MA8701 Advanced methods in statistical inference and learning

Example Data Analysis Report Project 1

Mette Langaas IMF/NTNU

23 January, 2021

Contents

Preface

These are my suggestions for what the Data Analysis Project 1 report could look like. I am worried that you spend too much work on the report. The main aim is to get hands on experience with the topics of Part 1. I would not expect that you spend many days of working on the report.

If a group hands in a report which get a fail grade, they will get feedback and will be able to resubmit the report.

Introduction

Here you describe what is the aim of the analysis.

The aim of the analysis is to find an interpretable model for estimating the level of prostate antigen PSA from 8 clinical measurements. Data are available from 97 males who were about to receive ratical prostatectomy.

I will use the standard data set on prostate cancer used in the ELS book - to show what a minimal solution could be.

Data

_Write

- a few words on the data set to be used and on
- preprocessing of the data. Hopefully that is not very much, mainly maybe centering of reponse (if regression) and standardize the covariates._

The data is presented in ELS page Example 2, pages 3-4, and downloaded from http://statweb.stanford.edu /~tibs/ElemStatLearn.1stEd/ with information in the file `prostate.info` and data in `prostate.data`.

Response:

- the log of PSA (lpsa)

Covariates:

- log cancer volume (lcavol)[continuous]
- log prostate weight (lweight)[continuous]
- age in years (age)[continuous]
- log of benign prostatic hyperplasia amount (lbph)[continuous]
- seminal vesicle invasion(svi) [binary: 0, 1]
- log of capsular penetration (lcp)[continuous]
- Gleason score (gleason) [ordinal: 6,7,8,8]
- percent of Gleason scores 4 or 5 (pgg45)[continuous]

Plan for statistical analyses

Write down what your plan is.

The data contains a column with information for use as traning and test data, with 67 observations for training and 30 for testing. The test data will be set aside for evaluation?

For model selection cross-validation will be used. Since interpretation is the prime interest and the data set is very small, we will not set aside a test set, and will therefore not have focus on comparing goodness of fit across different models.

We will start by presenting pairs plots of the response and covariates, together with summary statistics.

This is a regression problem.

1) As a baseline model we will fit a linear regression model to the data, with least squares. Firdt no variable selection will be performed, but the fitted model with be presented with confidence intervals and Wald p-values for regression coefficients. The the goodness of the model may be evaluated by the proportion of variability explained on the training data.
2) The second model to investigate is the ridge regression. 10-fold cross-validation will be used for model selection. The same strategy as for best subset selection will be used for model selection and regression coefficient presentation.
3) The third model is the lasso regression. The same strategy as for the best subset selection will be used for model selection and regression coefficient presentation.

The boostrapping will be performed with a loop jointly across all three analyses, after each of the analyses are presented.

Statistical analyses

Reading data

```
## [1] 97 11
```

```
##  [1] "X"       "lcavol"  "lweight" "age"     "lbph"    "svi"     "lcp"
##  [8] "gleason" "pgg45"   "lpsa"    "train"
```

```
##   X     lcavol  lweight age      lbph svi       lcp gleason pgg45       lpsa
## 1 1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0 -0.4307829
## 2 2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0 -0.1625189
## 3 3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20 -0.1625189
## 4 4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0 -0.1625189
## 5 5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0  0.3715636
## 6 6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0  0.7654678
##   train
## 1  TRUE
## 2  TRUE
## 3  TRUE
## 4  TRUE
## 5  TRUE
## 6  TRUE
```

Decriptive statistics

Traning data, sample size 67

lcavol

lweight

age

lbph

svi

lcp

gleason

pgg45

lpsa

min

-1.35

2.37

41.00

-1.39

0.00

-1.39

6.00

0.00

-0.43

median

1.47

3.60

65.00

-0.05

0.00

-0.80

7.00

15.00

2.57

mean

1.31

3.63

64.75

0.07

0.22

-0.21

6.73

26.27

2.45

max

3.82

4.78

79.00

2.33

1.00

2.66

9.00

100.00

5.48

sd

1.24

0.48

7.50

1.46

0.42

1.40

0.71

29.30

1.21

Test data, sample size 30

lcavol

lweight

age

lbph

svi

lcp

gleason

pgg45

lpsa

min

-0.78

2.87

43.00

-1.39

0.00

-1.39

6.00

0.00

0.77

median

1.44

3.65

64.00

0.44

0.00

-0.43

7.00

8.00

2.59

mean

1.43

3.71

61.90

0.16

0.20

-0.10

6.80

20.17

2.54

max

3.47

6.11

70.00

2.17

1.00

2.90

9.00

90.00

5.58

sd

1.04

0.54

7.04

1.44

0.41

1.41

0.76

25.55

1.04

Center reponses and standardize all covariates - will not use as factors but numeric for the two ordinal and binary covariates.

