Regularization Methods for Linear Regression

Variable selection et regularisation

M1 Math et Interactions – UEVE/ENSIIE

Autumn semester 2016

http://julien.cremeriefamily.info/teachings_M1MINT_Reg.html

ensiie
école nationale supérieure d'informatique pour l'industrie et l'entreprise

université evry val-d'essonne

Outline

Outline

Outline

Statistical Learning

Canonical scenario

1. an outcome measurement (or response, output)
   - either quantitative (expression level, tumor size, survival time, etc.)
   - or categorical (presence/absence of a gene or of a disease, etc.)
2. a set of features (or predictors, inputs)
   - clinical measurements (expression level, tumor size)
   - age, smoking or not, height, SNPs, etc.

Learning problem

Given a training set of data (observed inputs and outputs), we aim to

1. suggest a model,
2. learn this model on the training set,
3. test this model on new outcomes/features.

$\rightsquigarrow$ A "good" model should accurately predict new outcomes.

Notations

Let

- Y be the output random variable,
- $X = (X_1, \ldots, X_p)$ be the input random variables, where X_j is the j predictor.

The data

Given a sample $\{(y_i, x_i), i = 1, \ldots, n\}$ of i.id. realizations of (Y, X), denote

- $\mathcal{D} = \{i : (y_i, x_i) \in$ training set$\}$,
- $\mathcal{T} = \{i : (y_i, x_i) \in$ test set$\}$,
- $\mathbf{y} = (y_i)_{i \in \mathcal{D}}$, the response vector in $\mathbb{R}^{|\mathcal{D}|}$,
- $\mathbf{x}_j = (x_{ij})_{i \in \mathcal{D}}^{\mathsf{T}})$ the vector of data for the jth predictor in $\mathbb{R}^{|\mathcal{D}|}$,
- $\mathbf{X}$ the $n \times p$ data (or design) matrix on the training set whose jth row is $\mathbf{x}_j$,
- $(\mathbf{y}_{\mathcal{T}}, \mathbf{X}_{\mathcal{T}})$ are the test data.

Regression models

We seek a function f that predicts Y through X.

Proposition

The model $f(X) = \mathbb{E}[\,Y|X]$ *minimizes the squared error loss, that is,*

$$f(X) = \arg\min_{\varphi} \operatorname{err}(\varphi(X)), \quad \textit{with } \operatorname{err}(\varphi(X)) = \mathbb{E}[(\,Y - \varphi(X))^2].$$

$\rightsquigarrow$ *The best prediction of* Y *at any point* $X = x$ *is the conditional mean, when best is measured by average squared error.*

This leads to the regression model

$$Y = f(X) + \varepsilon,$$

where

- ε is an additive error with $\mathbb{E}[\varepsilon] = \mathbf{0}$, $\mathbb{V}[\varepsilon] = \sigma^2$,
- $f(x) = \mathbb{E}[\,Y|X = x]$ is the regression function.

Learning strategy

Problem

$\mathbb{P}(Y|X)$ and $\mathbb{P}(X)$ are unknown thus $\mathbb{E}(Y|X), \mathrm{err}(f(X))$ unreachable: one should estimate this.

Strategy

1. Fix a family $\mathcal{F}$ of models

   For the linear model, $\mathcal{F} = \left\{X^T\beta, \beta \in \mathbb{R}^p\right\}$.

2. Fit a model $\hat{f} \in \mathcal{F}$ on the training set $\mathcal{D}$

   With the least square, compute $\hat{\beta}^{\mathrm{ols}}$ *and* $\hat{f} = \hat{Y} = \mathbf{X}\hat{\beta}^{\mathrm{ols}}$

3. Estimate the prediction error with the test set $\mathcal{T}$.

   For instance, $\widehat{\mathrm{err}}(\mathbf{X}_{\mathcal{T}}\hat{\beta}^{\mathrm{ols}}) = \frac{1}{n}\left\|\mathbf{y}_{\mathcal{T}} - \mathbf{X}_{\mathcal{T}}\hat{\beta}^{\mathrm{ols}}_{\mathcal{D}}\right\|^2$.

