ASM Practice

Ridge Regression

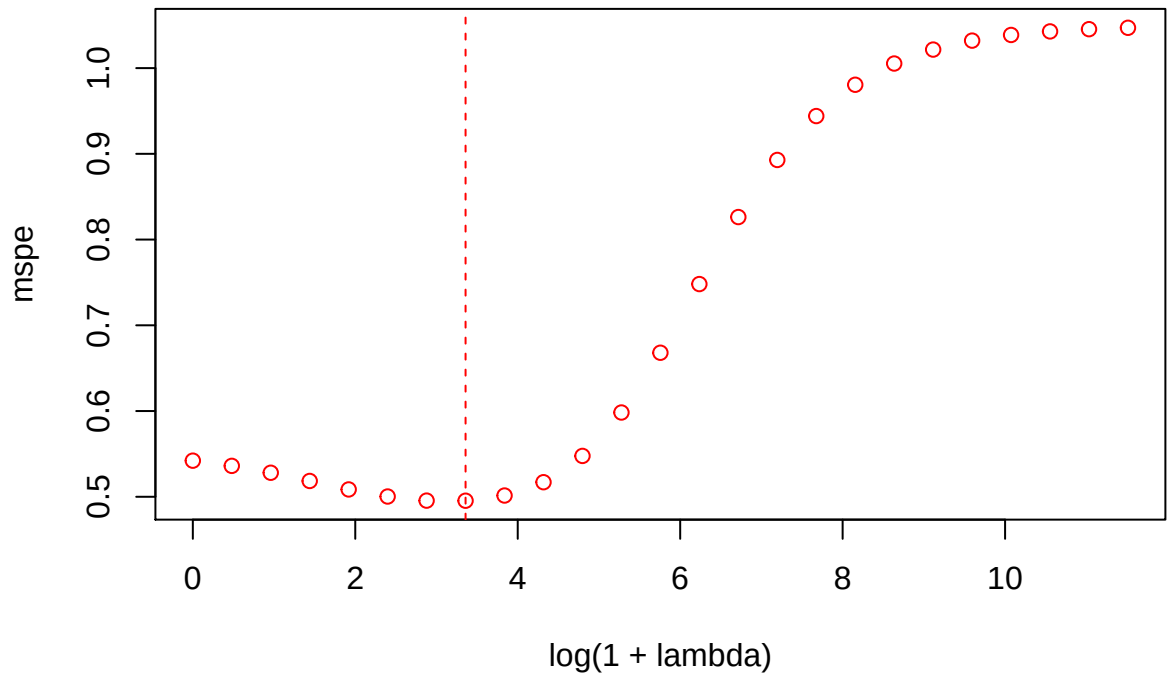
Maria Gkotsopoulou & Ricard Monge Calvo & Amalia Vradi

