Problem 1

a) Lowest training error

Best subset selection method would have the lowest training error. This is because it search all $\binom{p}{k}$ possible k-variables models, choosing the one with lowest training RSS while the other two methods just search a subset space of all possible k-variables models.

b) Lowest testing error

Since every method may lead to over-fitting with k variables, we could not tell which one has the lowest testing error.

c) TRUE or FALSE

i. TRUE: the k+1 variable model contains all k features chosen in the k variable model, plus the best additional feature.
ii. TRUE: the k variable model contains all but one feature in the k+1 best model which resulting in the smallest gain in RSS.
iii. FALSE: it is possible but not guaranteed that there are same sets in two methods.
iv. FALSE: it is possible but not guaranteed.
v. FALSE: Different from stepwise selection methods when choosing k+1-varialbes model, best subsets methods may change some variables in k-variables model in k+1-variables model. So this proposition is not guaranteed.

Problem 2

a) Lasso relative to least square

(iii) is correct: By adding a penalty parameter associate with $||\beta||_1$, lasso gains significant decreasing in variance(flexibility) by sacrificing a little increase in bias.

b) Ridge relative to least square

(iii) is also correct: By adding a penalty parameter associate with $||\beta||_2$, ridge gains decreasing in variance(flexibility) by sacrificing a little increase in bias. As long as it does not result in too high of a bias due to its added constraints, it will outperform least squares which might be fitting spurious parameters.

c) Non-linear model relative to least square

(ii) is correct: Non linear methods are generally more flexible than least squares. They perform better when the linearity assumption is broken, having more variance due to

their more sensitive fits to the underlying data, performing well will need to have a substantial drop in bias.

Problem 3

a) Training error

(iii) is correct. The training error would increase steadily as λ increase from 0. This is because the penalty λ works as a restrictor on least square regression, making the model less flexibility which leads to larger RSS on training set.

b) Testing error

(ii) is correct. The testing error would Decrease initially, and then eventually start increasing in a U shape since larger λ leads to less flexibility but more bias so that decreased RSS on testing sets. However eventually necessary coefficients will be removed from the model, and the test RSS will again increase, making a U shape.

c) Variance

(iv) is correct. The variance would Steadily decrease because larger λ leads to less flexibility so that smaller variance.

