MA 543/DS 502 – HW 4 responses

Problem 1)

a) iii. Relative to least squares, lasso is less flexible because of it's more restrictive method of estimating coefficients, potentially setting some of them to zero. Although increasing bias, this restrictive method diminishes the influence of variables with lesser significance on the model, leading to a lower variance. The model's prediction accuracy is therefore improved when this decrease in variance is greater than the increase in bias.

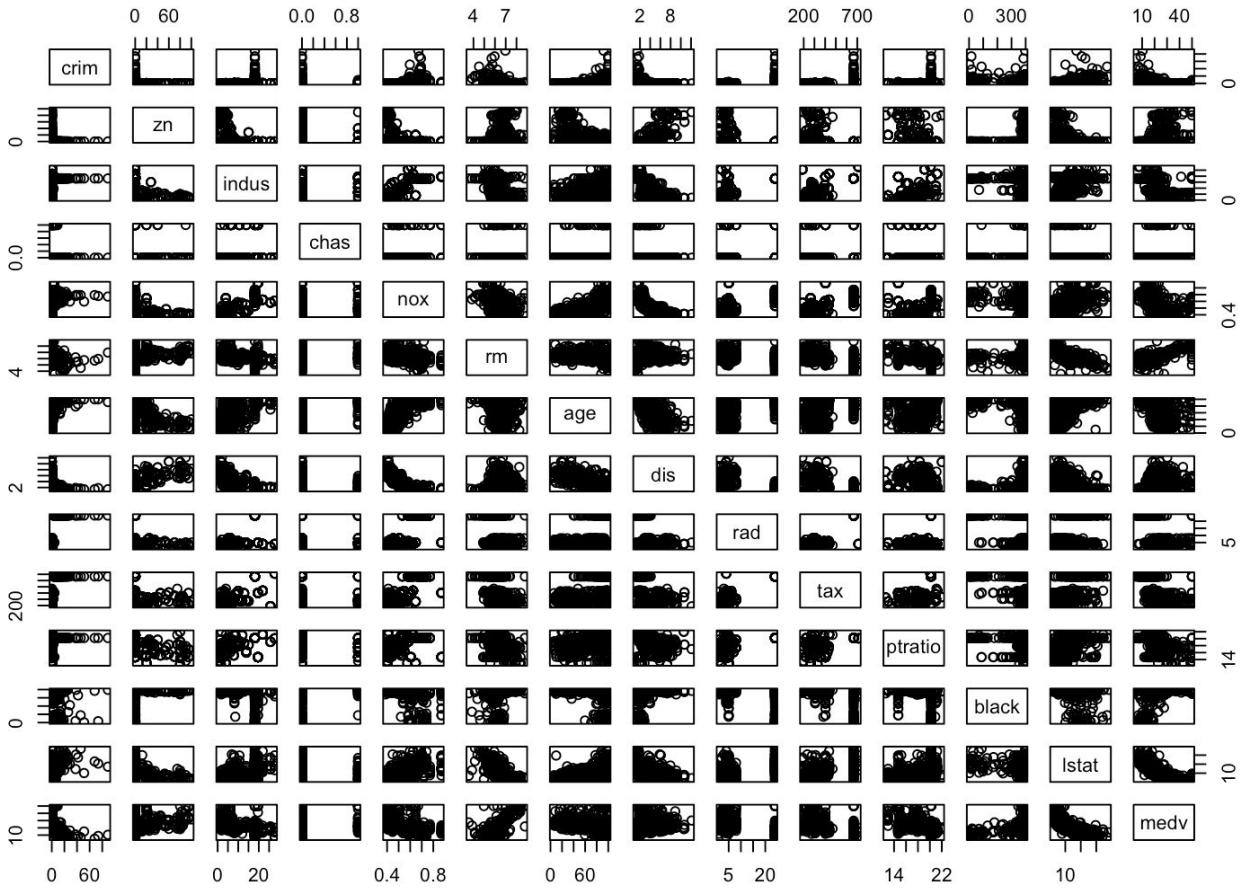
b) iii. Similarly to a), ridge regression is less flexible because of its restrictiveness, although it doesn't lower any coefficients all the way to zero. Bias increases and variance decreases in a similar manner to lasso, yielding the same kind of tradeoff. The model's prediction accuracy is also improved when this decrease in variance is greater than the increase in bias.

c) ii. Whereas least squares only allows for linear relationships, a method being non-linear allows for more flexible fits. This increased flexibility and closer fit to the data leads to a higher variance, so this variance increase should be less than the decrease in bias in order for a non-linear relationship to have improved prediction accuracy.