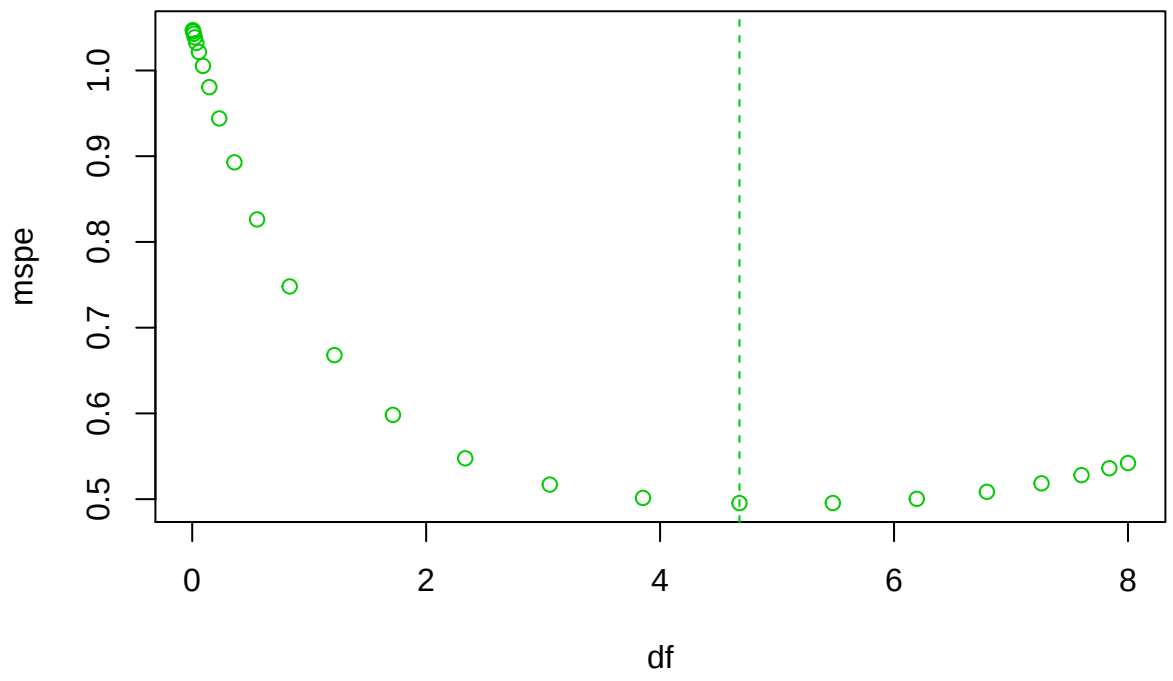
27/10/2019

Choosing the penalization parameter λ

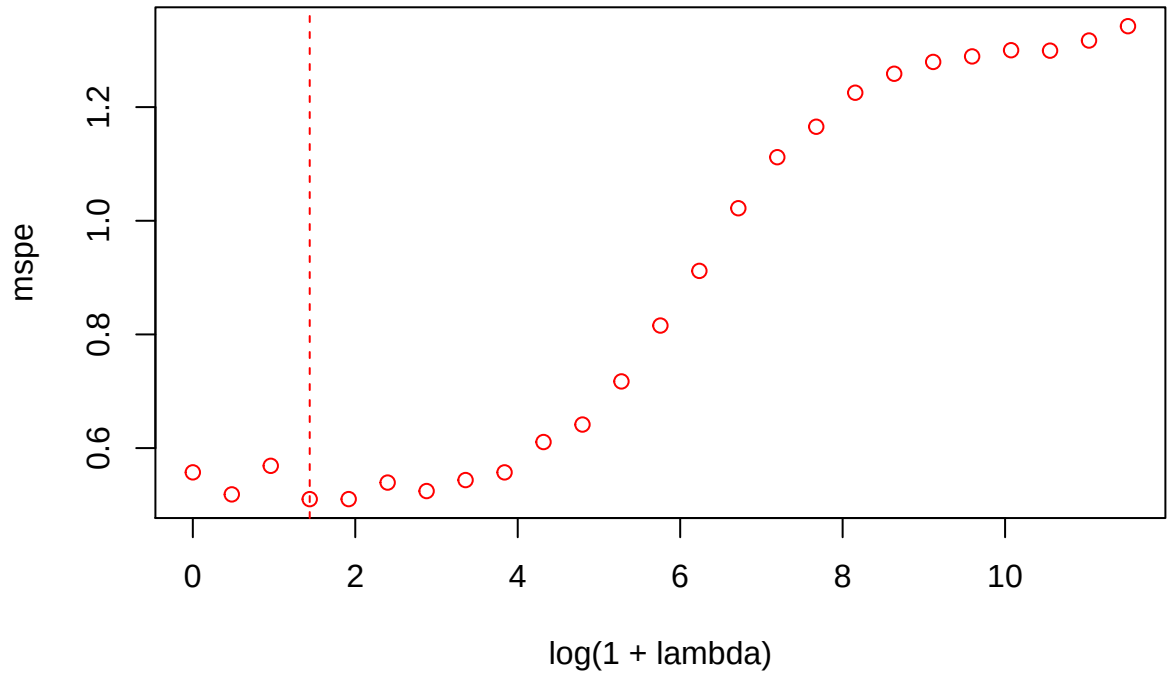
1. Ridge regression lambda search
2. Ridge regression lambda search with CV
3. Prostate data application