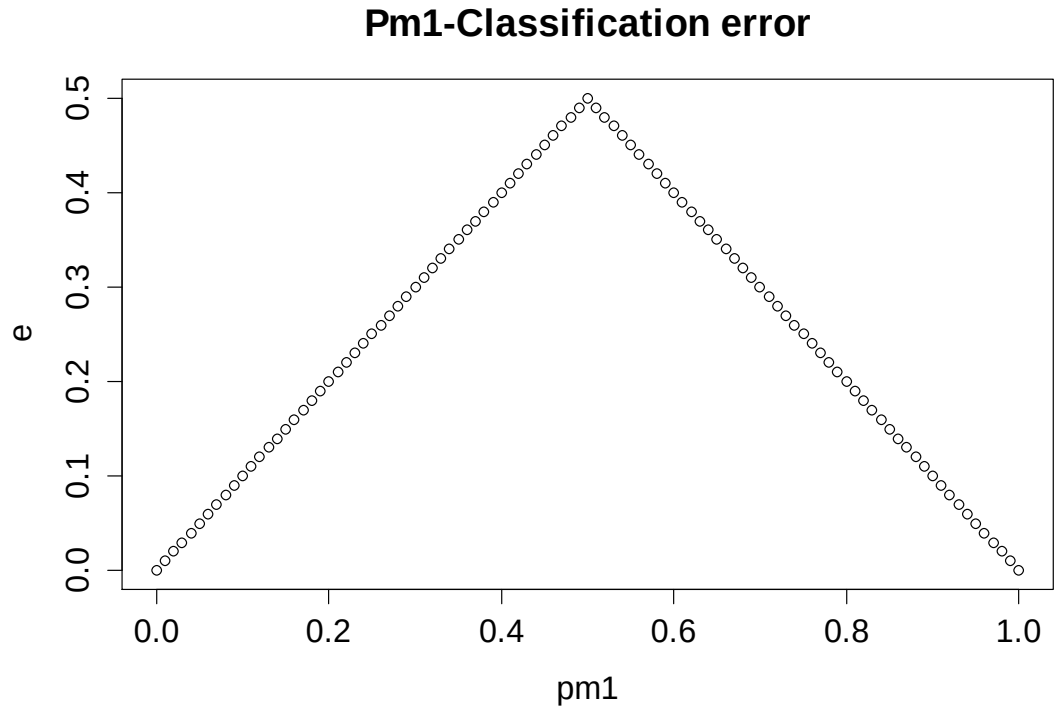
d) Bias

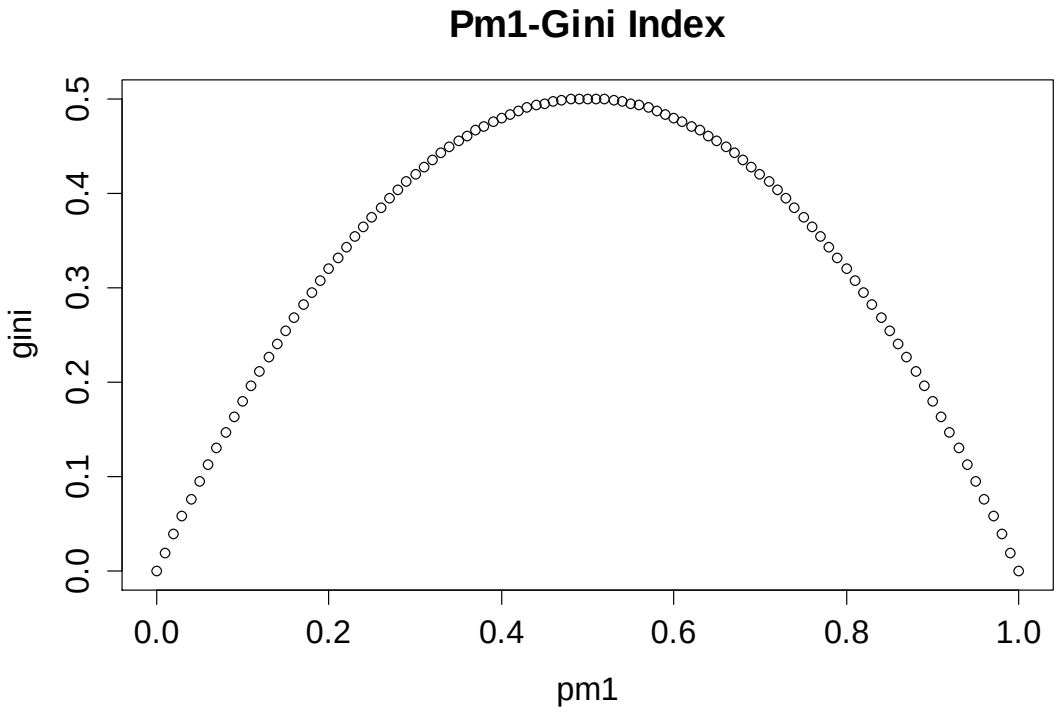
(iii) is correct. The bias would steadily increase because larger λ leads to less flexibility so that larger bias.