Problem 2)

a) and b)

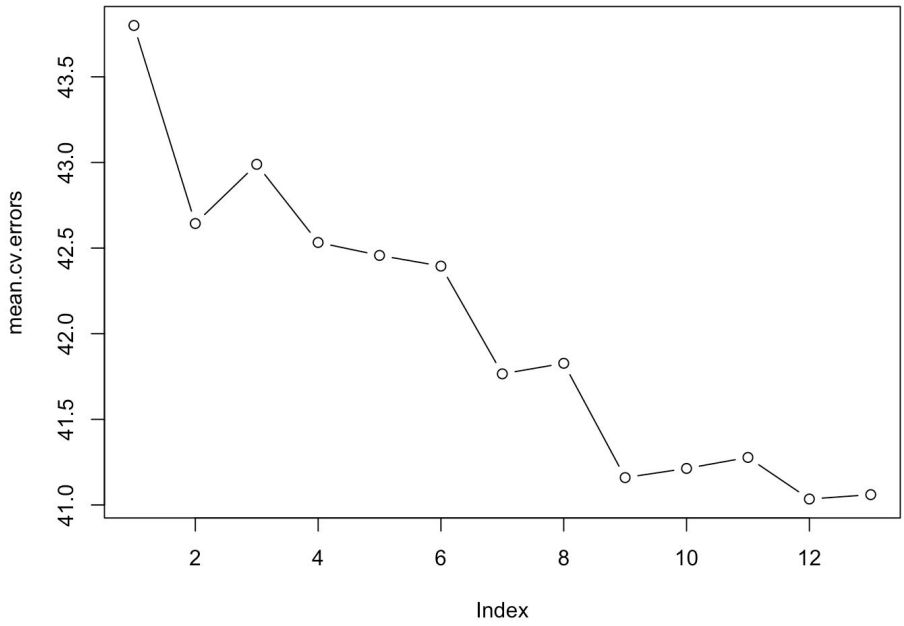
Boston dataset scatterplot pairs:



Boston dataset summary:

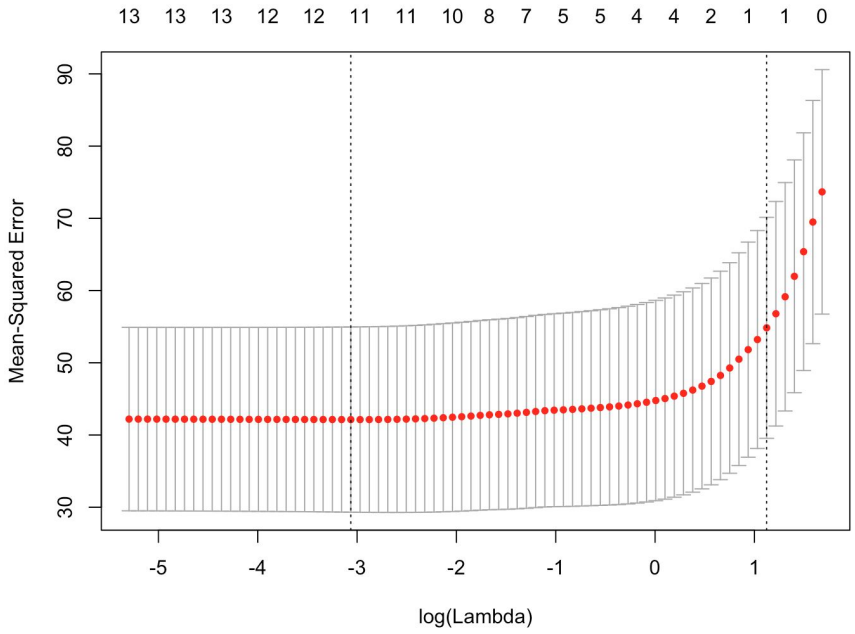
```
> summary(Boston)
      crim                zn              indus            chas              nox               rm             age        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      dis              rad              tax             ptratio          black            lstat            medv      
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

Best subset selection method:
Plot showing improving cross-validation mean square error with added parameters to the model:



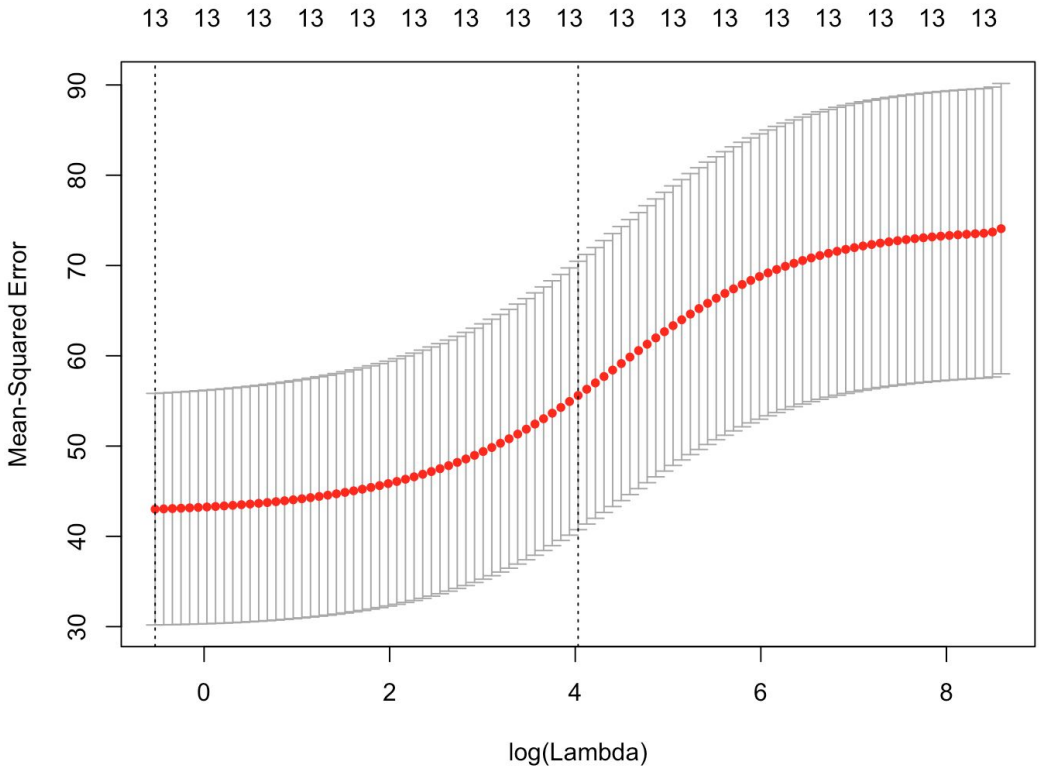
Based on the above best subset selection method, the most optimal simple prediction model would be one with 9-parameters. Furthermore, best subset selection output an average CV-MSE of **42.01192.**

Lasso:
Plot from performing lasso model fit:



According to the lasso model summary output, the only significant parameter is “rad”, with a coefficient of 0.2643196.
The lasso fitted model has an average CV-MSE of **45.34656.**

Ridge:
Plot from ridge fit:



Our ridge regression model outputs the following coefficients for our parameters:

```
15 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  1.017516864
(Intercept)  .
zn          -0.002805664
indus        0.034405928
chas        -0.225250602
nox          2.249887499
rm          -0.162546004
age          0.007343331
dis         -0.114928730
rad          0.059813844
tax          0.002659110
ptratio      0.086423005
black       -0.003342067
lstat        0.044495213
medv        -0.029124577
```

The ridge regression model has an average CV-MSE of **57.3594.**

PCR:

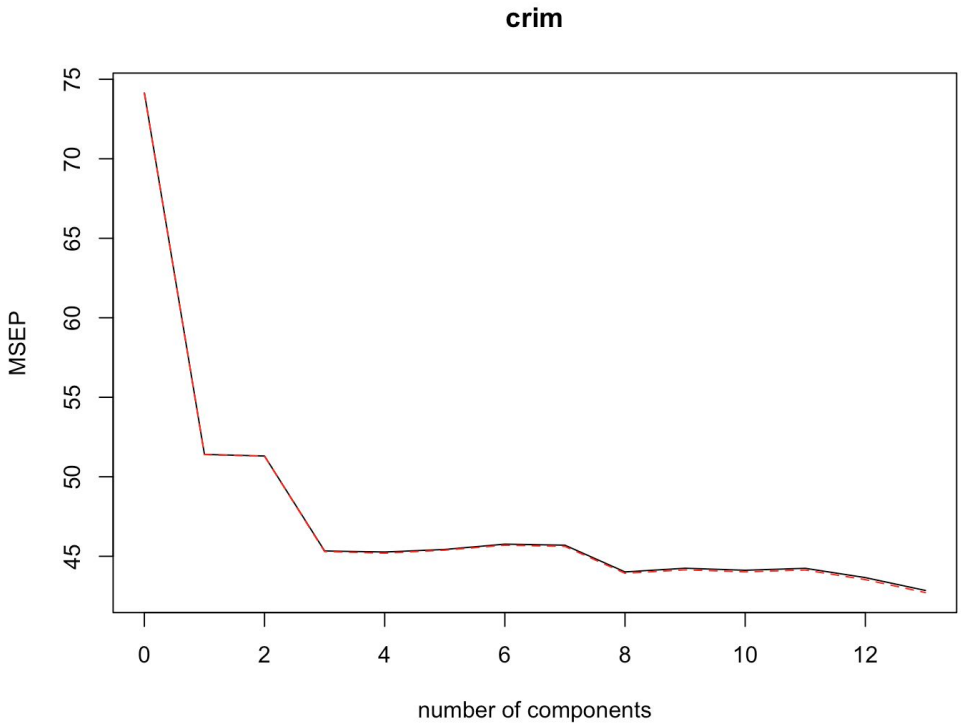
Summary of the PCR fit:

```
Data:   X dimension: 506 13
        Y dimension: 506 1
Fit method: svdpc
Number of components considered: 13

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
CV            8.61    7.170    7.163    6.733    6.728    6.740    6.765    6.760    6.634    6.652     6.642     6.652
adjCV         8.61    7.169    7.162    6.730    6.723    6.737    6.760    6.754    6.628    6.644     6.635     6.643
       12 comps  13 comps
CV        6.607     6.546
adjCV     6.598     6.536