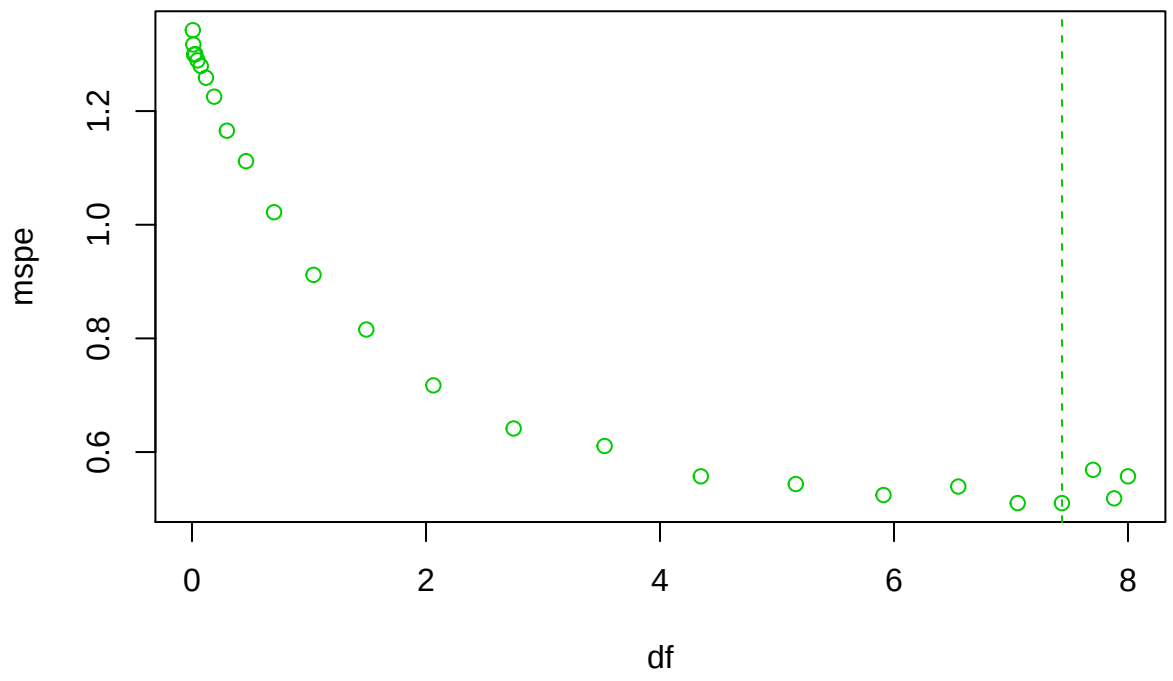
```
## With validtion data of size 30 instances.
```