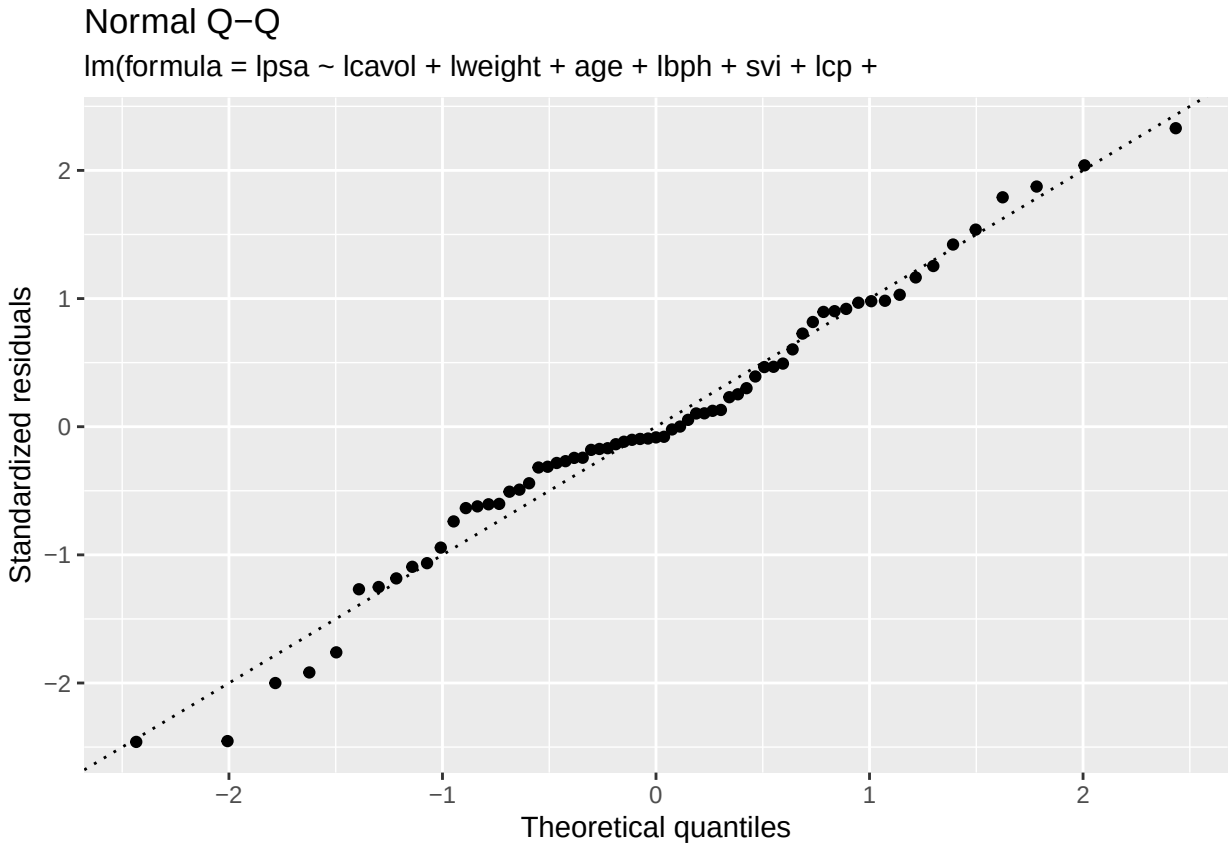
Least squares model fit

Add comments to what you do and what you find.