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
X       47.70    60.36    69.67    76.45    82.99    88.00    91.14    93.45    95.40     97.04     98.46     99.52
crim    30.69    30.87    39.27    39.61    39.61    39.86    40.14    42.47    42.55     42.78     43.04     44.13
      13 comps
X        100.0
crim      45.4
```

Validation plot:



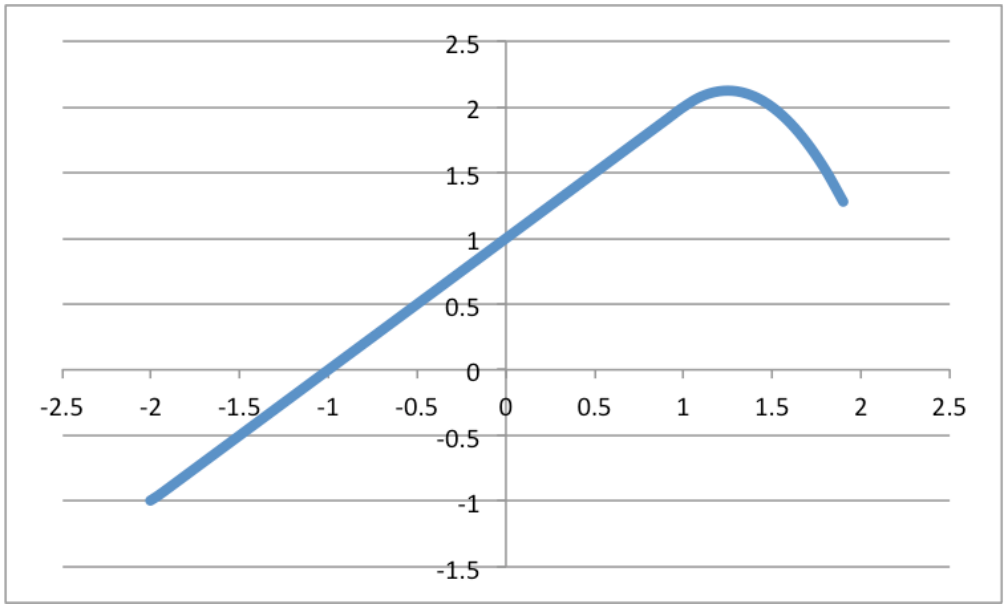
Therefore, after observing the resulting validation plot and summary of PCR fit, we can determine that a full 13-parameter model is most optimal, and a 13-parameter model would lead to an MSE of 6.536^2 = **42.719296.**

One objective of this problem was to see which of the chosen models had the best accuracy. After implementing the CV method and MSE accuracy work on the Boston dataset using best subset selection, lasso, ridge regression, and PCR, one can determine that both best subset and PCR output the most optimal models.

c)

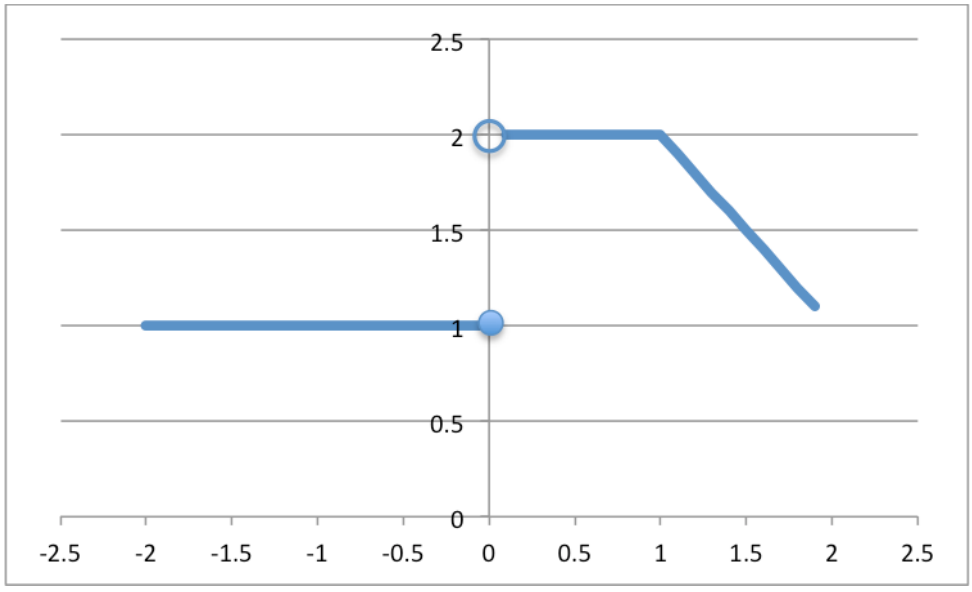
If one wanted to choose a simpler model for prediction containing 9-parameters, we would recommend the best subset selection method. However, the PCR fitted model utilizes all 13 parameters with almost the same level of accuracy.

Problem 3)



Y-intercept: (0,1)
X-intercept: (-1,0)
Slope [-2,1] = 1
Slope (1,2]: f'(x) = -4(x-1) (not a constant slope)

Problem 4)



Y-intercept: (0,1)
X-intercept: None
Slope [-2,1] = 0

Slope (1,2] = -1
Discontinuous at x=0

Problem 5)

a)

From using the cross-validation approach, **optimal degree = 9.**
In comparison with the result from ANOVA analysis:

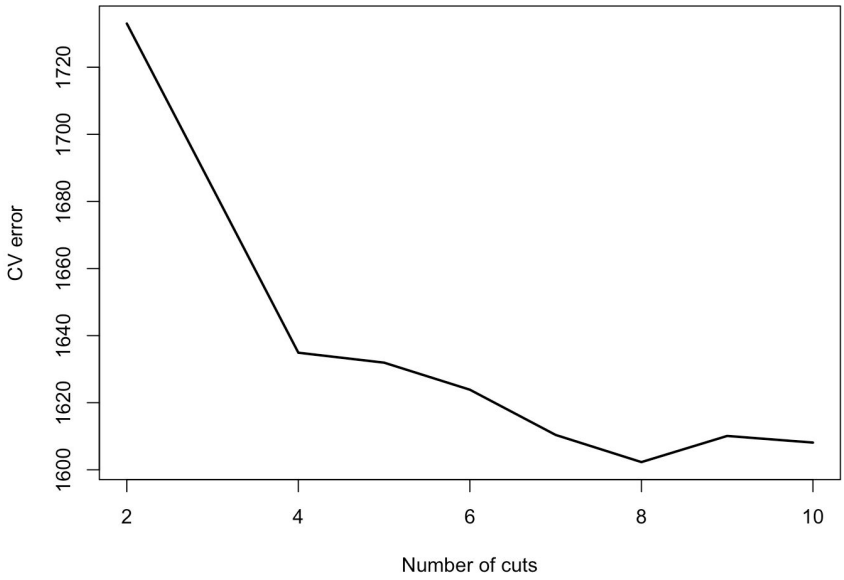
```
Analysis of Variance Table