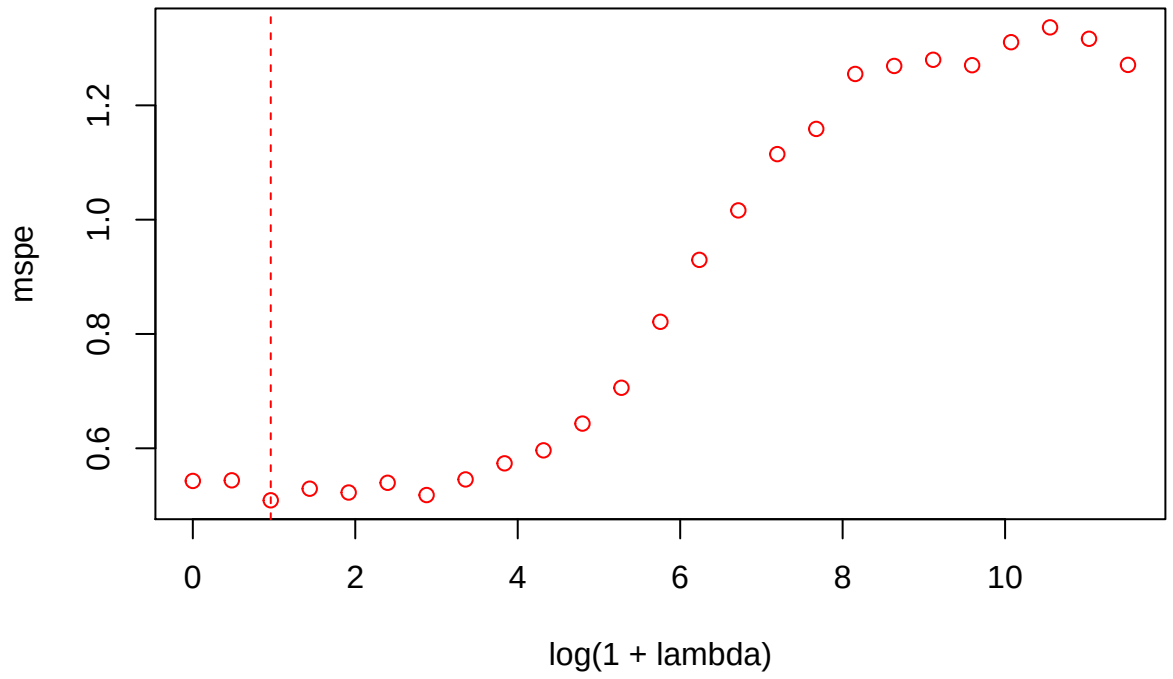


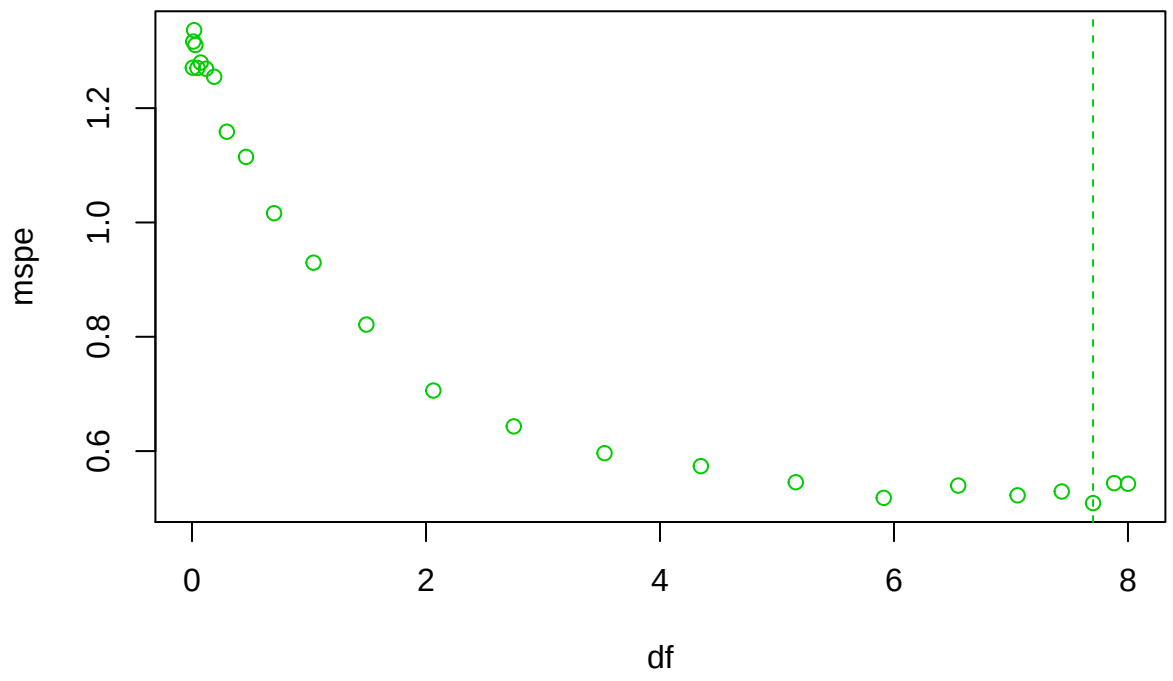



```
## With 5-fold and 10-fold Cross Validation respectively.
```