Learning strategy

Problem

$\mathbb{P}(Y|X)$ and $\mathbb{P}(X)$ are unknown thus $\mathbb{E}(Y|X), \mathrm{err}(f(X))$ unreachable: one should estimate this.

Strategy

1. Fix a family $\mathcal{F}$ of models
   For the linear model, $\mathcal{F} = \left\{X^T\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^p\right\}$.

2. Fit a model $\hat{f} \in \mathcal{F}$ on the training set $\mathcal{D}$
   With the least square, compute $\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$ *and* $\hat{f} = \hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$

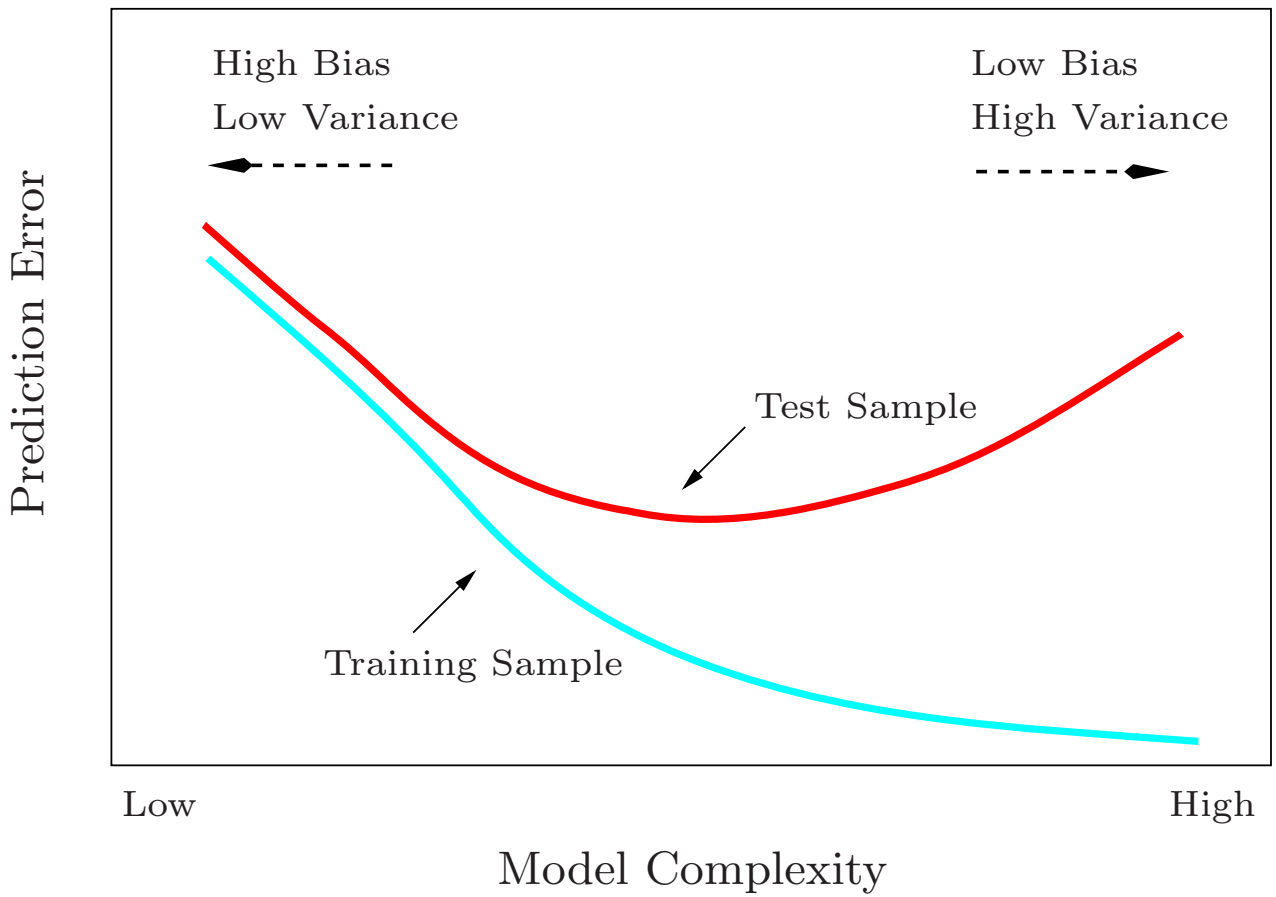
3. Estimate the prediction error with the test set $\mathcal{T}$.

