Problem 1: Bias-variance tradeoff

a) Na

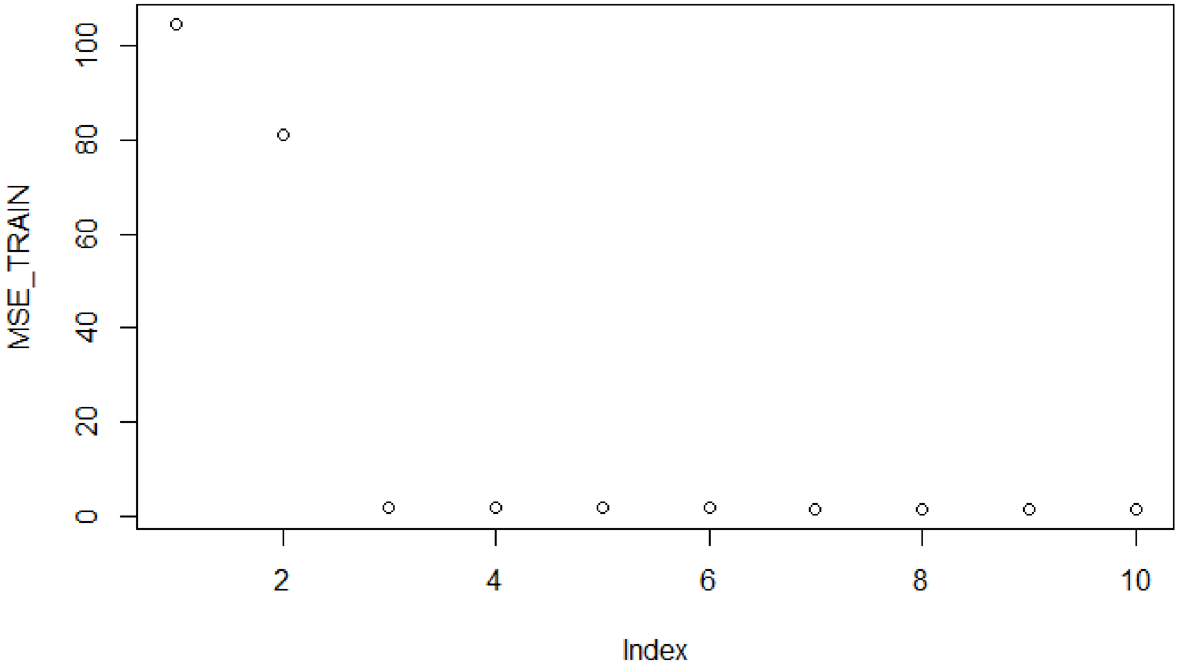
b) Na

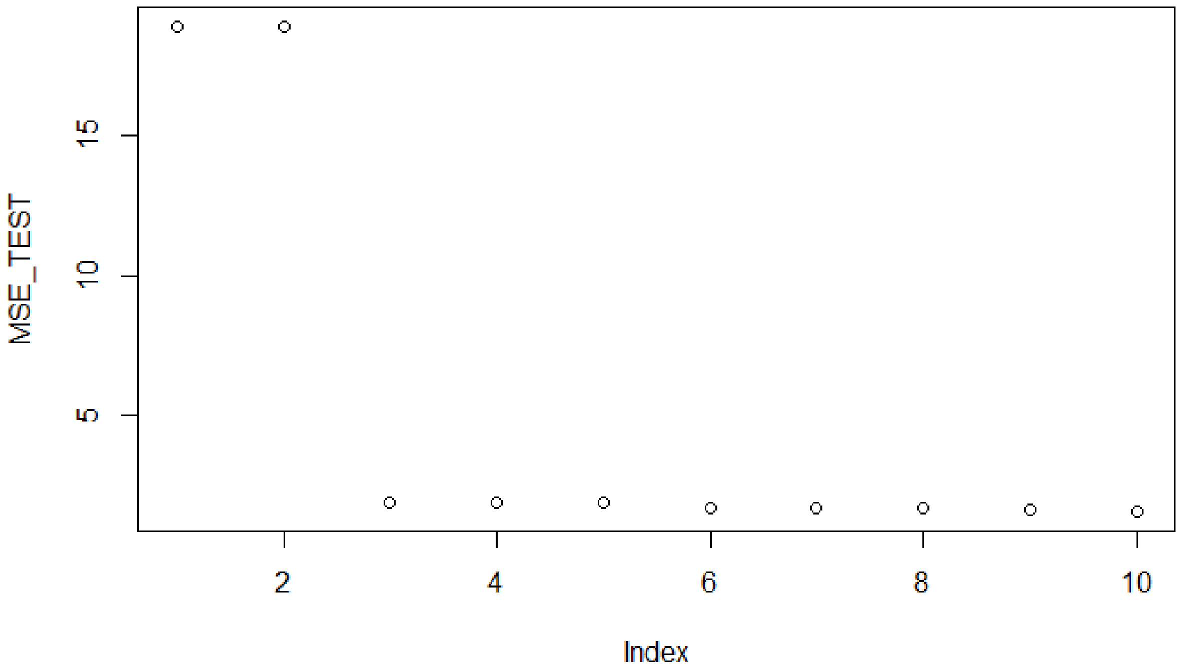
c) B0 = 1, B1 = 2, B2 = 3, B3 = 4

d) Na

e) Na

f)