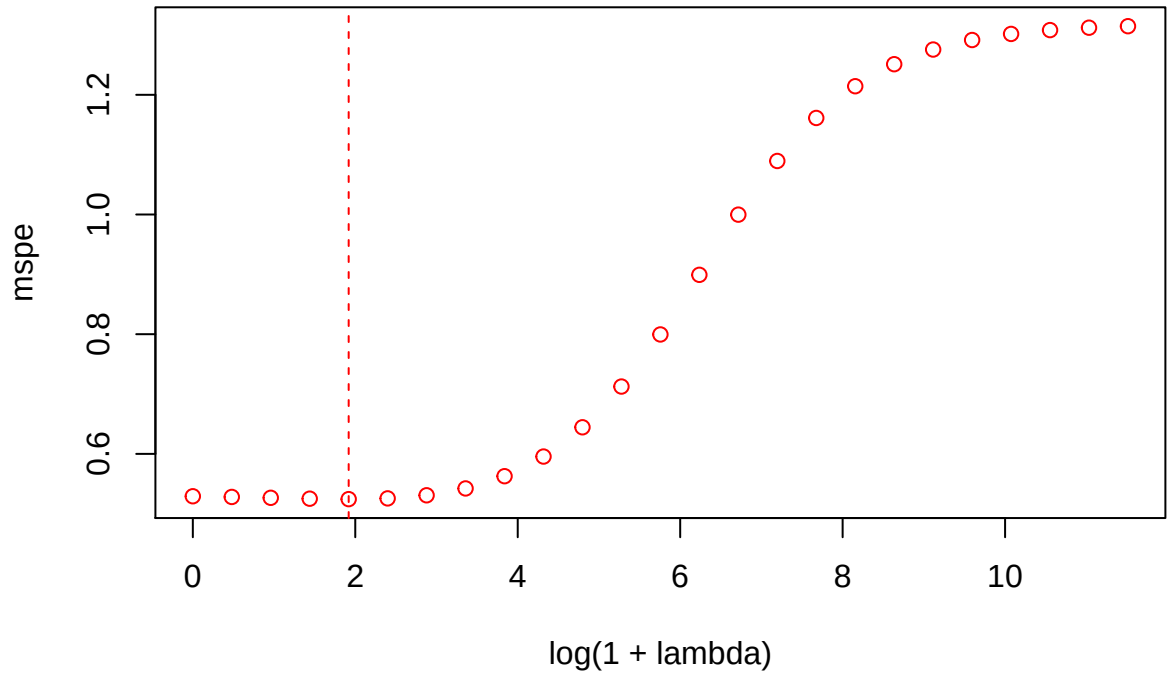


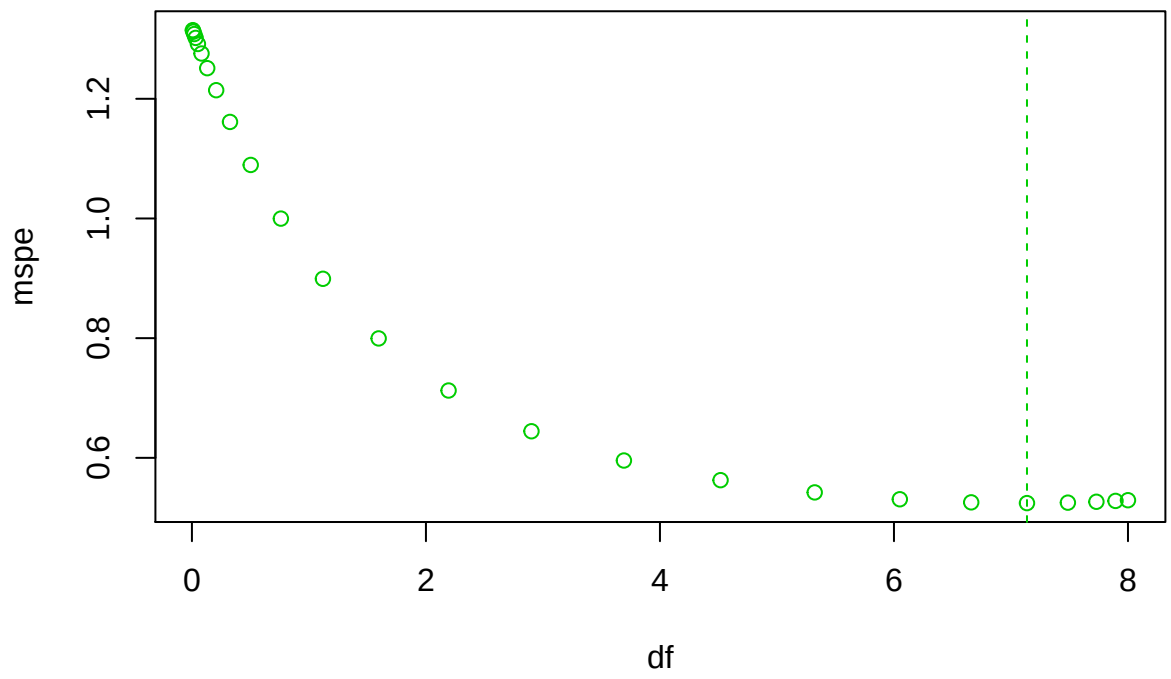

mspe
0.6
0.8
1.0
1.2