$$\text{For instance, } \hat{\mathrm{err}}(\mathbf{X}_{\mathcal{T}}\hat{\boldsymbol{\beta}}^{\mathrm{ols}}) = \frac{1}{n}\left\|\mathbf{y}_{\mathcal{T}} - \mathbf{X}_{\mathcal{T}}\hat{\boldsymbol{\beta}}^{\mathrm{ols}}_{\mathcal{D}}\right\|^2.$$

Bias/variance tradeoff

At an input point $X = x$,

$$\mathrm{err}(\hat{f}(x)) = \underbrace{\sigma^2}_{\substack{\text{incompressible} \\ \text{error}}} + \underbrace{\mathsf{bias}^2(\hat{f}(x)) + \mathbb{V}(\hat{f}(x))}_{\mathrm{MSE}(\hat{f}(x))}.$$



Linear regression

Prediction error

For a fixed $\mathbf{X}$, one has

$$\hat{\mathrm{err}}(\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ols}}) = \sigma^2 \frac{(p+1)}{n} + \sigma^2.$$

Gauss-Markov Theorem

$\hat{Y} = X^\intercal \hat{\boldsymbol{\beta}}^{\mathrm{ols}}$ is the BLUE: the best model (i.e. with the smallest variance) among unbiased estimators of $\boldsymbol{\beta}$.

$\rightsquigarrow$ Are they some cases where we should **trade some bias for smaller variance** ?

Outline

Collinearity in OLS: Gram-Schmidt procedure (I)

Regression by successive orthogonalizations

Gram-Schmidt orthogonalization

S0 Initialization

$\mathbf{z}_0 \leftarrow \mathbf{x}_0 (= \underline{1}_p)$;

S2 Regressions on an orthonormal basis

for $j = 1, \ldots, p$ **do**

  for $k = 1, \ldots, j-1$ **do**

    Regress $\mathbf{x}_j$ on $\mathbf{z}_k$

    $\gamma_{kj} \leftarrow \frac{\mathbf{z}_k^T \mathbf{x}_j}{\mathbf{z}_k^\intercal \mathbf{z}_k}$

  Update the residual $\mathbf{z}_j$

  $\mathbf{z}_j \leftarrow \mathbf{x}_j - \sum_{\ell=0}^{j-1} \gamma_{\ell k} \mathbf{z}_{\ell-1}$

S3 Compute the estimate $\hat{\beta}_p$

$\hat{\beta}_p \leftarrow \frac{\mathbf{z}_p^\intercal \mathbf{y}}{\mathbf{z}_p^\intercal \mathbf{z}_p}$.

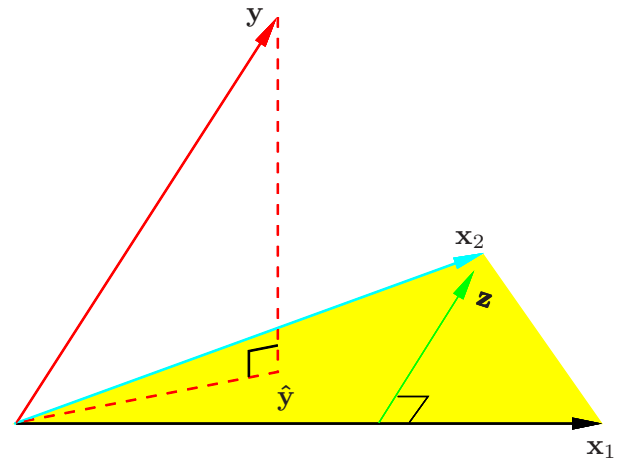


Figure: Example with two predictors

Step 2 can be written as (with $\mathbf{D}$ diagonal so as $\mathbf{D}_{jj} = \mathbf{z}_j^\intercal \mathbf{z}_j$)

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{\Gamma} = \mathbf{Q}\mathbf{R},$$

with $\mathbf{Q}$ orthogonal column-wise and $\mathbf{R}$ upper triangular.

