-linear model. The disadvantages of it is it need a large number of parameters, it is easy to overfit and the variance is large.

When we have high requirements for the interpretability of the model, less flexible approach is better than more flexible approach.

When we are more interested in the accuracy of the prediction, more flexible approach is preferred.

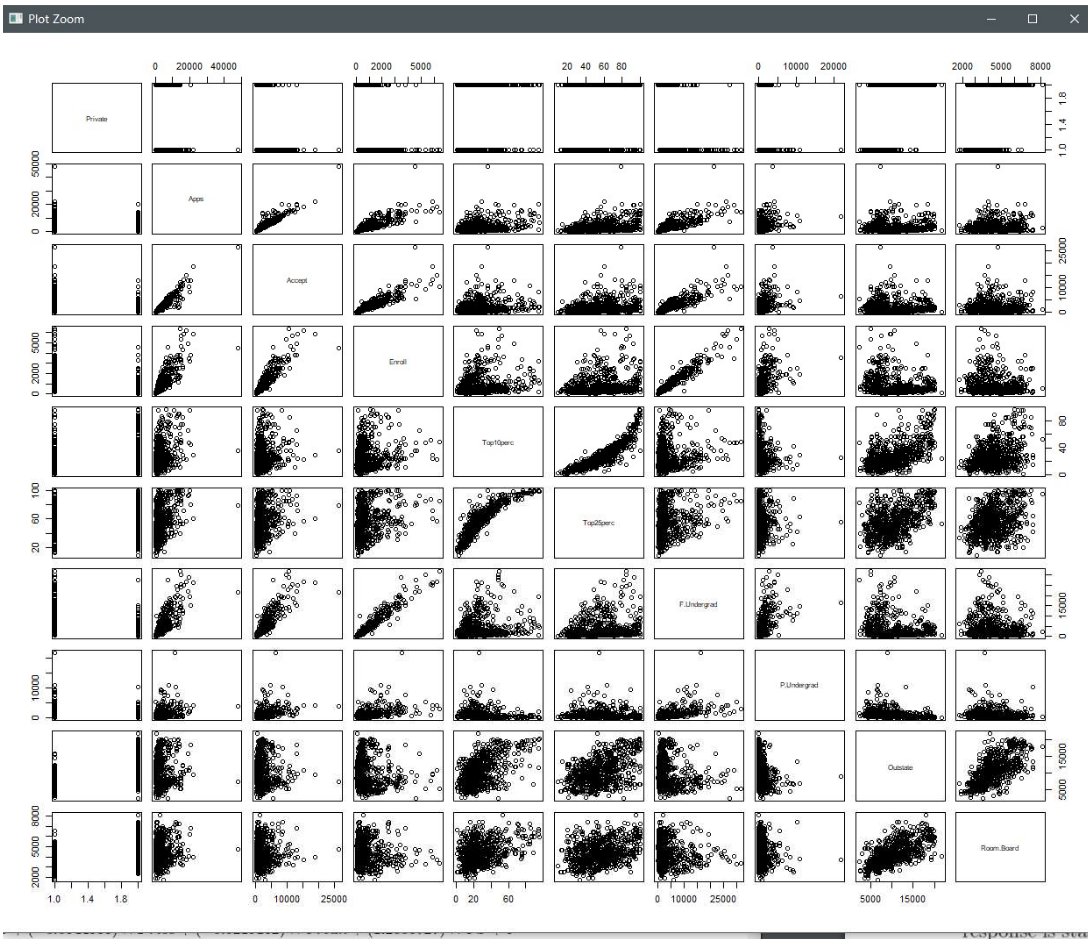
Q3

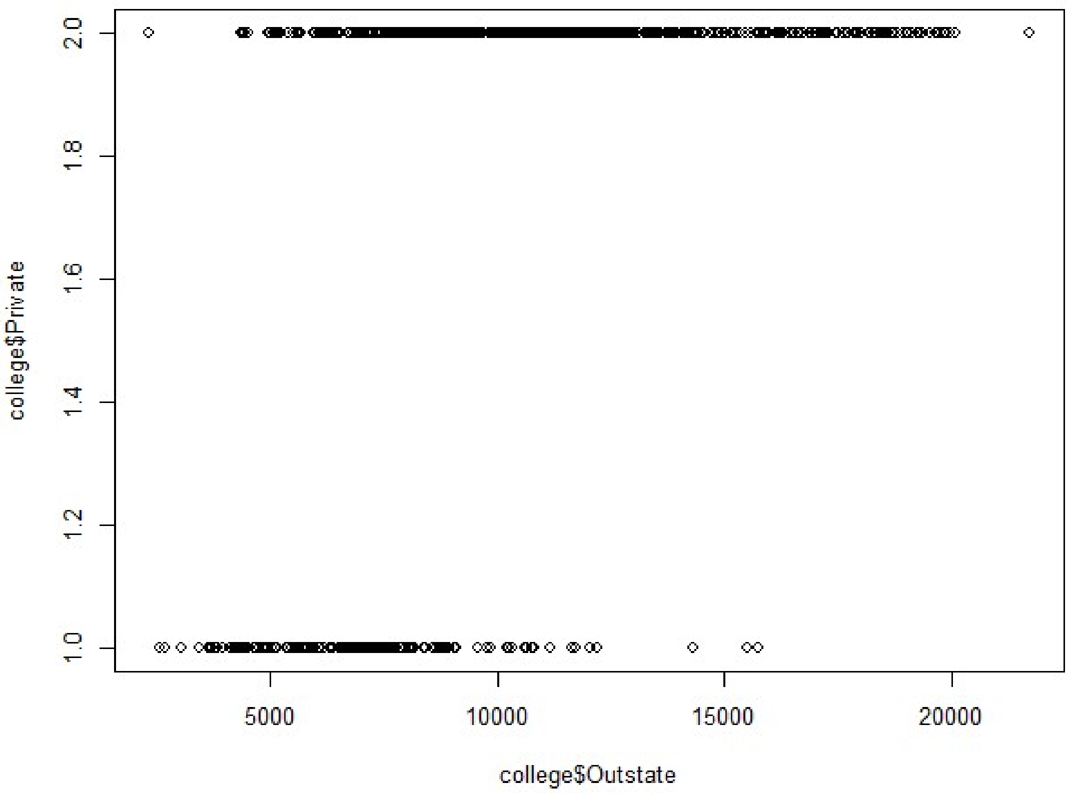
	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885
2	Adelphi University	Yes	2186	1924	512	16	29	2683
3	Adrian College	Yes	1428	1097	336	22	50	1036
4	Agnes Scott College	Yes	417	349	137	60	89	510
5	Alaska Pacific University	Yes	193	146	55	16	44	249
6	Albertson College	Yes	587	479	158	38	62	678
7	Albertus Magnus College	Yes	353	340	103	17	45	416
8	Albion College	Yes	1899	1720	489	37	68	1594
9	Albright College	Yes	1038	839	227	30	63	973
10	Alderson-Broaddus College	Yes	582	498	172	21	44	799
11	Alfred University	Yes	1732	1425	472	37	75	1830
12	Allegheny College	Yes	2652	1900	484	44	77	1707
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130
14	Alma College	Yes	1267	1080	385	44	73	1306
15	Alverno College	Yes	494	313	157	23	46	1317
16	American International College	Yes	1420	1093	220	9	22	1018
17	Amherst College	Yes	4302	992	418	83	96	1593
18	Anderson University	Yes	1216	908	423	19	40	1819
19	Andrews University	Yes	1130	704	322	14	23	1586