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45, data = strain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64870 -0.34147 -0.05424  0.44941  1.48675
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.347e-16  8.702e-02   0.000  1.00000
## lcavol       7.164e-01  1.335e-01   5.366 1.47e-06 ***
## lweight      2.926e-01  1.064e-01   2.751  0.00792 **
## age         -1.425e-01  1.021e-01  -1.396  0.16806
## lbph         2.120e-01  1.031e-01   2.056  0.04431 *
## svi          3.096e-01  1.254e-01   2.469  0.01651 *
## lcp         -2.890e-01  1.548e-01  -1.867  0.06697 .
## gleason     -2.091e-02  1.426e-01  -0.147  0.88389
## pgg45        2.773e-01  1.596e-01   1.738  0.08755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7123 on 58 degrees of freedom
## Multiple R-squared:  0.6944, Adjusted R-squared:  0.6522
## F-statistic: 16.47 on 8 and 58 DF,  p-value: 2.042e-12
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.174188643 0.17418864
## lcavol       0.449175029 0.98363900
## lweight      0.079689925 0.50559488
## age         -0.346964156 0.06186490
## lbph         0.005581926 0.41843328
## svi          0.058624481 0.56061459
## lcp         -0.598879566 0.02086834
## gleason     -0.306314488 0.26448745
## pgg45       -0.042112877 0.59680478
```

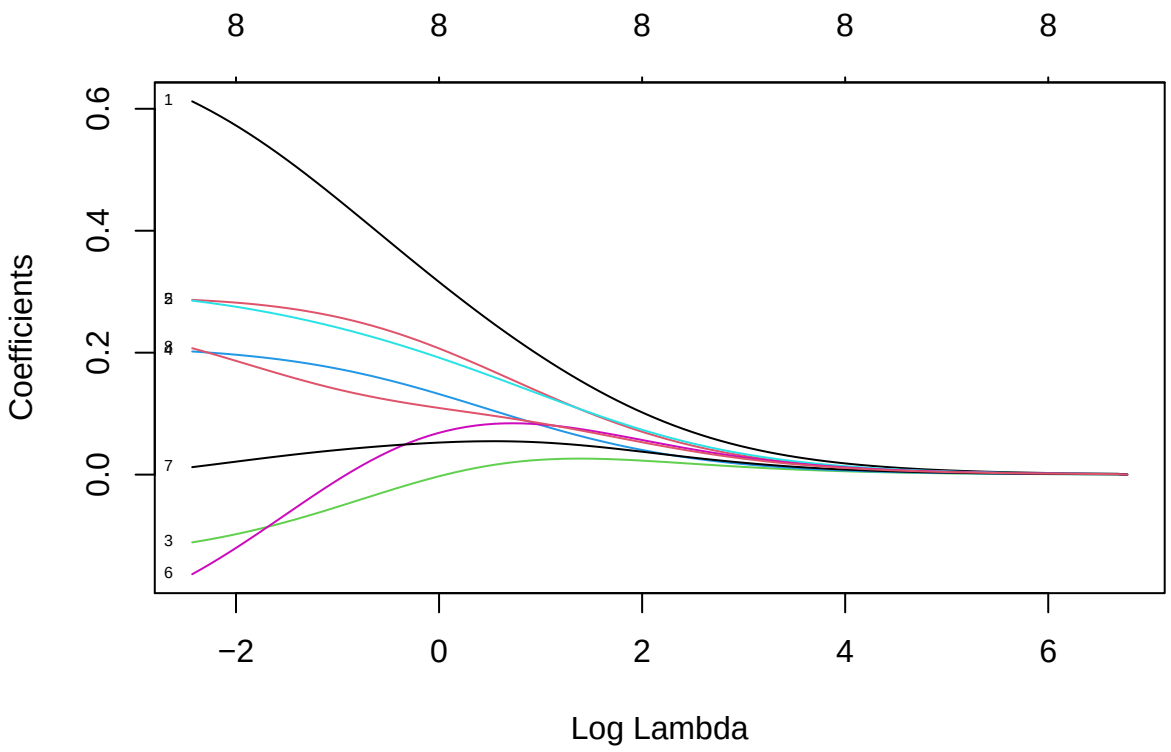



```
##
##  Anderson-Darling normality test
##
## data:  rstudent(full)
## A = 0.57619, p-value = 0.1293
```

Ridge regression

Add comments to what you do and what you find.