0
2
4
6
8
df

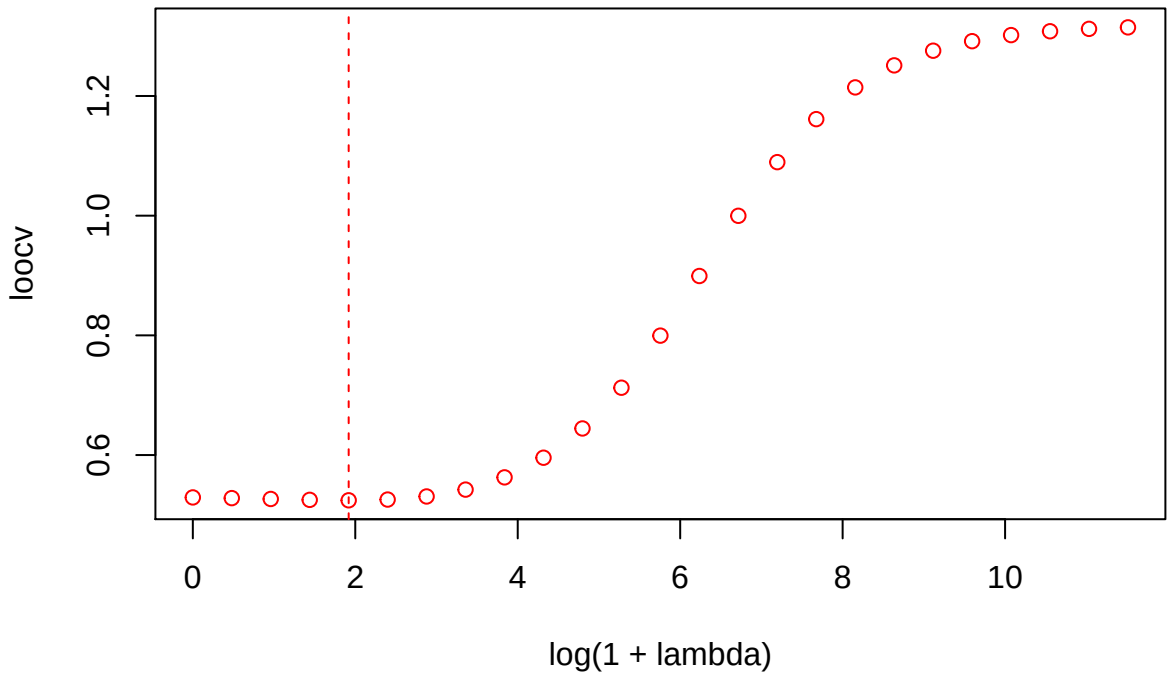

mspe
0.6
0.8
1.0
1.2
0
2
4
6
8
10
log(1 + lambda)