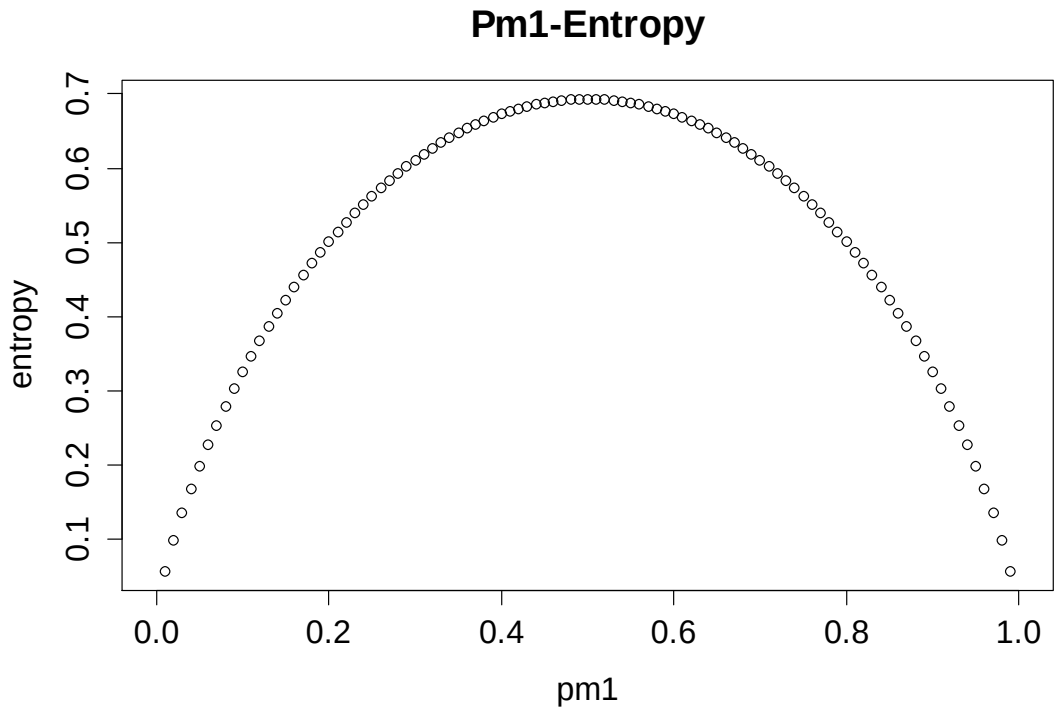
e) Irreducible error

(v) is correct. The irreducible error would stay constant because this part of error could not be reduced by any methods.

Problem 4


Pm1-Classification error
e
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
pm1


Pm1-Gini Index
gini
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
pm1



Problem 5

0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75

a) Majority vote approach

Prob	0.1	0.15	0.2	0.55	0.6	0.65	0.7	0.75
class	Green	Green	Green*2	Red	Red*2	Red	Red	Red

So final classification is RED based on majority vote.

b) Average approach

The average of the 10 prediction value is 0.45, so the final classification is GREEN based on average approach.

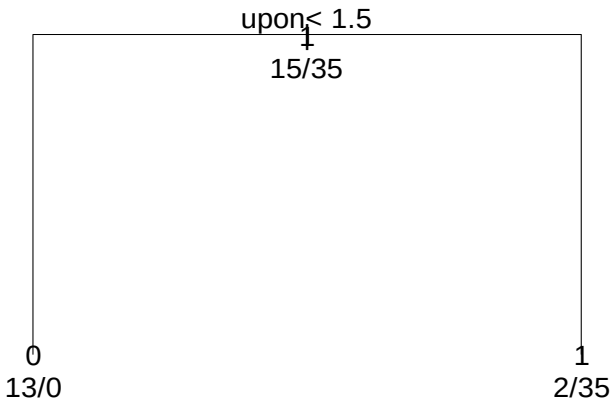
Problem 6

a) Gini Index split

1-"Hamilton";0-"Non-Hamilton"

		Predicted Class		Total
		0	1	
True Class	0	TN:10(0.909)	FP:1(0.091)	11
	1	FN:0(0.000)	TP:16(1.000)	16
Tatal		10	17	

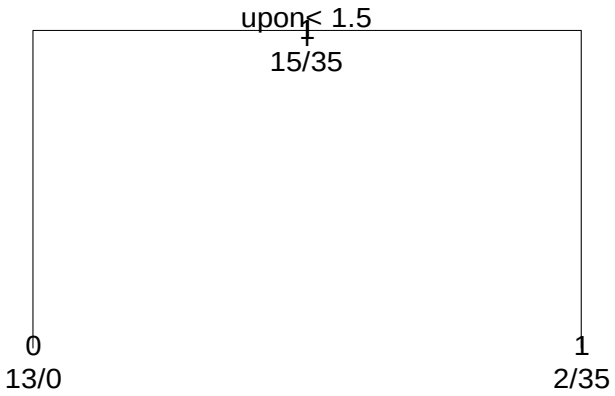
Classification Tree by Gini split



b) Information Gain split

		Predicted Class		Total
		0	1	
True Class	0	TN:10(0.909)	FP:1(0.091)	11
	1	FN:0(0.000)	TP:16(1.000)	16
Tatal		10	17	