	row.names
1	Abilene Christian University
2	Adelphi University
3	Adrian College
4	Agnes Scott College
5	Alaska Pacific University
6	Albertson College
7	Albertus Magnus College
8	Albion College
9	Albright College
10	Alderson-Broaddus College
11	Alfred University
12	Allegheny College
13	Allentown Coll. of St. Francis de Sales
14	Alma College
15	Alverno College
16	American International College
17	Amherst College
18	Anderson University
19	Andrews University

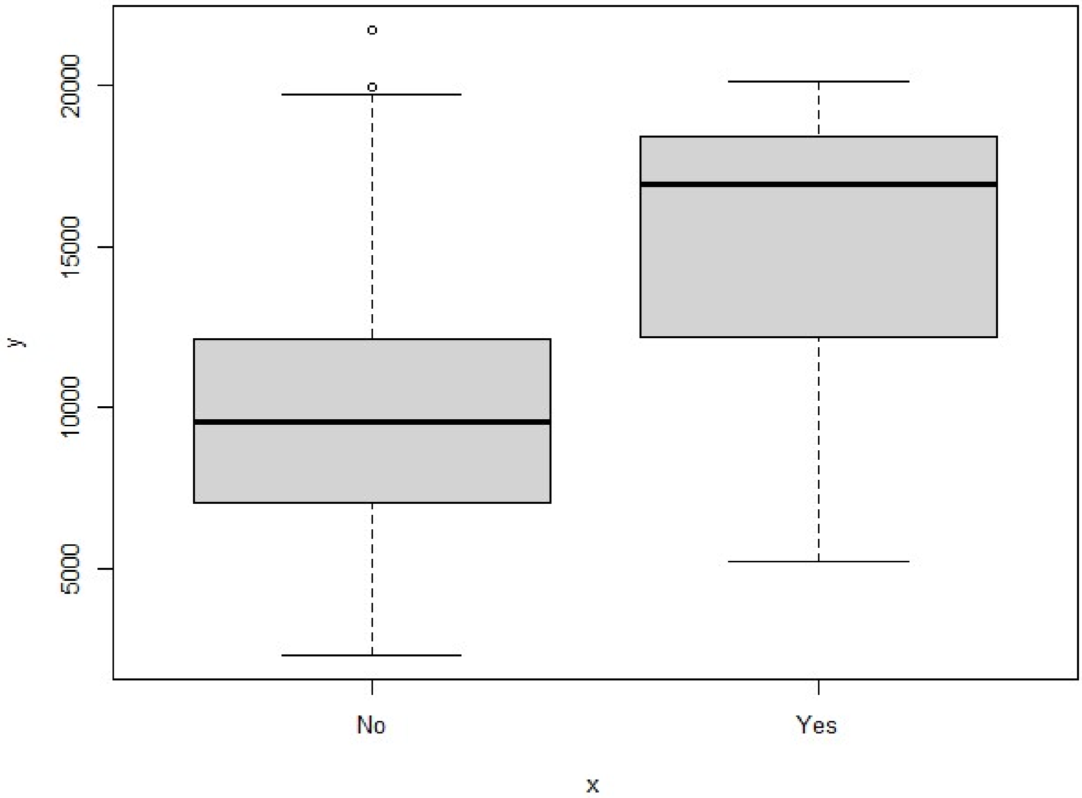
	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc
1	Abilene Christian University	Yes	1660	1232	721	23	52
2	Adelphi University	Yes	2186	1924	512	16	29
3	Adrian College	Yes	1428	1097	336	22	50
4	Agnes Scott College	Yes	417	349	137	60	89
5	Alaska Pacific University	Yes	193	146	55	16	44
6	Albertson College	Yes	587	479	158	38	62
7	Albertus Magnus College	Yes	353	340	103	17	45
8	Albion College	Yes	1899	1720	489	37	68
9	Albright College	Yes	1038	839	227	30	63
10	Alderson-Broaddus College	Yes	582	498	172	21	44
11	Alfred University	Yes	1732	1425	472	37	75
12	Allegheny College	Yes	2652	1900	484	44	77
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64
14	Alma College	Yes	1267	1080	385	44	73
15	Alverno College	Yes	494	313	157	23	46
16	American International College	Yes	1420	1093	220	9	22
17	Amherst College	Yes	4302	992	418	83	96
18	Anderson University	Yes	1216	908	423	19	40
19	Andrews University	Yes	1130	704	322	14	23