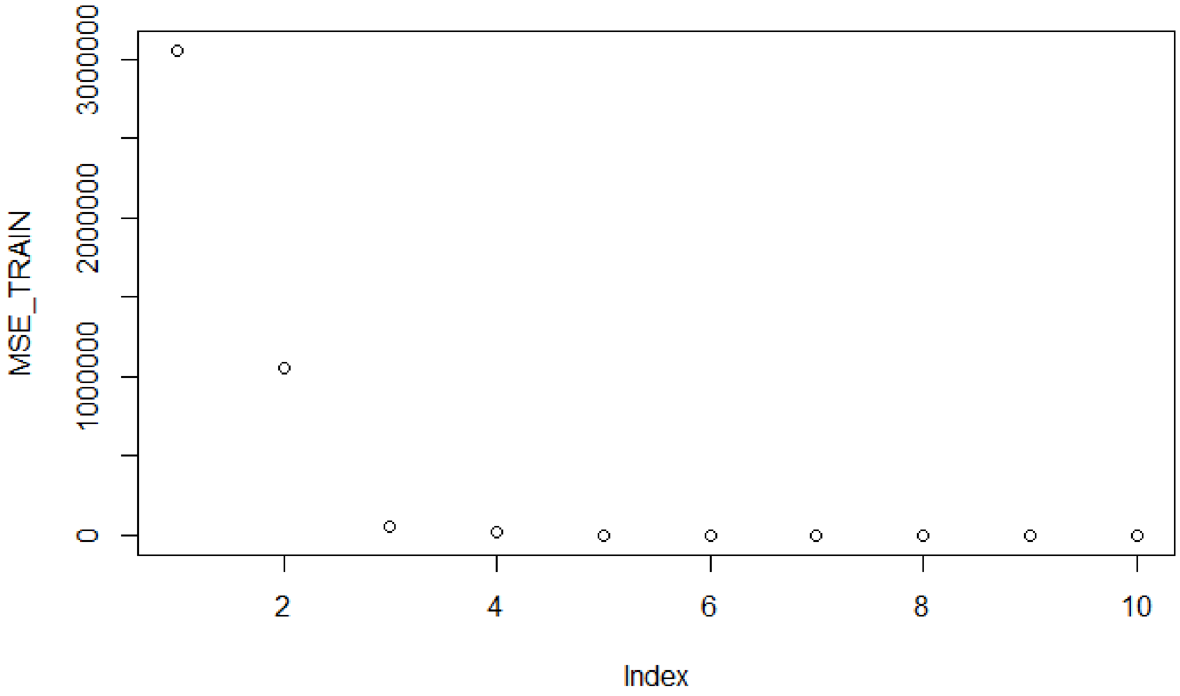


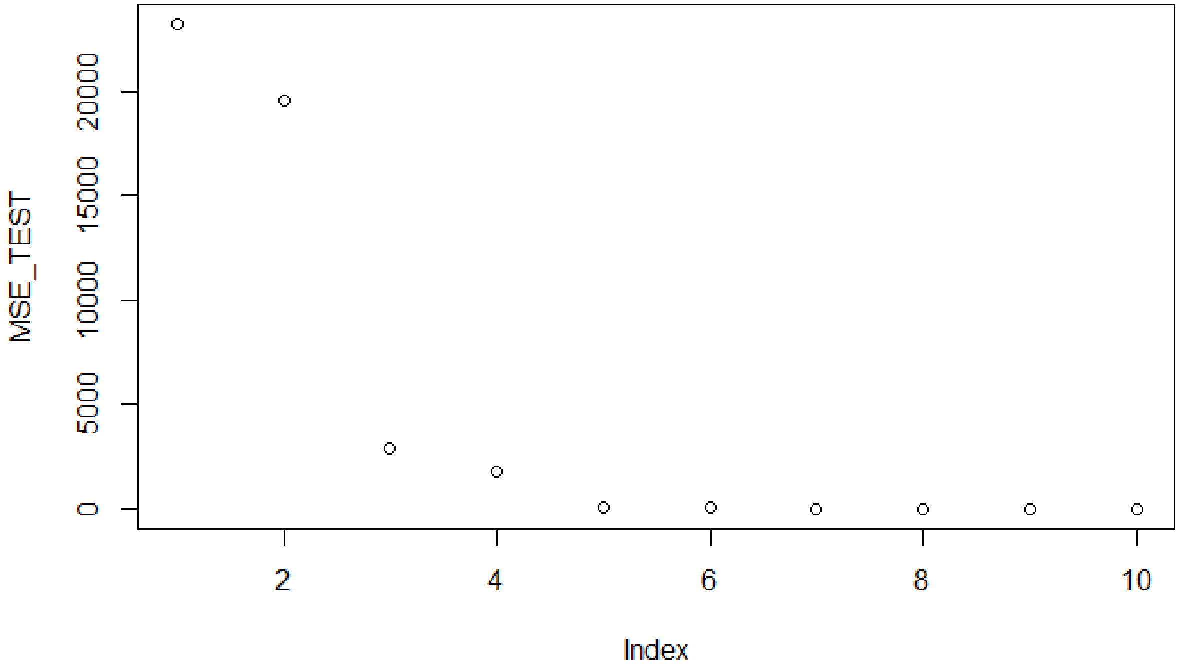


Training MSE tends to have a larger MSE compared to test MSE. However, a general trend that I'm noticing is that the MSE value will decrease (producing less errors) as we increase the number of predictors in our model.

g) The model with the highest number of predictors will have the smallest training and test mse. As the number of predictors increases, bias will decrease and variance will increase as we are overfitting the model leading to our model creating assumptions based on no real relationship.