```
ridgefit=glmnet(x=strainx,y=strainy,alpha=0)
plot(ridgefit,xvar="lambda",label=TRUE)
```



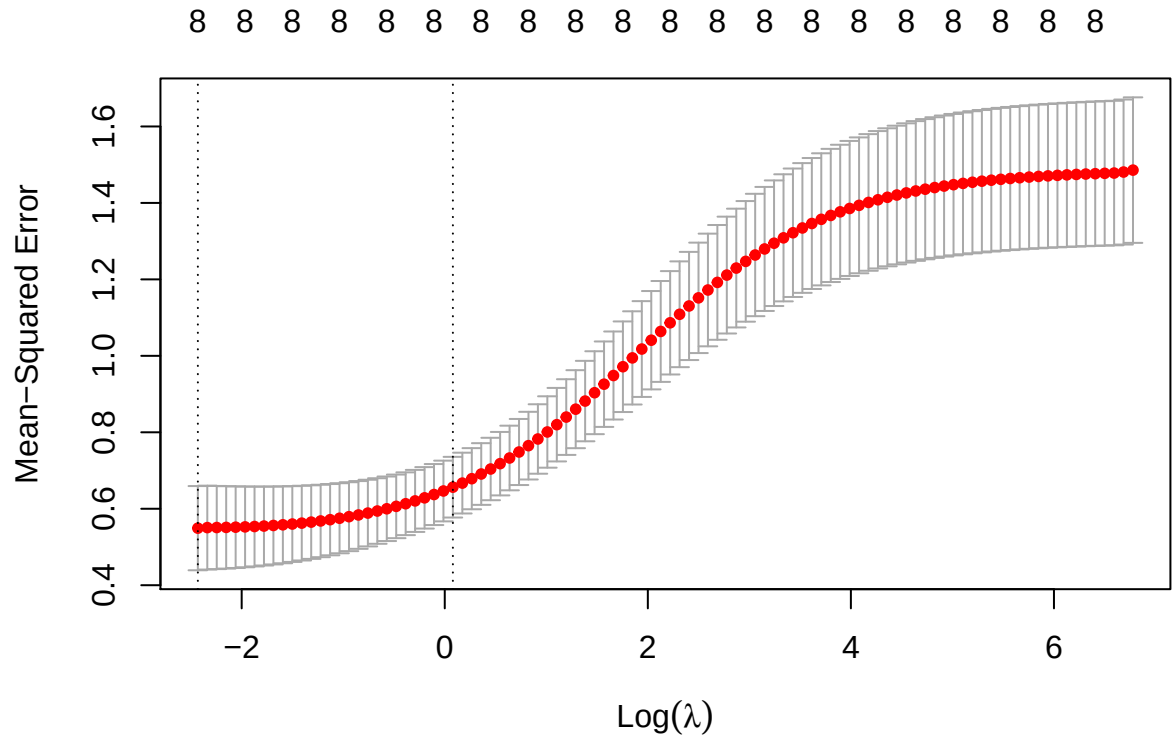

```
cv.ridge=cv.glmnet(x=strainx,y=strainy,alpha=0)
print(paste("The lamda giving the smallest CV error",cv.ridge$lambda.min))
```

```
## [1] "The lamda giving the smallest CV error 0.08788804212004"
```

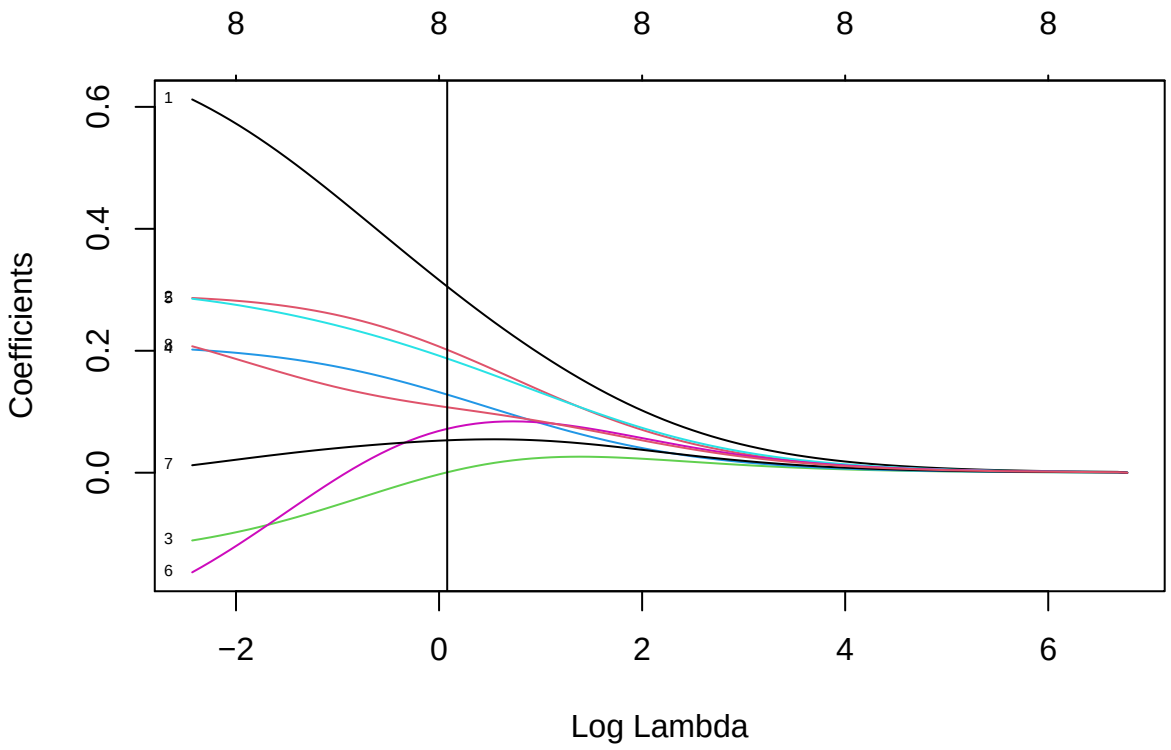
```
print(paste("The 1sd err method lambda",cv.ridge$lambda.1se))
```