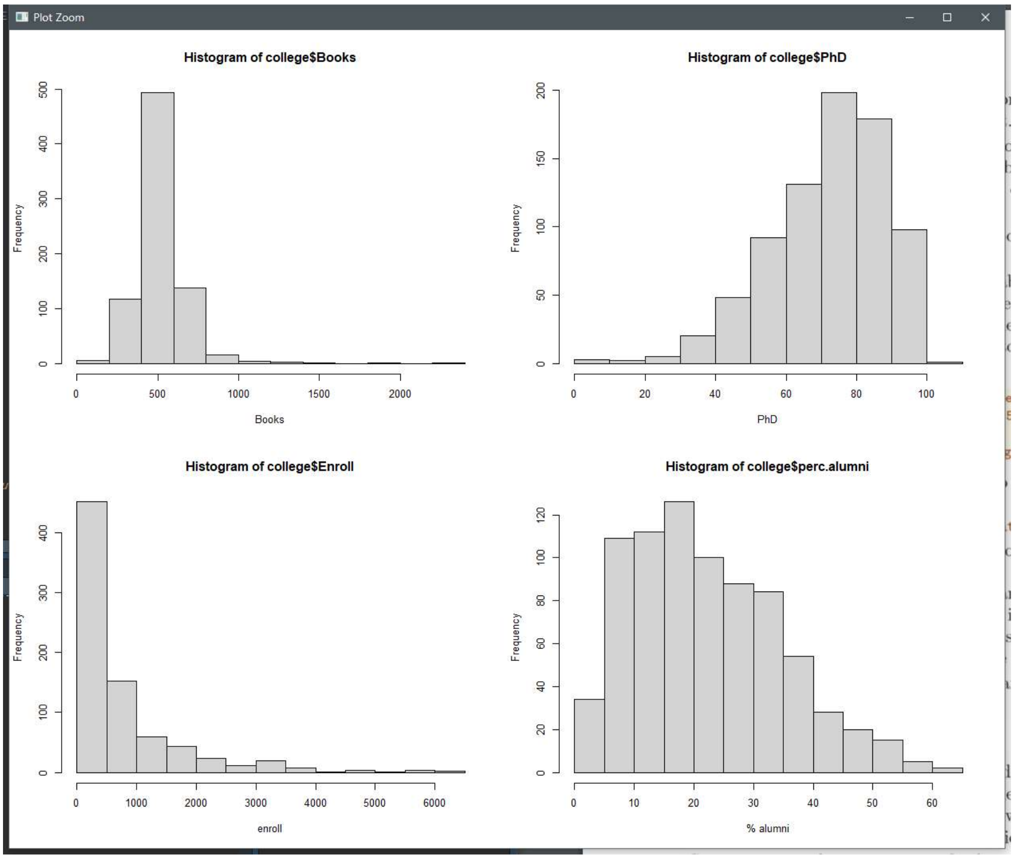
```
 Private        Apps           Accept          Enroll        Top10perc    
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00  
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00  
           Median : 1558   Median : 1110   Median : 434   Median :23.00  
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56  
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00  
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00  
   Top25perc      F.Undergrad     P.Undergrad         Outstate    
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340  
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320  
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990  
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441  
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925  
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700  
   Room.Board       Books           Personal         PhD        
 Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00  
 1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00  
 Median :4200   Median : 500.0   Median :1200   Median : 75.00  
 Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66  
 3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00  
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00  
    Terminal       S.F.Ratio      perc.alumni        Expend     
 Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186  
 1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751  
 Median : 82.0   Median :13.60   Median :21.00   Median : 8377  
 Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660  
 3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830  
 Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233  
   Grad.Rate     
 Min.   : 10.00  
 1st Qu.: 53.00  
 Median : 65.00  
 Mean   : 65.46  
 3rd Qu.: 78.00  
 Max.   :118.00  
```


Plot Zoom
Private
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board


college$Private
2.0
1.8
1.6
1.4
1.2
1.0
5000
10000
15000
20000
college$Outstate

```
> summary(college$Elite)
 No Yes
699  78
```





Q4

(a)

iii. is correct.

Based on the given condition we can derive:

$$Y = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \text{ x } IQ - 10GPA \text{ x } Gender$$

For female we can get:

$$Y_F = 85 + 10GPA + 0.07IQ + 0.01GPA \text{ x } IQ$$

For male we can get:

$$Y_M = 50 + 20GPA + 0.07IQ + 0.01GPA \text{ x } IQ$$

For fixed value of GPA and IQ without specific value, we can not get result that female or male is better. So i and ii are incorrect. For answer iii, 50 + 20GPA > 85 + 10 GPA we can get solution that GPA > 3.5, so if GPA is high enough, males can earn more than females. So answer iii is correct.

(b)

85 + 10 x 4 + 0.07 x 110 + 0.01 x 4 x 110 = 137.1 thousands of dollar = 137100$

(c)

False

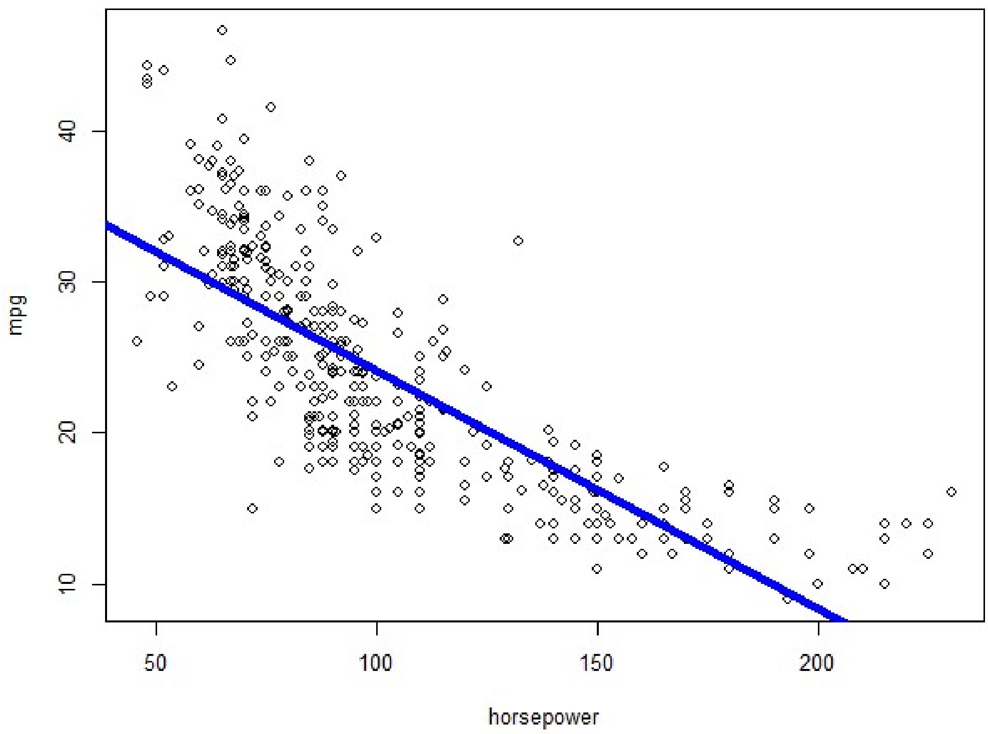
The small coefficient does not indicate the less infect, the degree of influence of the coefficients on the results needs to be judged according to the p-value, so we need to check the p-value of the interaction.