Classification Tree for Info split



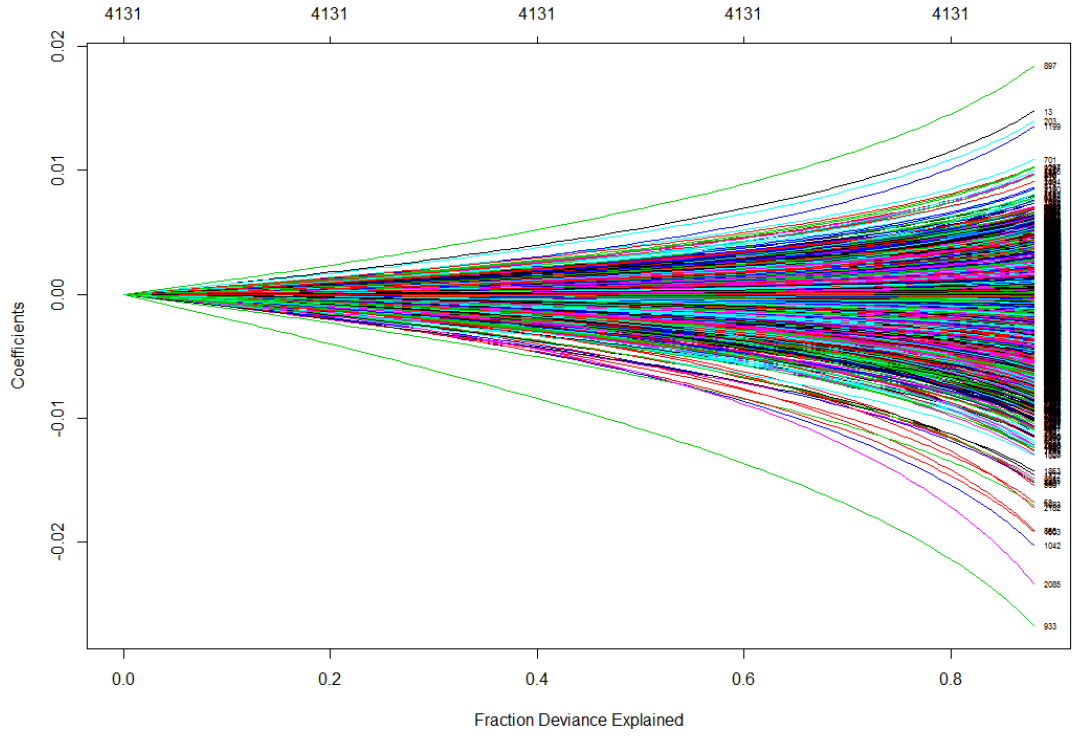
There is no significant difference between the result under these two split, either on the tree themselves or the prediction errors.

Problem 7

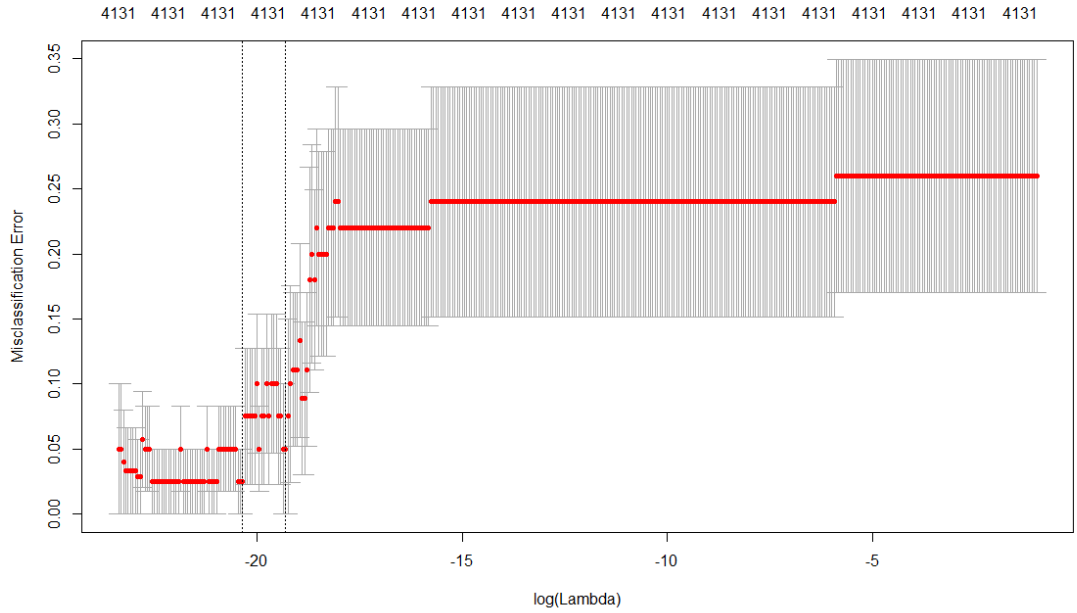
a) Why should we use regularized regression in this problem?