```
## [1] "The 1sd err method lambda 1.08352485910158"
```

```
plot(cv.ridge)
```

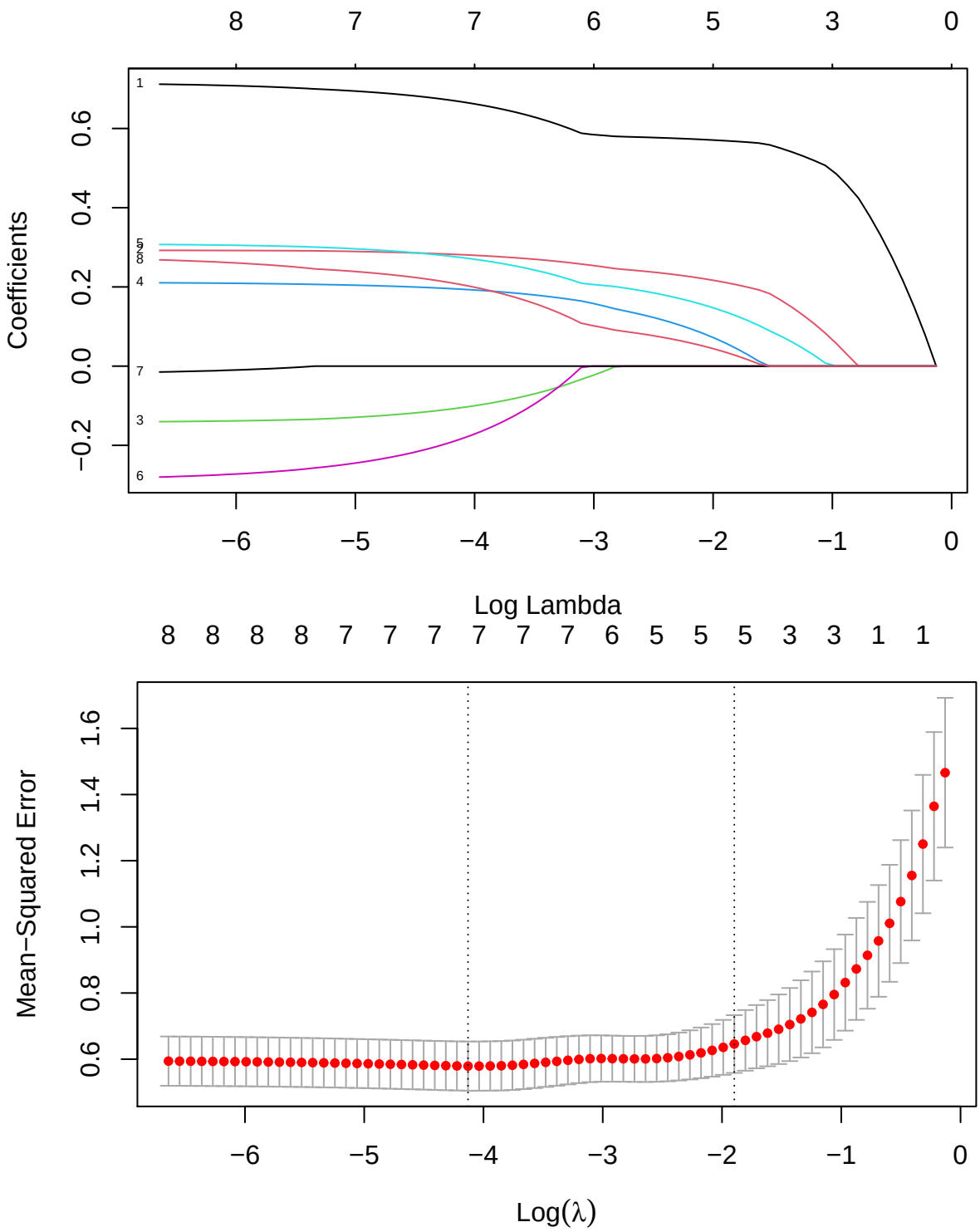


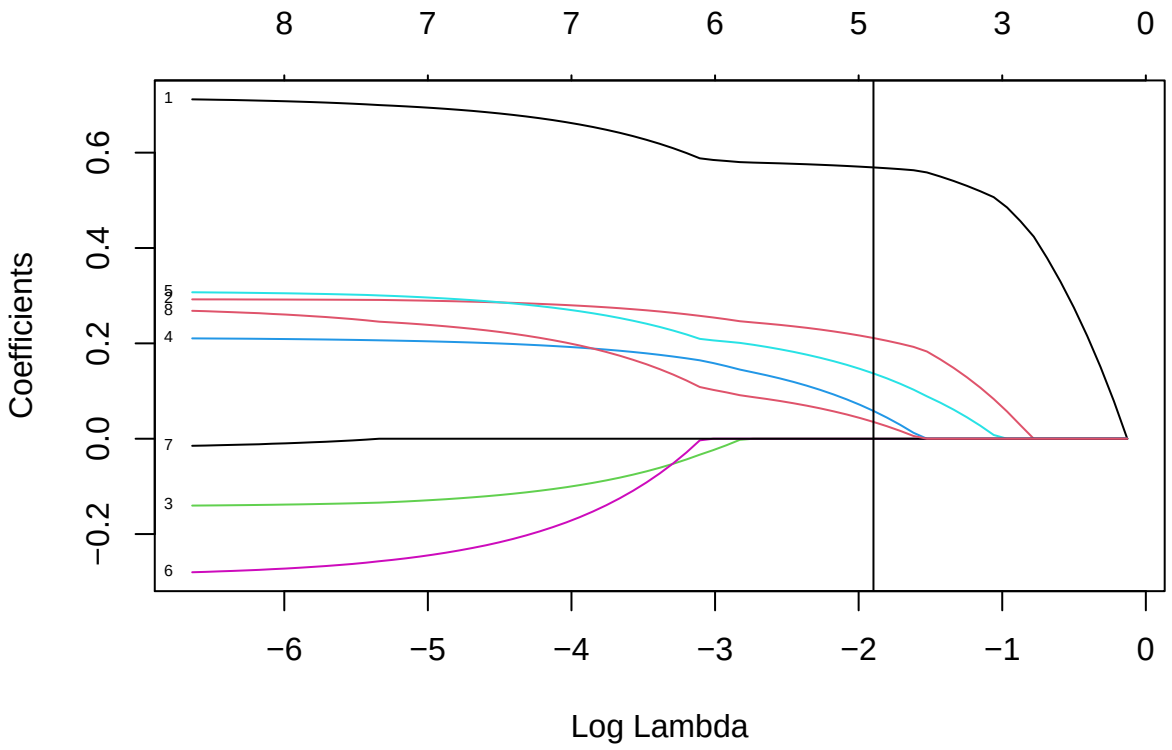