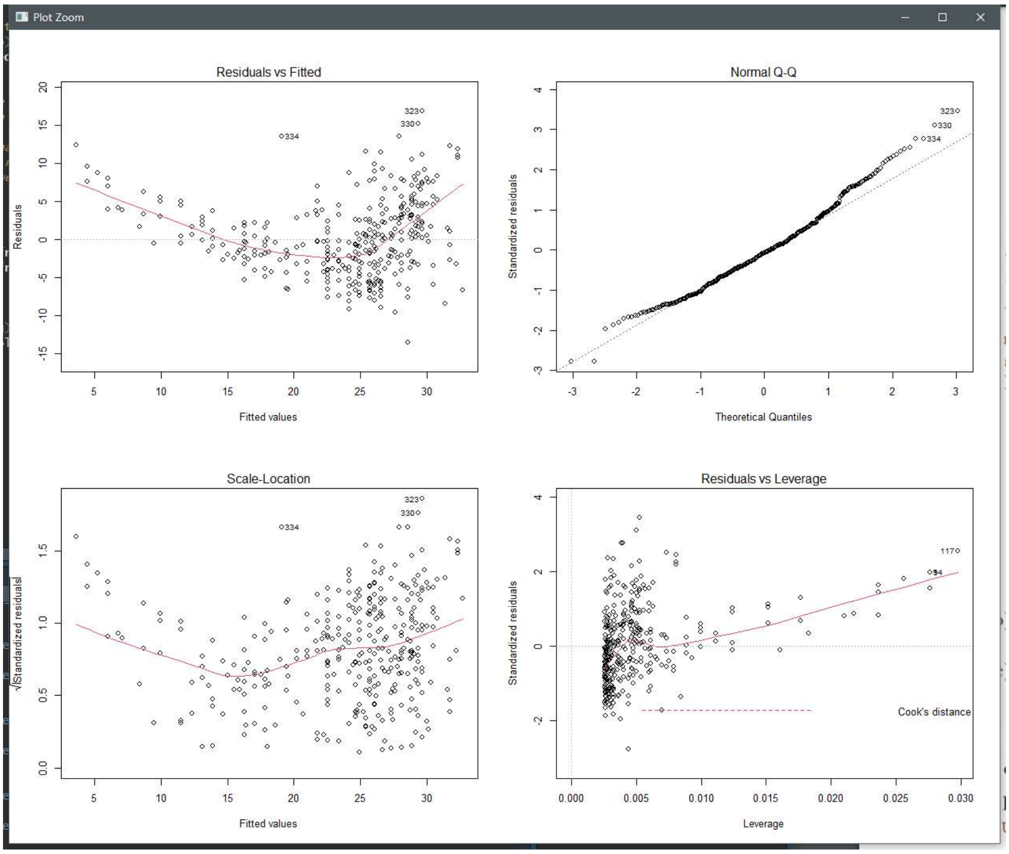
Q5

(a)

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5710  -3.2592  -0.3435   2.7630  16.9240 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
> predict(fit, data.frame(horsepower = 98), interval = 'confidence')
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = 'prediction')
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```





Q6

(a)

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b)

Price: when price increase 1$, sales will decrease 0.05449 unit sales, all other predictors are fixed.

Urban: The average unit sales in urban area was 21.9161 units less than that in rural areas, all other predictors are fixed.

US: Average unit sales in a United State store are 1200.573 units more than in a non-US store, all other predictors are fixed.

(c)

$$Sales = 13.043469 - 0.054459\ \times Price - 0.021916\ \times Urban + 1.200573\ \times US +\ \varepsilon$$

(d)

Price and US

(e)

```
Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2354 
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

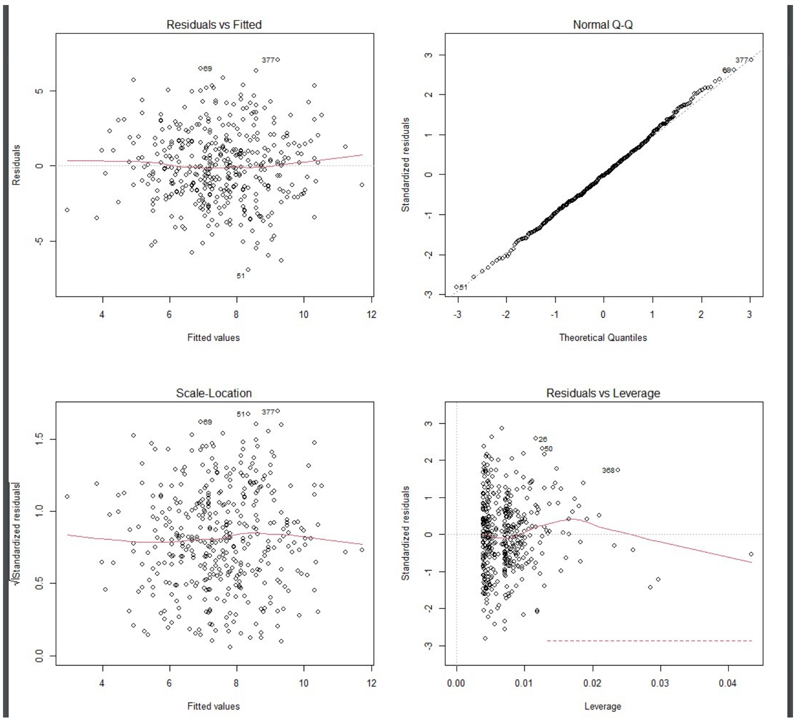
(f)

We can see the R-square values, the performance of both models was modest, only 23.93% change in response explained.

(g)

```
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

(f)



We can find few outliers in the Normal Q-Q figure, and some leverage points in Residuals vs Leverage figure.

Q7

(a)

$$\text{Min}: (y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 + y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

(b)

$$\frac{\partial}{\partial \hat{\beta}_1} = 2\hat{\beta}_1 x_1^2 - 2x_1 y_1 + 2\hat{\beta}_2 x_1^2 + 2\hat{\beta}_1 x_2^2 - 2x_2 y_2 + 2\hat{\beta}_2 x_2^2 + 2\lambda\hat{\beta}_1 = 0$$

$$\hat{\beta}_1 x_1^2 + \hat{\beta}_2 x_1^2 + \hat{\beta}_1 x_2^2 + \hat{\beta}_2 x_2^2 + 2\lambda\beta_1 = x_1 y_1 + x_2 y_2$$

$$\hat{\beta}_1 (x_1^2 + x_2^2 + \lambda) + \hat{\beta}_2 (x_1^2 + x_2^2) = x_1 y_1 + x_2 y_2 .$$

$$\frac{\partial}{\partial \hat{\beta}_2} \to \hat{\beta}_1 (x_1^2 + x_2^2) + \hat{\beta}_2 (x_1^2 + x_2^2 + \lambda) = x_1 y_1 + x_2 y_2$$

$$\hat{\beta}_1 \lambda = \hat{\beta}_2 \lambda \Rightarrow \hat{\beta}_1 = \hat{\beta}_2 .$$