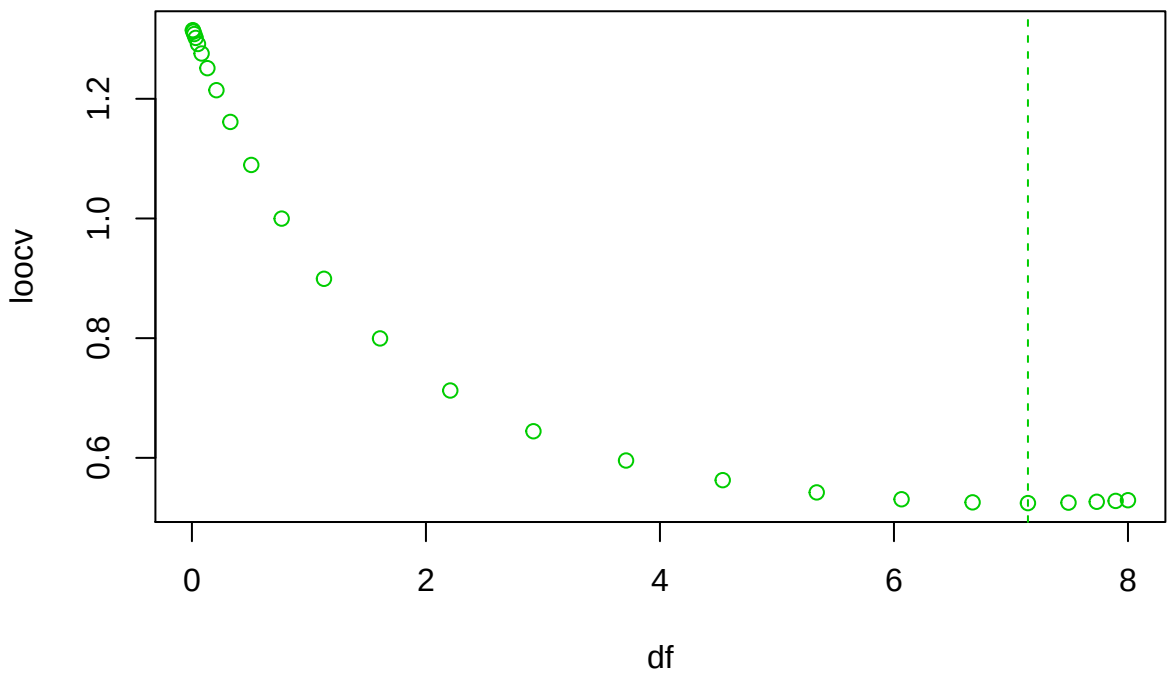


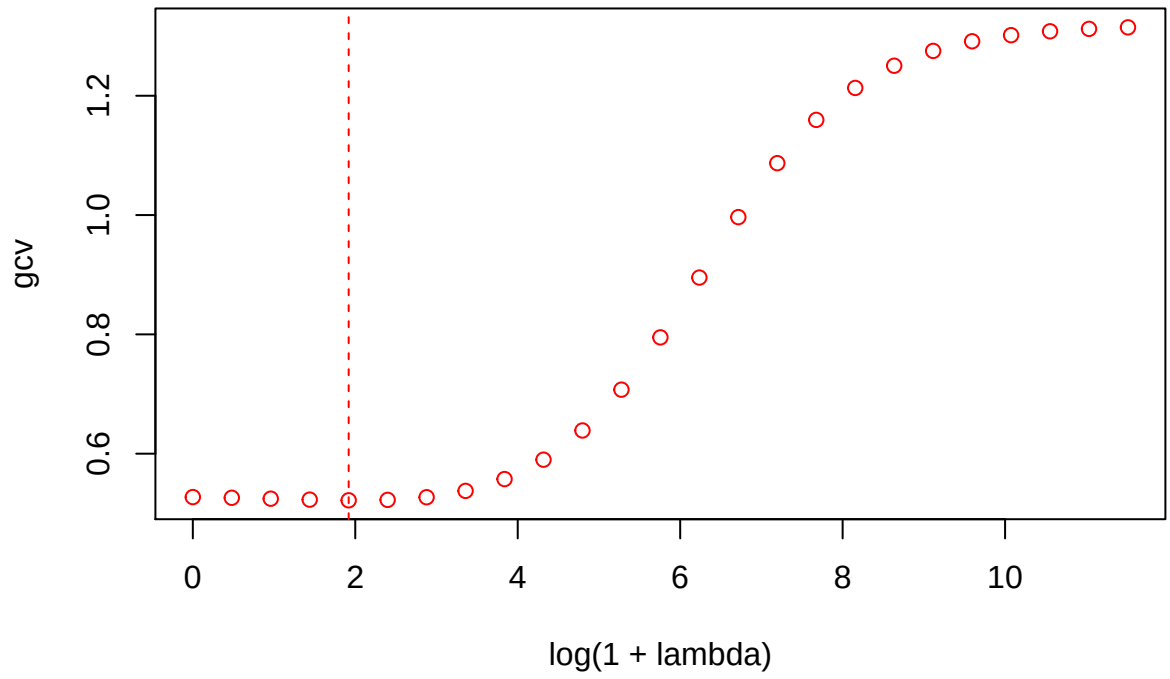

```
## With LOOCV (from n-CV and estimate)  and GCV estimate respectively.
```