```
# use 1sd error rule default
plot(ridgefit,xvar="lambda",label=TRUE);
abline(v=log(cv.ridge$lambda.1se));
```



Lasso regression

Add comments to what you do and what you find.


8 7 7 6 5 3 0
Coefficients
0.6
0.4
0.2
0.0
−0.2
1
5
2
8
4
7
3
6
−6 −5 −4 −3 −2 −1 0
Log Lambda
8 8 8 8 7 7 7 7 7 7 6 5 5 5 3 3 1 1
Mean−Squared Error
1.6
1.4
1.2
1.0
0.8
0.6
−6 −5 −4 −3 −2 −1 0
Log(λ)



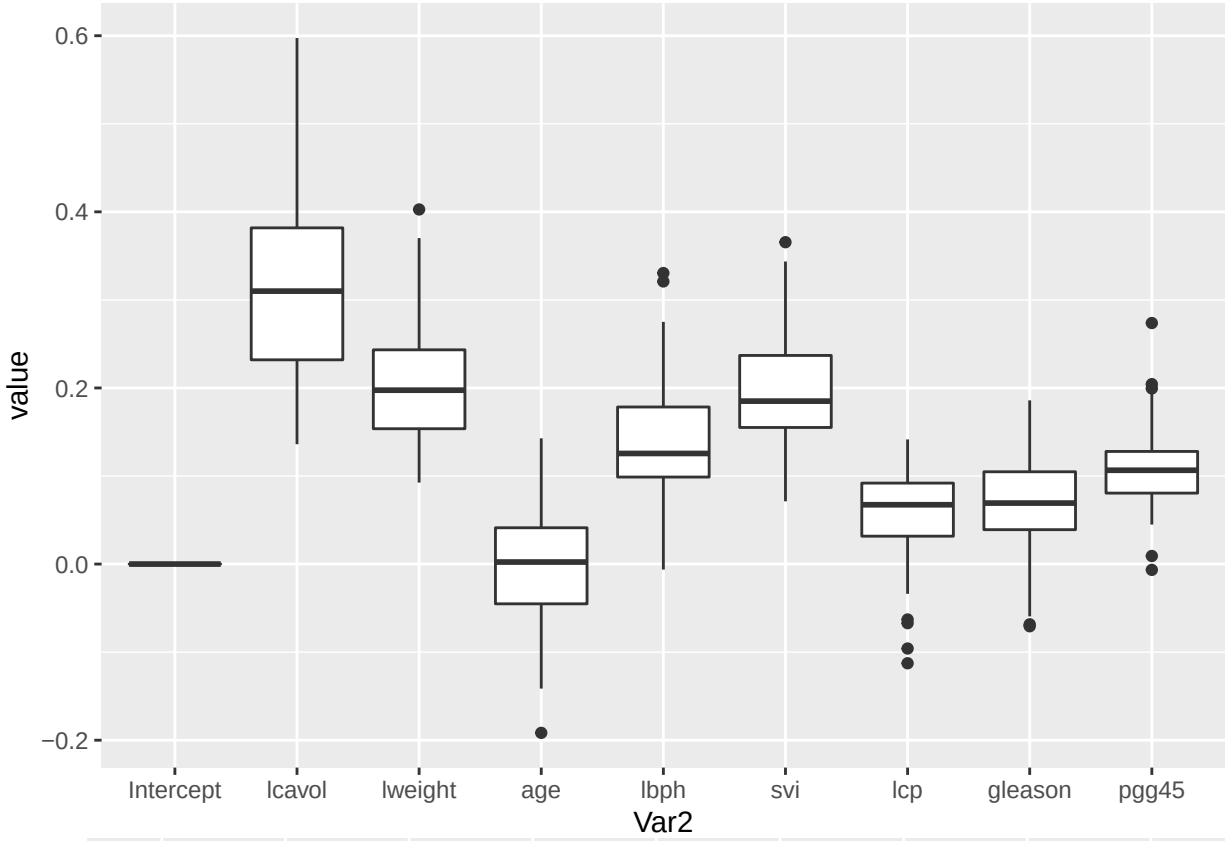

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept) -5.112829e-17
## lcavol       5.690116e-01
## lweight      2.111485e-01
## age          .
## lbph         5.829639e-02
## svi          1.369689e-01
## lcp          .
## gleason      .
## pgg45        3.522616e-02
```