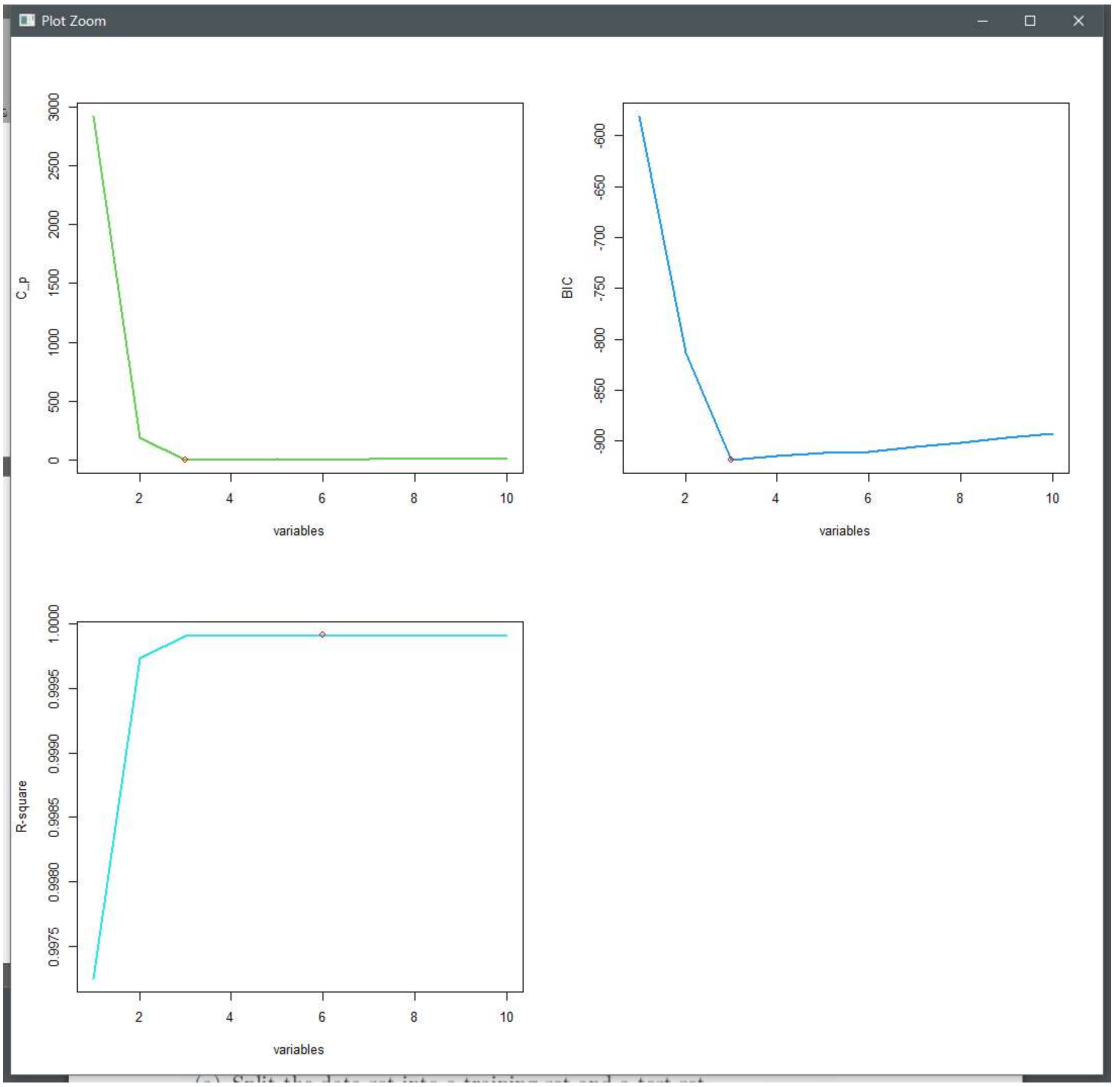
(c)

$$(y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)^2 + (y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$$

(d)

Q8

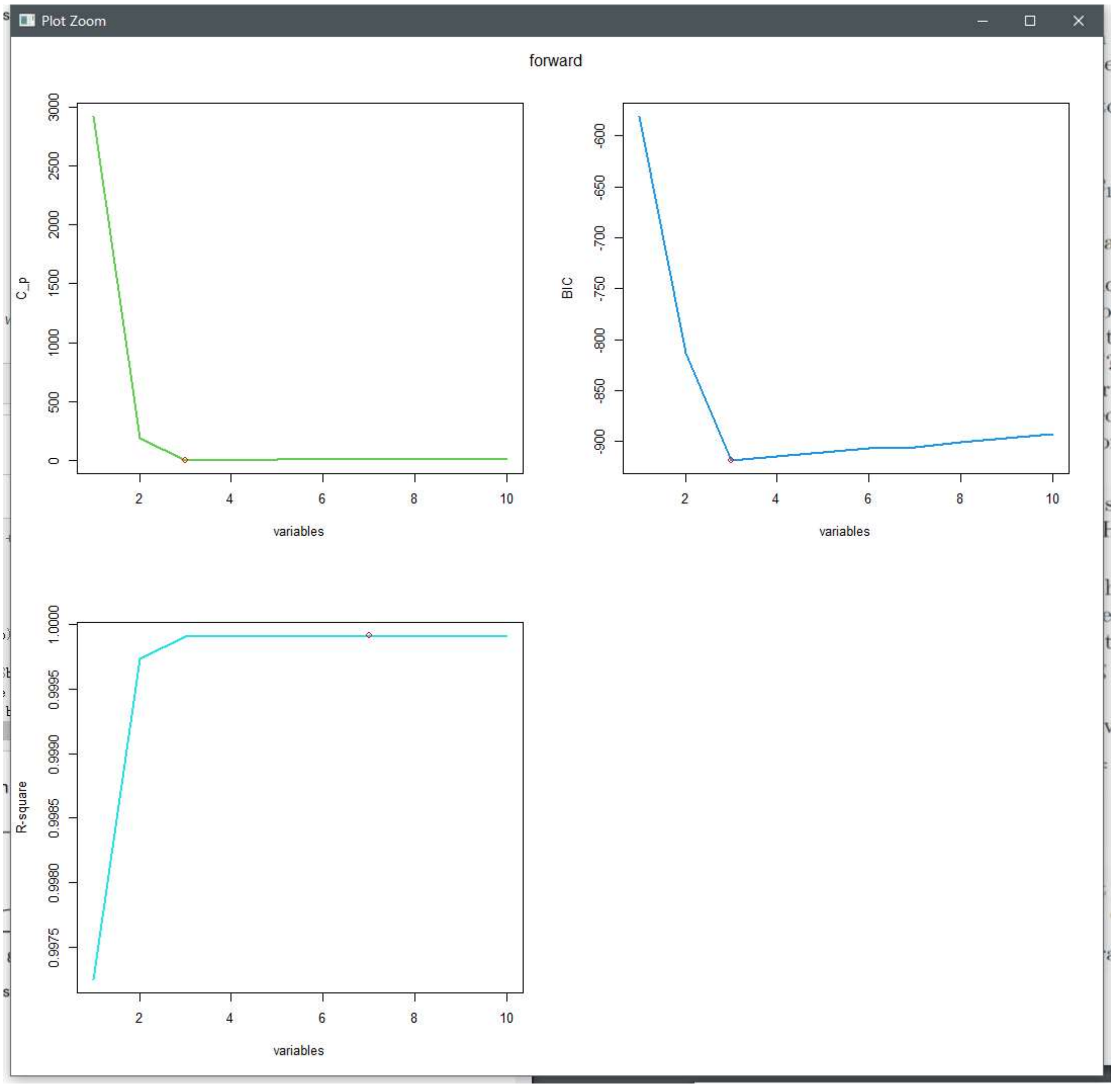
(abc)




```
(Intercept)           x      I(x^2)      I(x^3)
  10.985970    9.270659   -1.110683   16.917415
> coef(fit, which.min(fit_sum$cp))
(Intercept)           x      I(x^2)      I(x^3)
  10.985970    9.270659   -1.110683   16.917415
> coef(fit, which.max(fit_sum$adjr2))
  (Intercept)             x         I(x^2)         I(x^3)         I(x^5)
11.0756208964  8.8183396623 -1.2659454518 17.6960852006 -0.2967115416
       I(x^7)        I(x^10)
 0.0303448847  0.0002644141
```