h) B4 = 5, B5 = 6, B6 = 7, B7 = 8

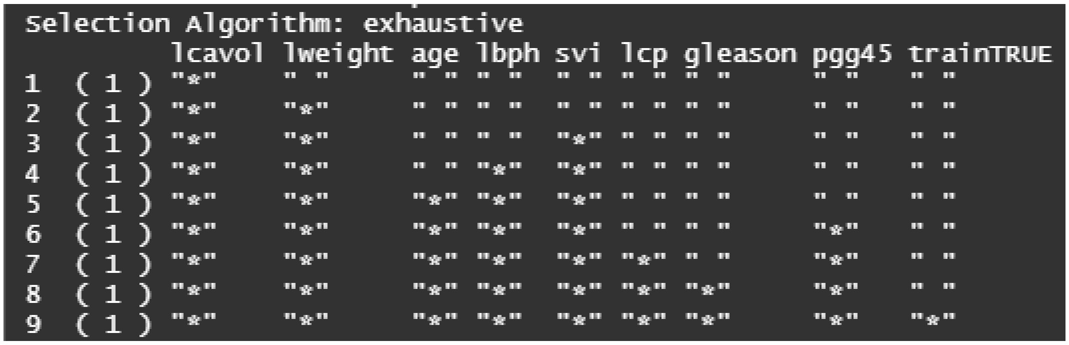

MSE_TRAIN
0
1000000
2000000
3000000
Index
2
4
6
8
10


MSE_TEST
0
5000
10000
15000
20000
Index
2
4
6
8
10

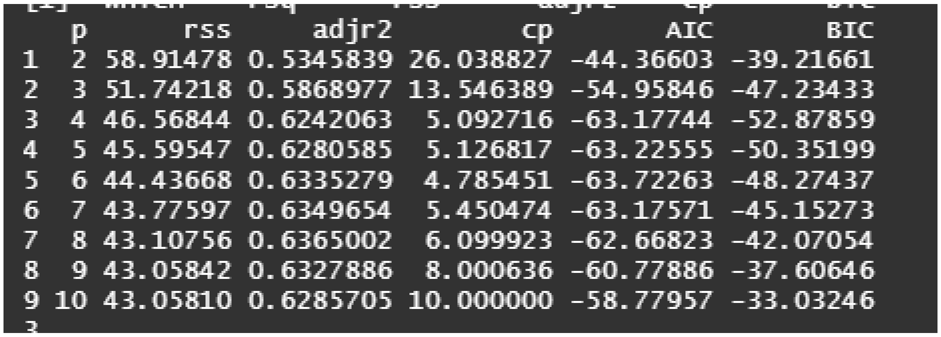
"h)" has an identical behavior compared to "f)". The MSE value for each similar increased, but train MSE is still larger than train MSE and the over behavior (where the MSE will decrease as you increase the the number of predictors) has not changed.