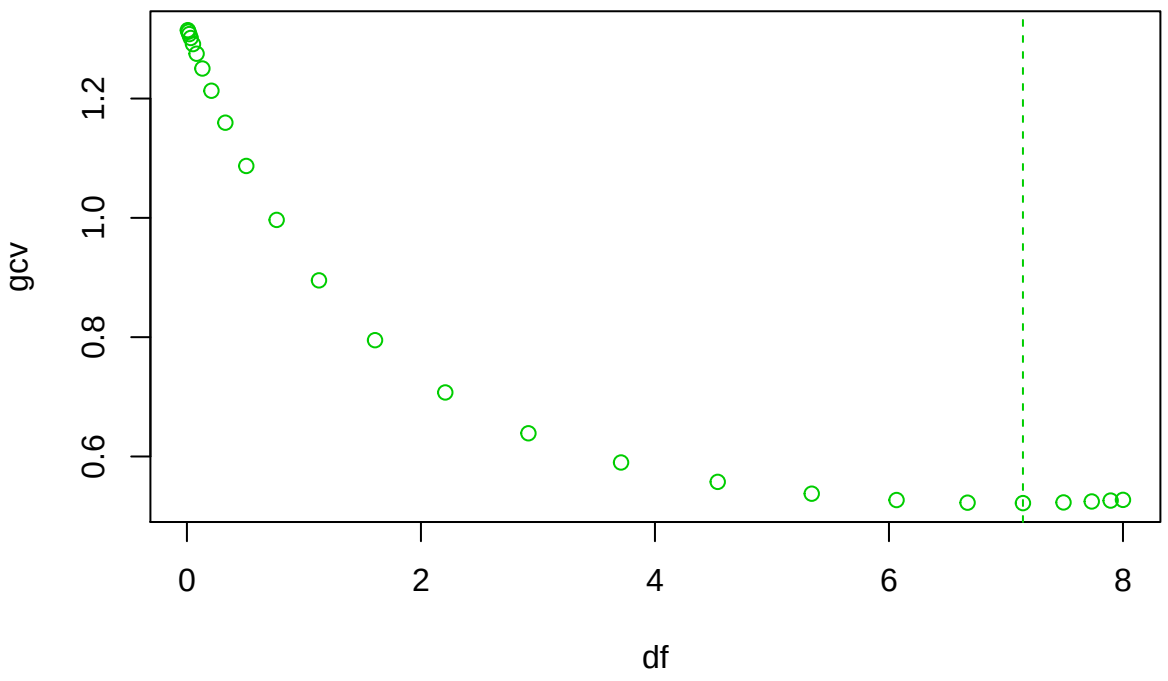



mspe
0.6
0.8
1.0
1.2
0
2
4
6
8
df


loocv
0.6
0.8
1.0
1.2
0
2
4
6
8
10
log(1 + lambda)

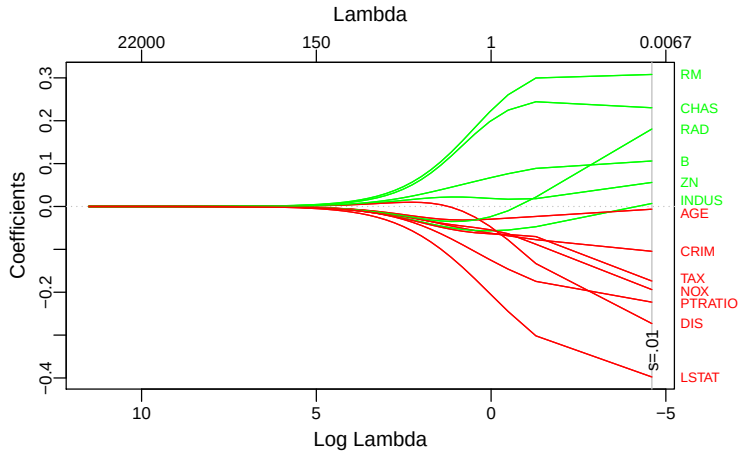

loocv
0.6
0.8
1.0
1.2
0
2
4
6
8
df


gcv
0.6
0.8
1.0
1.2
0
2
4
6
8
10
log(1 + lambda)



Ridge regression for the Boston Housing data

We start by scaling and splitting the Boston dataset to training and test using a 2/3 ratio. Since *CHAR* is a factor variable we do not include it in the *scale* function. First we need to tune the parameter λ. To do this we use 10 fold corss validation performed by *cv.glmnet*.



To select the best model, we now use 10x10-CV using the lambda that best minimised the error in cross-validation, which is 0.0100502.

So our final model has Df= 13 which is the number of non-zero coefficients and $\%Dev$= 0.759315 is the percent deviance explained, which is quite good.

In terms of interpreting the coefficients, we observe that each additional room (RM) is associated with an

increase in the house price, on average. This is quite straightforward, in principle, since it is to be expected that the larger the house, loosely speaking, the more expensive it will be. In addition, we see that an increase in *RAD* (index of accessibility to radial highways) is associated with an increse in *MEDV*. So basically, if we were to think of the town as a graph we would be capturing the connectivity degree of a specific suburb; so a remote node would have a lower value. Moreover, an increase in *CHAS* would mean that it will take the value of 1 is associate with an increase in *MEDV*, so ultimately if the Charles River passes through this suburb then this signals a higher house price, on average.

On the other hand, an increase in *LSTAT* (% lower status of the population) is associated with a decrease in the house price, on average. Most interestinglty though is that the increase in *PTRATIO* (pupil-teacher ratio by town) is associated with a decrease in the house price, on average. So in other words, the education offering of a town increases its value. Also, an increase in *DIS* (weighted distances to five Boston employment centres) is associated with a decrease in *MEDV*, on average. So, having to do a larger commute to work signals a lower house price. Another reasonable result is the fact that an increase in *NOX* (nitric oxides concentration) is associated with a decrease in *MEDV*, so air pollution is a detractor to house price.

Furthermore, we see that neither *AGE* (proportion of owner-occupied units built prior to 1940) nor *INDUS* (roportion of non-retail business acres per town) seem to have a considerable effect on *MEDV*.

Finally, we obtain the train and test error.

Table 1: Model Errors Summary

regression_method	train_error	test_error
ridge regression	0.245	0.309

The difference between train and test errors is not that large, even if the test set is relatively small, and thus subject to a great variance.