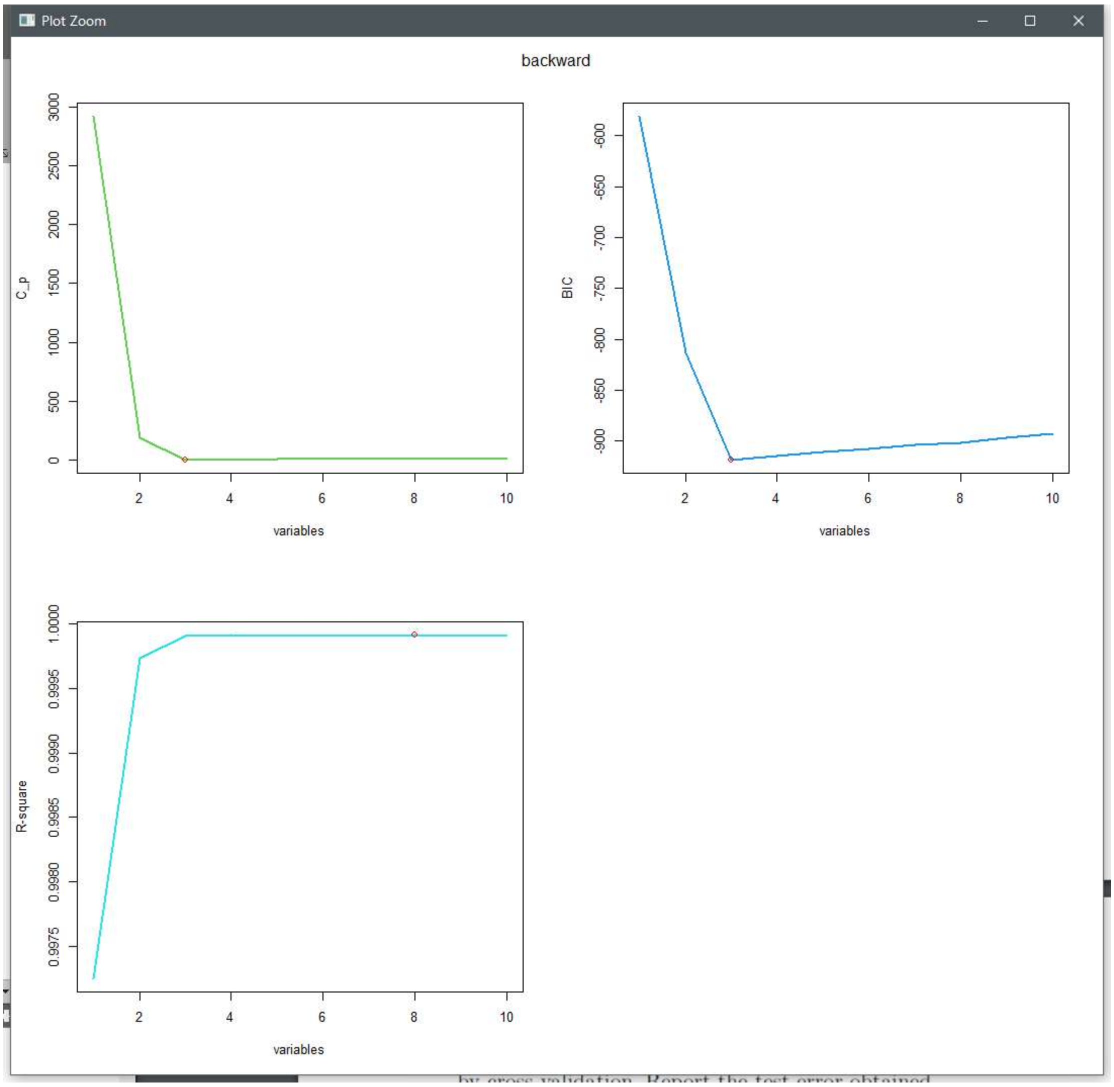
Cp and BIC choose three variables, R-square choose six variables

(d)




```
(Intercept)           x      I(x^2)      I(x^3)
  10.985970    9.270659   -1.110683   16.917415
> coef(fit, which.min(fit_sum$cp))
(Intercept)           x      I(x^2)      I(x^3)
  10.985970    9.270659   -1.110683   16.917415
> coef(fit, which.max(fit_sum$adjr2))
  (Intercept)             x        I(x^2)        I(x^3)
11.0834585325  8.8364132372 -1.2881201701 17.6667681176
       I(x^4)        I(x^5)        I(x^7)       I(x^10)
 0.0056041202 -0.2868687407  0.0294916189  0.0002568588
```