Problem 2: Best subset selection

a)



```
Selection Algorithm: exhaustive
         lcavol lweight age lbph svi lcp gleason pgg45 trainTRUE
1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "   " "
2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "   " "
3  ( 1 ) "*"    "*"     " " " "  "*" " " " "     " "   " "
4  ( 1 ) "*"    "*"     " " "*"  "*" " " " "     " "   " "
5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "   " "
6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     "*"   " "
7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"   " "
8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"   " "
9  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"   "*"
```



```
   p      rss     adjr2        cp       AIC       BIC
1  2 58.91478 0.5345839 26.038827 -44.36603 -39.21661
2  3 51.74218 0.5868977 13.546389 -54.95846 -47.23433
3  4 46.56844 0.6242063  5.092716 -63.17744 -52.87859
4  5 45.59547 0.6280585  5.126817 -63.22555 -50.35199
5  6 44.43668 0.6335279  4.785451 -63.72263 -48.27437
6  7 43.77597 0.6349654  5.450474 -63.17571 -45.15273
7  8 43.10756 0.6365002  6.099923 -62.66823 -42.07054
8  9 43.05842 0.6327886  8.000636 -60.77886 -37.60646
9 10 43.05810 0.6285705 10.000000 -58.77957 -33.03246
```

Smallest BIC

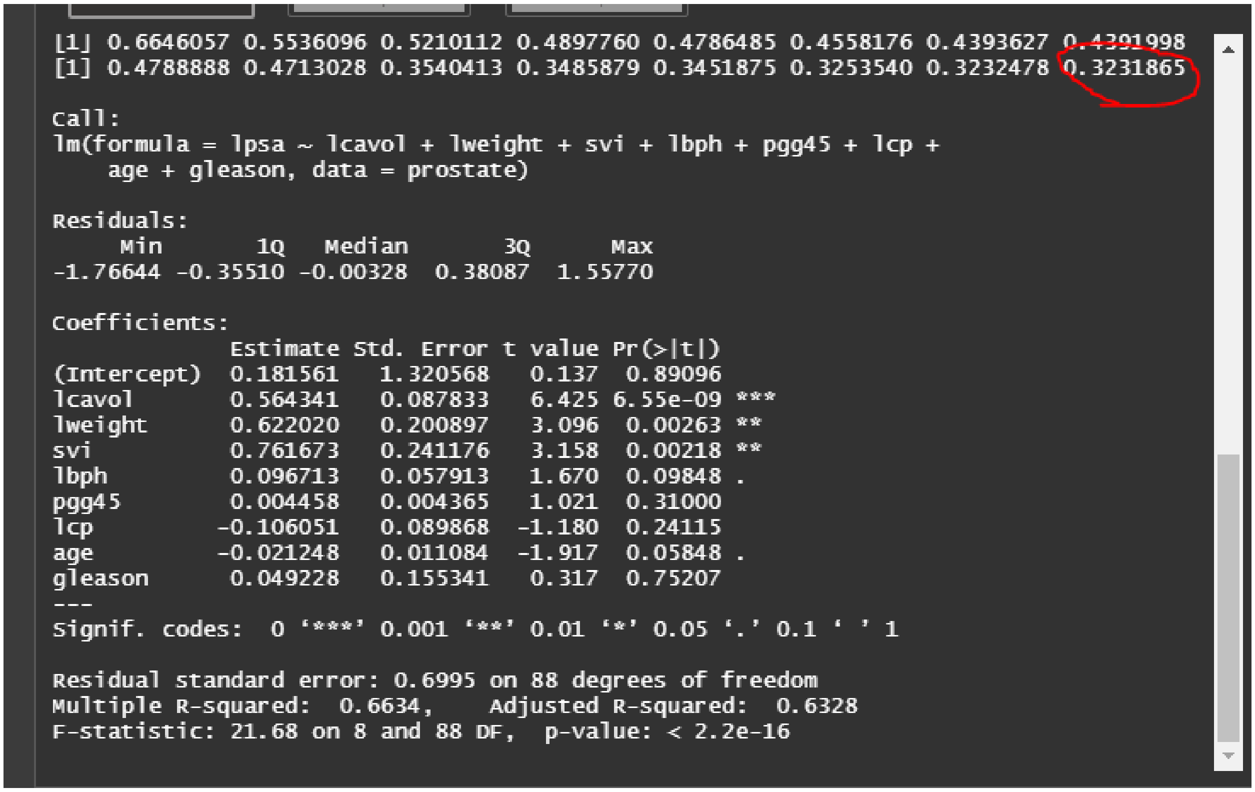
- Model 3 (first pic)
- Line 3 (second pic)