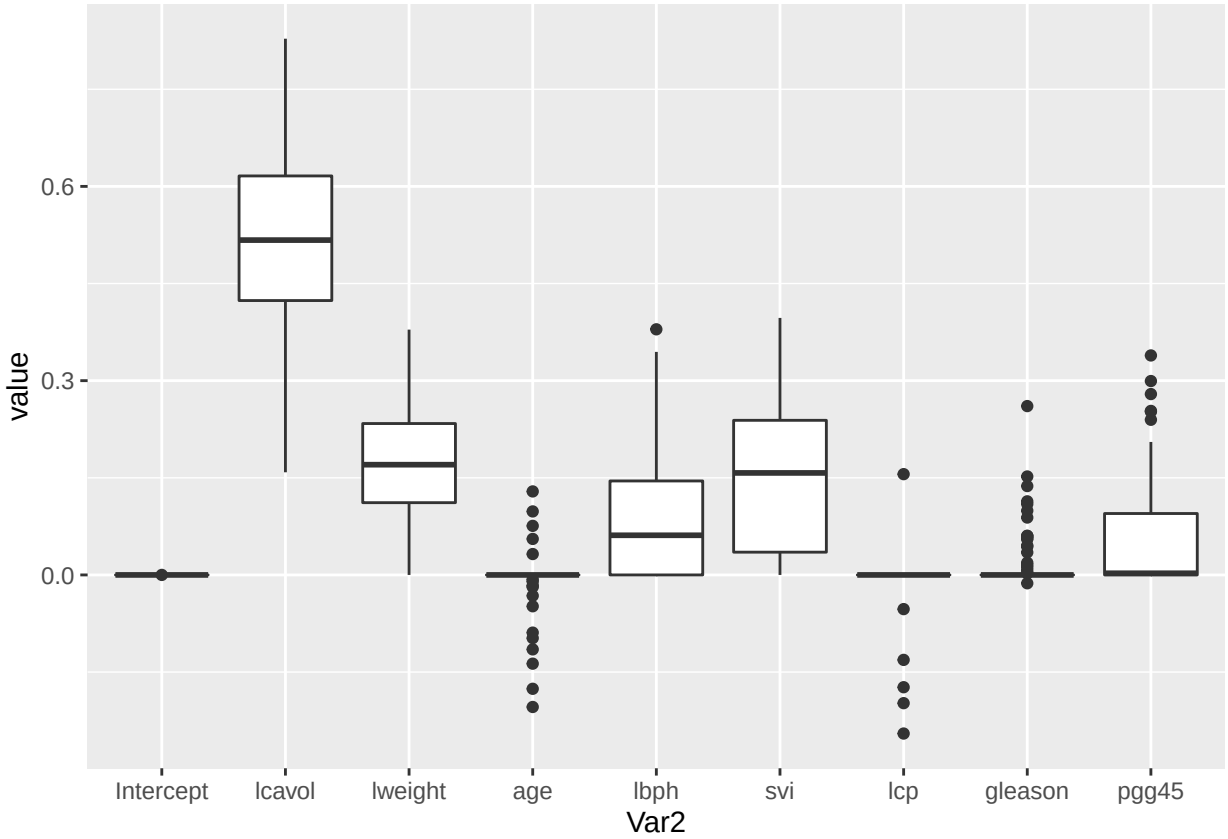
Inference for the model selection procedures