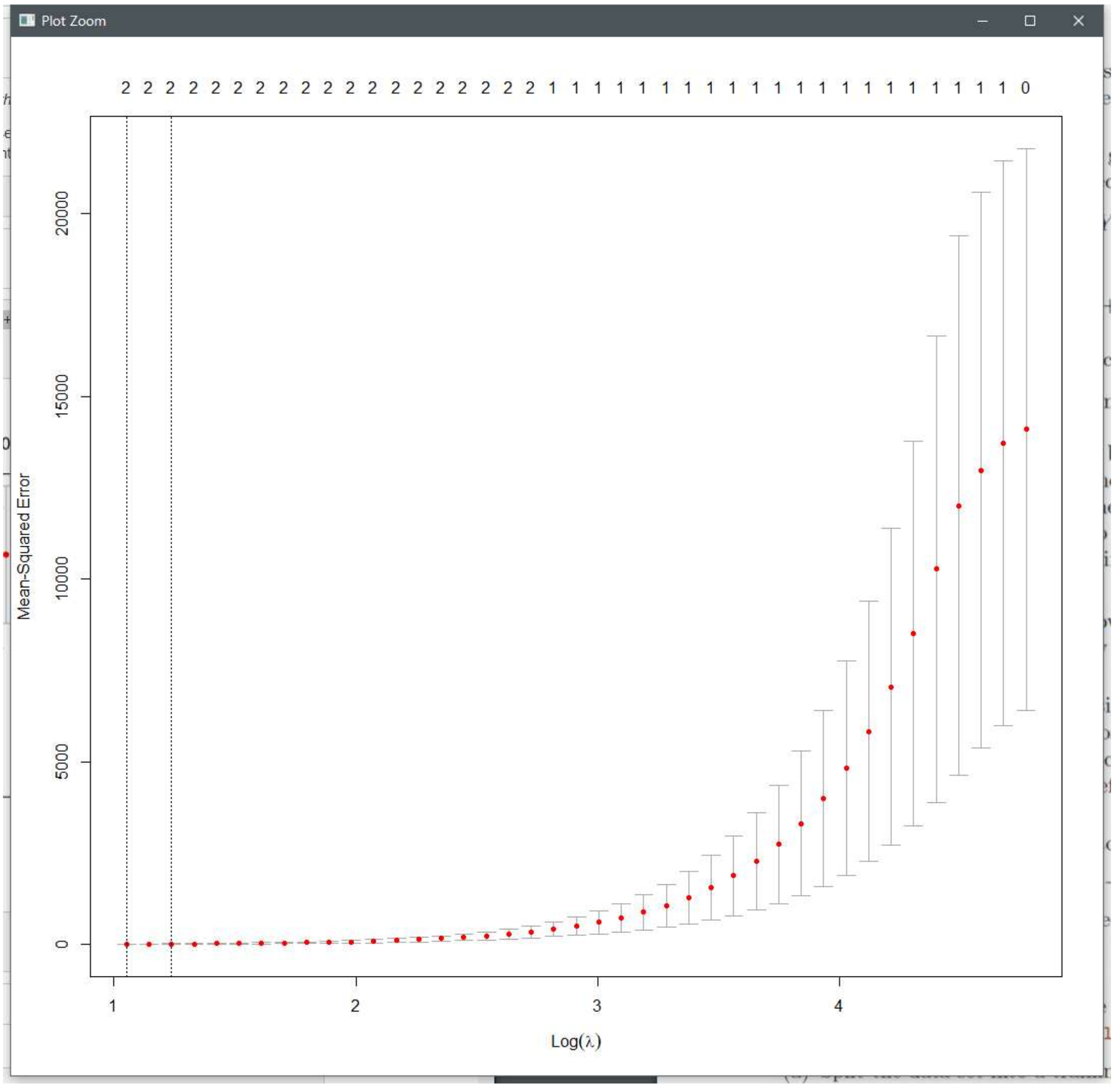
Forward: Cp and BIC choose three variables, R-square chooses seven variables




```
(Intercept)           x      I(x^2)      I(x^3)
  10.985970    9.270659   -1.110683   16.917415
> coef(fit, which.min(fit_sum$cp))
(Intercept)           x      I(x^2)      I(x^3)
  10.985970    9.270659   -1.110683   16.917415
> coef(fit, which.max(fit_sum$adjr2))
 (Intercept)            x        I(x^2)        I(x^3)
11.169697970  8.953192737 -1.813029864 17.285694620
      I(x^4)       I(x^6)        I(x^7)        I(x^9)
 0.427962425 -0.088113386 -0.045301345  0.006181310
     I(x^10)
 0.001147168
```

Backward: Cp and BIC choose three variables, R-square chooses eight variables

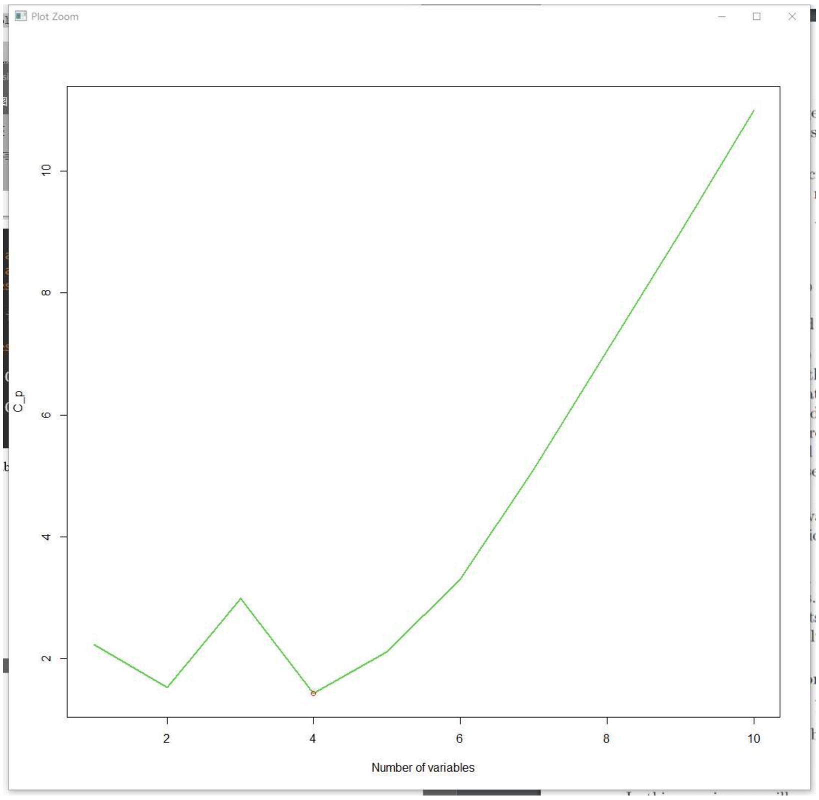
(e)