Model  1: wage ~ poly(age, 1)
Model  2: wage ~ poly(age, 2)
Model  3: wage ~ poly(age, 3)
Model  4: wage ~ poly(age, 4)
Model  5: wage ~ poly(age, 5)
Model  6: wage ~ poly(age, 6)
Model  7: wage ~ poly(age, 7)
Model  8: wage ~ poly(age, 8)
Model  9: wage ~ poly(age, 9)
Model 10: wage ~ poly(age, 10)
   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1    2998 5022216
2    2997 4793430  1    228786 143.7638 < 2.2e-16 ***
3    2996 4777674  1     15756   9.9005  0.001669 **
4    2995 4771604  1      6070   3.8143  0.050909 .
5    2994 4770322  1      1283   0.8059  0.369398
6    2993 4766389  1      3932   2.4709  0.116074
7    2992 4763834  1      2555   1.6057  0.205199
8    2991 4763707  1       127   0.0796  0.777865
9    2990 4756703  1      7004   4.4014  0.035994 *
10   2989 4756701  1         3   0.0017  0.967529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As can be seen above, the ANOVA analysis table also states that a **9 degree polynomial model** would be the optimal fit.
Therefore, we chose to fit a 9-degree polynomial model:

b)

The CV method determined that the model with 8 cuts is the one with the minimum most error.



We then fit a step function to our Wage dataset model, based on the optimal 8 cuts. The following is our plot fit:

Problem 6)

Predicting exam scores based off of hours spent studying and hours spent sleeping.