under construction

We now turn to bootstrapping for running an outer loop to get confidence intervals for regression parameters for the methods where model selection is used.

Should the lambda grid be the same for the loop or different for each in loop? HTW outside, else inside?






```
##           Intercept    lcavol   lweight          age       lbph        svi
```

```
## 2.5%  -3.128341e-16 0.1471355 0.1032684 -0.130755005 0.04730076 0.09442785
## 50%    3.779317e-17 0.3099514 0.1974597  0.002326308 0.12558689 0.18506050
## 97.5%  2.969954e-16 0.5125144 0.3425933  0.124033136 0.27411376 0.31936769
##               lcp     gleason      pgg45
## 2.5%  -0.06511798 -0.04749396 0.04528144
## 50%    0.06734899  0.06927480 0.10653913
## 97.5%  0.13440402  0.15066705 0.19763541
```

```
##           Intercept    lcavol     lweight         age       lbph       svi
## 2.5%  -2.717079e-16 0.2471296 0.003977292 -0.12656186 0.00000000 0.0000000
## 50%    4.160033e-17 0.5170494 0.170255962  0.00000000 0.06125431 0.1575293
## 97.5%  2.713226e-16 0.7853046 0.354220752  0.06620787 0.27924373 0.3659929
##              lcp   gleason       pgg45
## 2.5%  -0.1533521 0.0000000 0.000000000
## 50%    0.0000000 0.0000000 0.002598508
## 97.5%  0.0000000 0.1260287 0.266936884
```

```
##                    2.5%          50%        97.5%          2.5%          50%
## Intercept -3.128341e-16 3.779317e-17 2.969954e-16 -2.717079e-16 4.160033e-17
## lcavol     1.471355e-01 3.099514e-01 5.125144e-01  2.471296e-01 5.170494e-01
## lweight    1.032684e-01 1.974597e-01 3.425933e-01  3.977292e-03 1.702560e-01
## age       -1.307550e-01 2.326308e-03 1.240331e-01 -1.265619e-01 0.000000e+00
## lbph       4.730076e-02 1.255869e-01 2.741138e-01  0.000000e+00 6.125431e-02
## svi        9.442785e-02 1.850605e-01 3.193677e-01  0.000000e+00 1.575293e-01
## lcp       -6.511798e-02 6.734899e-02 1.344040e-01 -1.533521e-01 0.000000e+00
## gleason   -4.749396e-02 6.927480e-02 1.506671e-01  0.000000e+00 0.000000e+00
## pgg45      4.528144e-02 1.065391e-01 1.976354e-01  0.000000e+00 2.598508e-03
##                   97.5%
## Intercept 2.713226e-16
## lcavol    7.853046e-01
## lweight   3.542208e-01
## age       6.620787e-02
## lbph      2.792437e-01
## svi       3.659929e-01
## lcp       0.000000e+00
## gleason   1.260287e-01
## pgg45     2.669369e-01
```

Discussion

Strenghts

Weaknesses

Try out gleason as factor? How about standardization of the factor?

References

- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.