Smallest AIC

- Model 5 (first pic)
- Line 5 (second pic)

The table shows the best models for each model size (numbers of predictors). An example is model 5 with the predictors of lcavol, lweight, age, lbph, svi as these are the best combination to produce the most optimized model.
Model 3 (lcavol, lweight, svi) has the smallest BIC values of -52.87859.
Model 5 (lcavol, lweight, age, lbph, svi) has the smallest AIC values of -63.72263.
Model 3 is the best value as the Mallow's Cp is the smallest and it also has smaller AIC and BIC.

b) Model 8 has the smallest test MSE value of 0.3231865



```
[1] 0.6646057 0.5536096 0.5210112 0.4897760 0.4786485 0.4558176 0.4393627 0.4391998
[1] 0.4788888 0.4713028 0.3540413 0.3485879 0.3451875 0.3253540 0.3232478 0.3231865

Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lbph + pgg45 + lcp + 
    age + gleason, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.76644 -0.35510 -0.00328  0.38087  1.55770 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.181561   1.320568   0.137  0.89096    
lcavol       0.564341   0.087833   6.425 6.55e-09 ***
lweight      0.622020   0.200897   3.096  0.00263 ** 
svi          0.761673   0.241176   3.158  0.00218 ** 
lbph         0.096713   0.057913   1.670  0.09848 .  
pgg45        0.004458   0.004365   1.021  0.31000    
lcp         -0.106051   0.089868  -1.180  0.24115    
age         -0.021248   0.011084  -1.917  0.05848 .  
gleason      0.049228   0.155341   0.317  0.75207    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6995 on 88 degrees of freedom
Multiple R-squared:  0.6634,	Adjusted R-squared:  0.6328 
F-statistic: 21.68 on 8 and 88 DF,  p-value: < 2.2e-16
```

c)

i) 0.56287586 0.46116249 0.20101350 0.09150810 0.07521495 0.06563046 0.05874743 0.03735541

ii) Model 8 has the smallest coefficients (lcavol, weight, age, lbph, svi, lcp, gleason, pgg45).

Problem 3: Cross-validation

a) It works by taking the number of observations (the numbers of rows/n) and randomly splitting them into k (non overlapping groups with the length of n/k). The k group will become the validation sets and and the left over will become the training sets. The test error is then estimated by averaging the k resulting MSE estimates.

b)

i. The validation set approach?

A disadvantages of the validation set approach relative to k-fold cross-validation is the validation estimate of the test error rate can be highly variable (depends on which observations are included in the training/validation set). Another disadvantage is that only a subset of the observations are used to fit the model, so the validation set error may overestimate the test error rate for the model fit on the entire data set.

ii. LOOCV?

LOOCV has less bias. We repeatedly fit the statistical learning method using training data that contains n-1 obs., i.e. almost all the data set is used LOOCV produces a less variable MSE. The validation approach produces different MSE when applied repeatedly due to randomness in the splitting process, while performing LOOCV multiple times will always yield the same results, because we split based on 1 obs. each time LOOCV is computationally intensive (disadvantage). We fit each model n times.

c) Na

d) Na

e) They are very similar, but different. This could be due to different seeds.

f) Model 2 from the second iteration has the smallest LOOCV error as additional variables like x3 and x4 do not seem to reduce the LOOCV.

g) The p value is very small, thus the coefficient estimates are statistically significant and close to the true coefficients.

Problem 4: Concept Review

a) True, because all the models have the same number of predictors. That mean it will be possible for us to get different results as Mallow’s Cp, AIC, BIC, and adjusted R^2 and different combinations of coefficients will create different results for each factor.

b) False, RSS4 will be lower because the higher numbers of predictors will produce lower error term, lower rss.

c) False. RSS4 will be lower because the higher numbers of predictors will produce lower error term, lower rss.