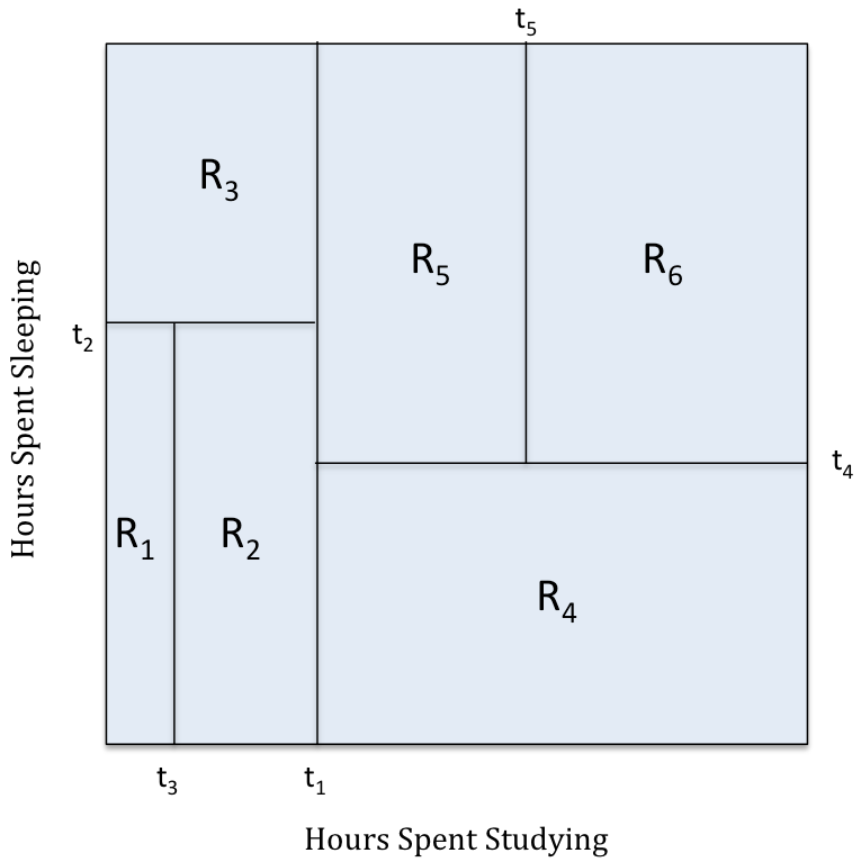
$t_1 = 3$

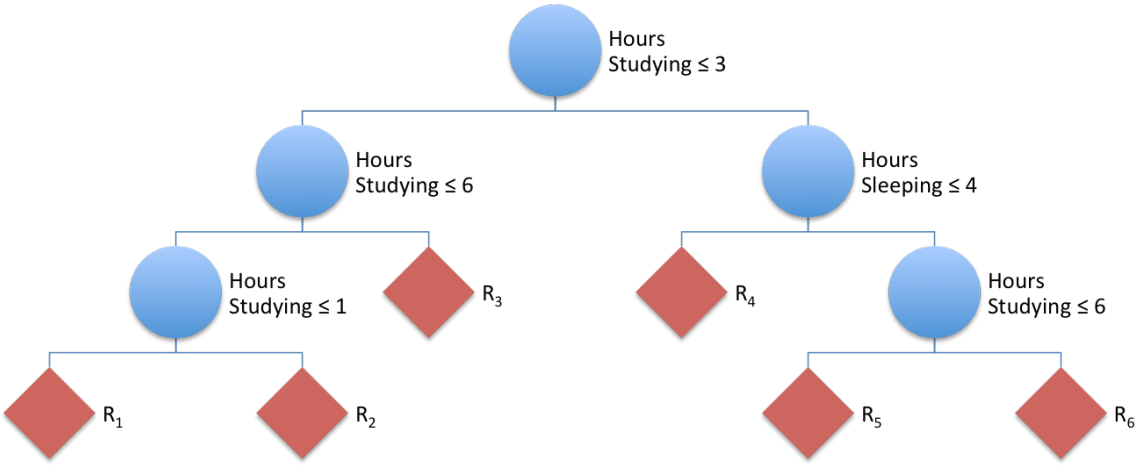
$t_2 = 6$

$t_3 = 1$

$t_4 = 4$

$t_5 = 6$

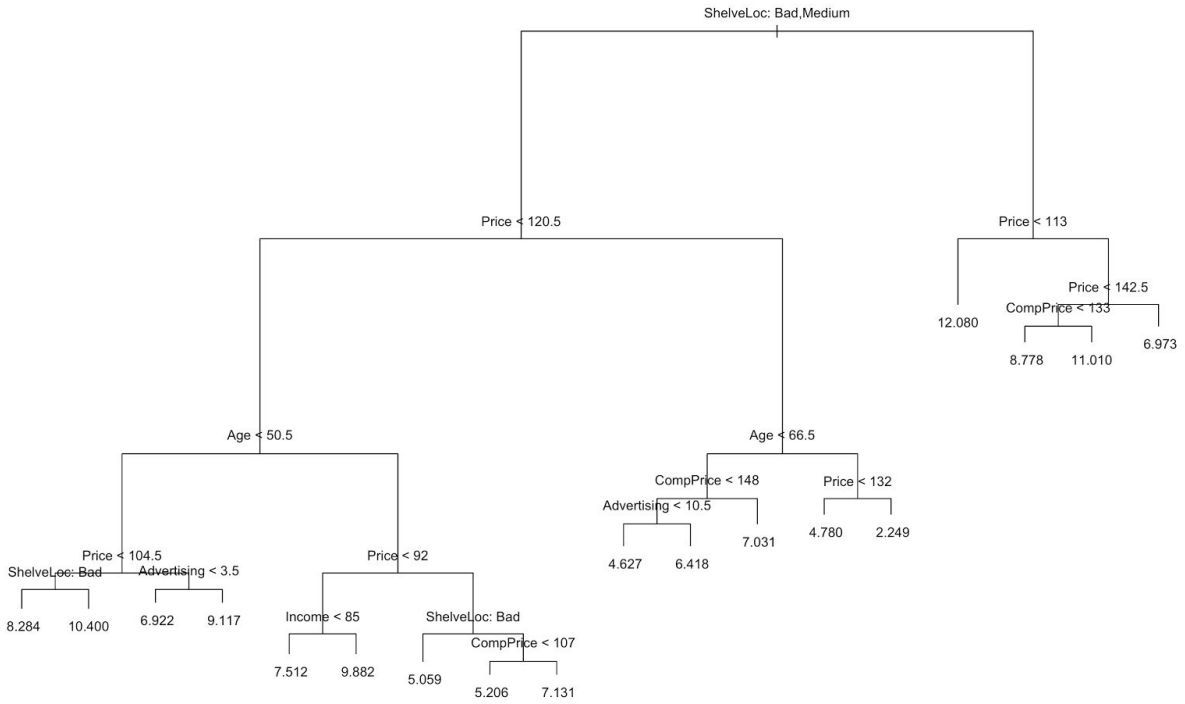




Problem 7)

b)

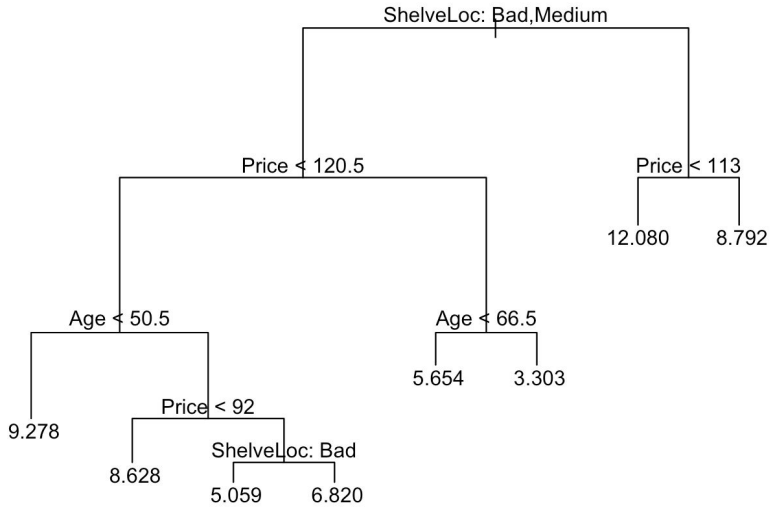
Regression tree plot:



The resulting test MSE is **4.148897.**

c)

Using CV approach, the optimal level of tree complexity is **8.**
New MSE from pruning is **5.334766.**
Tree from pruning:



Pruning the tree did NOT improve the MSE.

d)

Using the bagging approach, the resulting MSE is **2.604369.** This is indeed an improvement from the previous CV approach and regression tree.

We get the following importance table:

	%IncMSE	IncNodePurity
CompPrice	14.4124562	133.731797
Income	6.5147532	74.346961
Advertising	15.7607104	117.822651
Population	0.6031237	60.227867
Price	57.8206926	514.802084
ShelveLoc	43.0486065	319.117972
Age	19.8789659	192.880596
Education	2.9319161	39.490093
Urban	-3.1300102	8.695529
US	7.6298722	15.723975

Therefore, the parameters "**Price**", "**ShelvLoc**", and "**Age**" are the most important variables.

e)

Using random forests to analyze the data, the resulting MSE is **2.802383.**

This is not as good as the bagging methodology, but still an improvement over regression tree and the CV approach.

Importance table:

	%IncMSE	IncNodePurity
CompPrice	12.0259791	124.81403
Income	5.5542673	106.15418
Advertising	12.0466048	136.15204
Population	0.3136897	81.68162
Price	45.9639857	457.15711
ShelveLoc	36.2789679	271.76488
Age	20.8537727	196.72182
Education	2.9005332	54.16980
Urban	-0.6888196	11.86848
US	6.9739759	23.64075

Therefore, just like previously, "Price", "ShelveLoc", and "Age" are the most important variables.