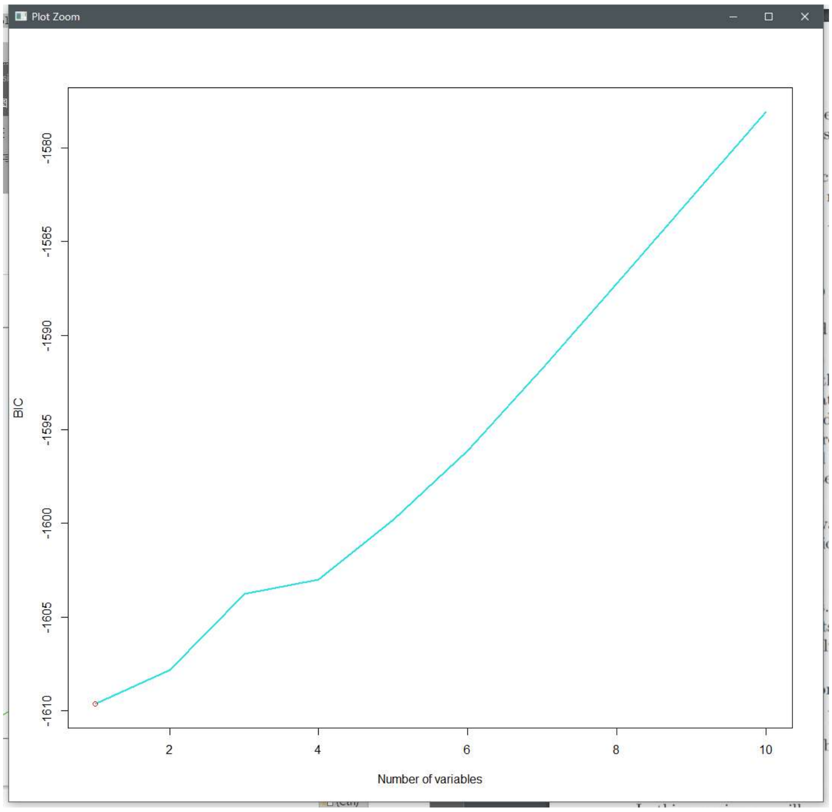


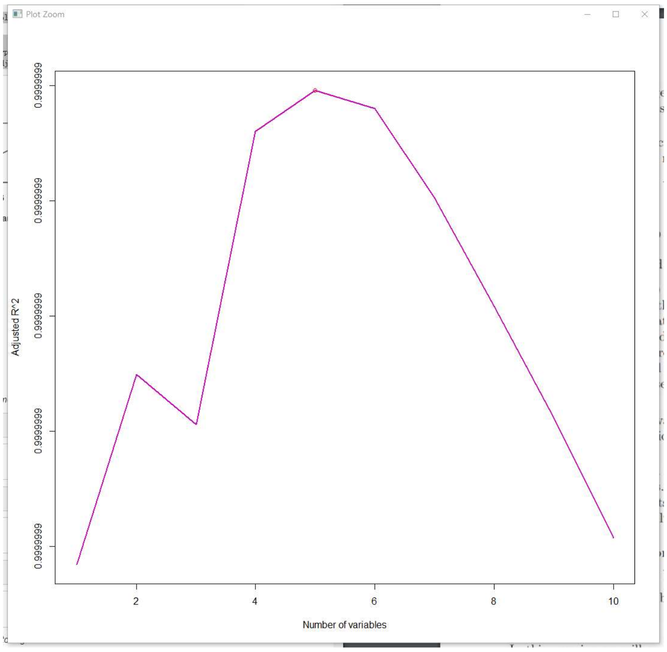

```
[1] 2.862286
> fit_lasso = glmnet(x, y, alpha = 1)
> fit_lasso = glmnet(x, y, alpha = 1)
> predict(fit_lasso, s = best, type = "cofficients")[1:11,]
Error in match.arg(type) :
  'arg' should be one of “link”, “response”, “coefficients”, “nonz
ero”, “class”
> predict(fit_lasso, s = best, type = "coefficients")[1:11,]
(Intercept)           x      I(x^2)      I(x^3)      I(x^4)
   9.764047    6.786404    0.000000   16.993473    0.000000
     I(x^5)      I(x^6)      I(x^7)      I(x^8)      I(x^9)
   0.000000    0.000000    0.000000    0.000000    0.000000
    I(x^10)
   0.000000
```

the lasso method choose X, X^3 as variables

(f)


Plot Zoom
C_p
10
8
6
4
2
2
4
6
8
10
Number of variables


Plot Zoom
BIC
-1580
-1585
-1590
-1595
-1600
-1605
-1610
2
4
6
8
10
Number of variables




```
(Intercept)      I(x^7)
  10.894301    7.000206
> coef(regfit, 4)
 (Intercept)       I(x^3)        I(x^5)        I(x^7)        I(x^9)
10.866144369 -0.656811356  0.448536350  6.924419804  0.003624614
> coef(regfit, 5)
 (Intercept)            x       I(x^3)        I(x^5)        I(x^7)        I(x^9)
10.844011820  0.589072718 -1.430738478  0.734030753  6.886852543  0.005205529
```

```
(Intercept)           x      I(x^2)      I(x^3)      I(x^4)      I(x^5)
  -1.741343    0.000000    0.000000    0.000000    0.000000    0.000000
     I(x^6)      I(x^7)      I(x^8)      I(x^9)     I(x^10)
   0.000000    6.796146    0.000000    0.000000    0.000000
```

BIC chooses 1 variable model and lasso also chooses 1-variable model.

Q9

(b)

Test Error

```
[1] 1026096
```

(c)

Test error

```
[1] 1026069
```

(d)

```
[1] 1026036
> predict(fit_lasso, s = best_lasso, type = "coefficients")
19 x 1 sparse Matrix of class "dgCMatrix"
                         1
(Intercept)    37.86520037
(Intercept)     .
PrivateYes   -551.14946609
Accept          1.74980812
Enroll         -1.36005786
Top10perc      65.55655577
Top25perc     -22.52640339
F.Undergrad     0.10181853
P.Undergrad     0.01789131
Outstate       -0.08706371
Room.Board      0.15384585
Books          -0.12227313
Personal        0.16194591
PhD           -14.29638634
Terminal       -1.03118224
S.F.Ratio       4.47956819
perc.alumni    -0.45456280
Expend          0.05618050
Grad.Rate       9.07242834
```

(e)

```
[1] 0.9104228
> ridge_r2
[1] 0.9104252
> lasso_r2
[1] 0.910428
```

R_square for these models are similar, all these three models can predict with high accuracy.

Q10

(a)

We have equation

$$p(X) = e(\beta_0 + \beta_1 X_1 + \beta_2 X_2) / (1 + e(\beta_0 + \beta_1 X_1 + \beta_2 X_2))$$

We have $\hat{\beta}_0 = -6$ $\hat{\beta}_1 = 0.05$ $\hat{\beta}_2 = 1$ $X_1 = 40$ $X_2 = 3.5$.

$$P(X) = e(-6 + 0.05 \times 40 + 1 \times 3.5) / (1 + e(-6 + 0.05 \times 40 + 1 \times 3.5)) = 0.3775$$

So the probability is 37.75%

(b)

$0.5 = e(-6 + 0.05 \times X_1 + 1 \times 3.5) \,/\, (1 + e(-6 + 0.05 \times X_1 + 1 \times 3.5))$

$e(0.05X_1 - 2.5) = 0.5(1 + e(0.05X_1 - 2.5))$

$e(0.05X_1 - 2.5) = 0.5 + 0.5e(0.05X_1 - 2.5)$

$0.5e(0.05X_1 - 2.5) = 0.5$.

$e(0.05x_1 - 2.5) = 1$.

Taking log on both side.

$0.05x_1 - 2.5 = 0 \log(1) = 0$

$X_1 = 50$.

So studentes need to study 50 hours.