Since there are 4875 words work as variables, the full model would contain at most 4876 coefficients (p=4875). However, the number of observations is so small compared with p, such that $n << p$. So if we use unregularized regression model, there would be serious overfitting problem which leads to large variance on testing sets because they use all variables with no constrains. On the contrary, the regularized regression add constrains on the coefficients, greatly reducing variance by sacrificing a little bit increasing on bias, which guarantees a better precision on model predictions.

b)



Since Ridge regression could not shrink coefficients equal to zero exactly, λ should be very small so that the penalty term will not act beyond least square terms. We let $\lambda \in [e^{-30}, e^{-1}]$ artificially because the default scope of λ is not small enough . And we use 10-fold cross validation method to search the best lambda.




```
> cv.ridge$lambda.min
[1] 1.448546e-09
```

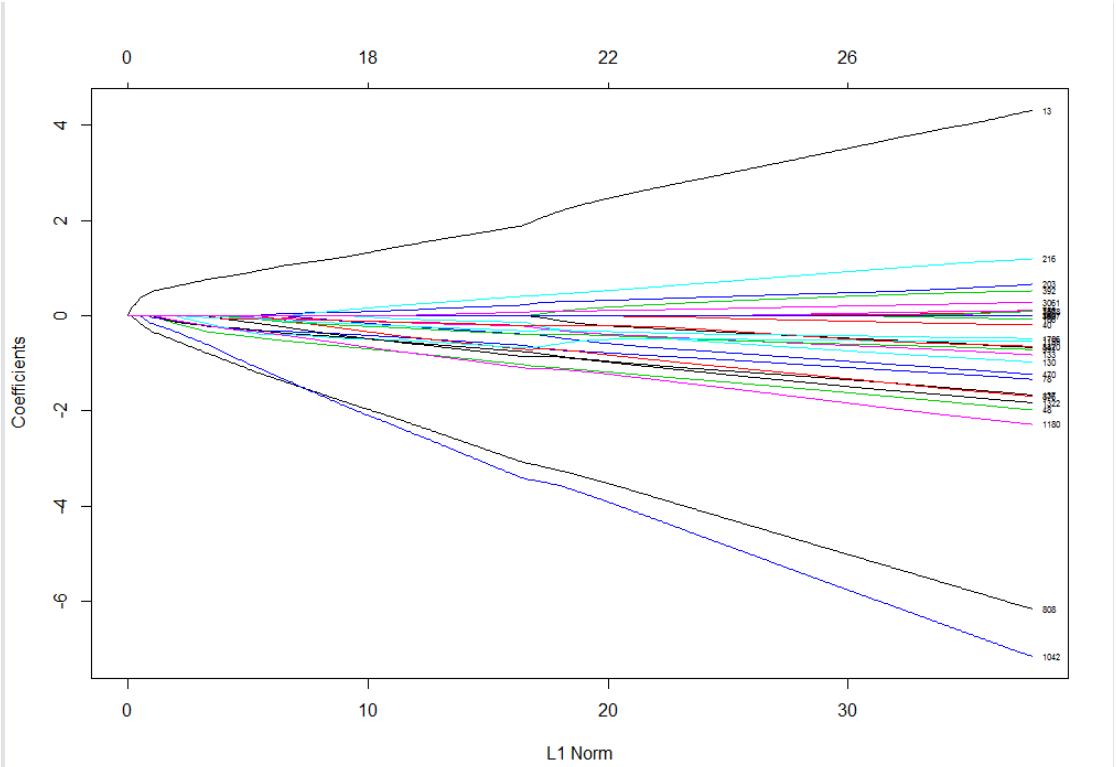
The classification result on testing sets is:

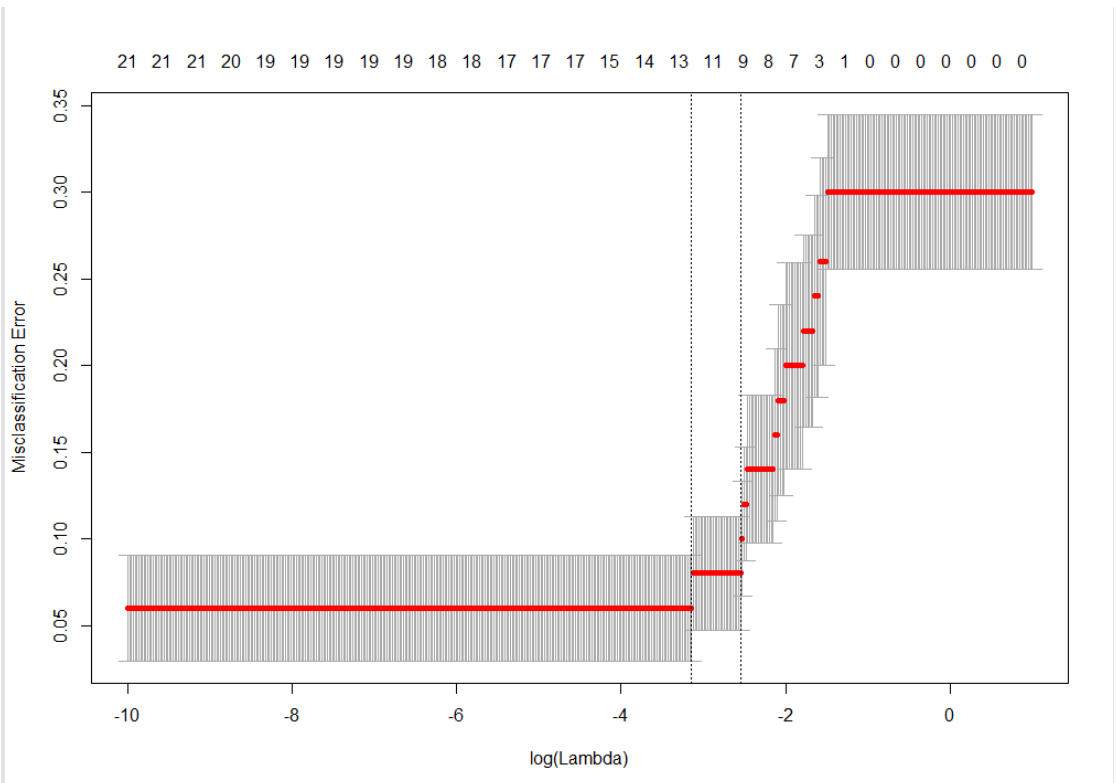
		Predicted Class		Total
		0	1	
True Class	0	TN:10(0.9090)	FP:1(0.0909)	11
	1	FN:1(0.0625)	TP:15(0.937500)	16
Tatal		11	16	

The 10 most important words are

upon	power	will	nation	can	may	everi	feder	particular	union
1.685	1.406	1.364	1.299	1.114	1.0361	0.8443	0.8353	0.8073	0.7905

c)





We find the best lambda by using 10-fold cross validation.

```
> cv.lasso$lambda.min
[1] 0.06457613
```

The classification result on testing sets is:

		Predicted Class		Total
		0	1	
True Class	0	TN:9(0.8182)	FP:2(0.1818)	11
	1	FN:0(0.0000)	TP:16(1.0000)	16
Tatal		9	18	

The 10 most important words are

whilst	upon	februari	form	although	lesser	sever	within	anim	nobl
0.9402	0.8248	0.8022	0.383	0.2687	0.2554	0.2540	0.2427	0.072	0.067

The word "upon" is the only same word from the two classification methods. However, it is definitely the only word that used for classification in tree method, which indicates that this word "upon" is the most useful variable that could distinguish the works from Hamilton and Madison. The more the weight of "upon" is, the smaller the misclassification error is.