Collinearity in OLS: Gram-Schmidt (II)

Insights brought by the QR factorization

Estimator and fitted values via the QR factorization

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = \mathbf{R}^{-1}\mathbf{Q}^{\intercal}\mathbf{y},$$

$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{Q}^{\intercal}\mathbf{y}.$$

We can permute the columns of $\mathbf{X}$ during the Gram-Schmidt orthogonalization, thus

- $\hat{\beta}_j$ is the additional contribution of $\mathbf{x}_j$ on $\mathbf{y}$ once $\mathbf{x}_j$ has been adjusted,
- The variance of $\hat{\beta}_p$ can be written

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_p) = \frac{\sigma^2}{\|\mathbf{z}_p\|_2^2},$$

$\rightsquigarrow$ thus collinear predictors lead to **bad estimation of β.**.

Collinearity in OLS: interpretability (I)

Conditional dependency: no **direct links** between variables

X and Y are independent conditional on Z ($X \perp\!\!\!\perp Y|Z$) iff

$$\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z) \times \mathbb{P}(Y|Z).$$

Partial covariance/correlation

Its is the covariance once removed the effect of another variable

$$\text{cov}(X, Y|Z) = \text{cov}(X, Y) - \text{cov}(X, Z)\text{cov}(Y, Z)/\mathbb{V}(Z),$$

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}.$$

Gaussian case

If X, Y, Z are jointly Gaussian, then

$$\text{cov}(X, Y|Z) = 0 \Leftrightarrow \text{cor}(X, Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

Collinearity in OLS: interpretability (I)

Conditional dependency: no **direct links** between variables

X and Y are independent conditional on Z ($X \perp\!\!\!\perp Y|Z$) iff

$$\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z) \times \mathbb{P}(Y|Z).$$

Partial covariance/correlation

Its is the covariance once removed the effect of another variable

$$\text{cov}(X, Y|Z) = \text{cov}(X, Y) - \text{cov}(X, Z)\text{cov}(Y, Z)/\mathbb{V}(Z),$$

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}.$$

Gaussian case

If X, Y, Z are jointly Gaussian, then

$$\text{cov}(X, Y|Z) = 0 \Leftrightarrow \text{cor}(X, Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

Collinearity in OLS: interpretability (I)

Conditional dependency: no **direct links** between variables

X and Y are independent conditional on Z ($X \perp\!\!\!\perp Y|Z$) iff

$$\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z) \times \mathbb{P}(Y|Z).$$

Partial covariance/correlation

Its is the covariance once removed the effect of another variable

$$\text{cov}(X, Y|Z) = \text{cov}(X, Y) - \text{cov}(X, Z)\text{cov}(Y, Z)/\mathbb{V}(Z),$$

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}.$$

Gaussian case

If X, Y, Z are jointly Gaussian, then

$$\text{cov}(X, Y|Z) = 0 \Leftrightarrow \text{cor}(X, Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

Collinearity in OLS: interpretability (II)

Assume that (X, Y) is a Gaussian vector in the linear model

$$Y = X^\intercal \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Then

$$Y = \sum_{j=1}^{p} X_j \mathrm{cor}(X_j, Y | X_k, k \neq j) \frac{\sigma}{\sqrt{\mathbb{V}(X_j)}} + \varepsilon.$$

$\rightsquigarrow$ β_j is proportional to the partial correlation between X_j and Y
i.e. the effect of X_j on Y once removed the other effects.

$$\mathrm{cov}(\hat{\beta}_i^{\mathrm{ols}}, \hat{\beta}_j^{\mathrm{ols}}) \propto -\mathrm{cor}(X_i, X_j | X_k, k \neq i, j),$$

$\rightsquigarrow$ Predictors with strong relationships induce negative covariance on the associated coefficients. . .

Collinearity in OLS: interpretability (II)

Assume that (X, Y) is a Gaussian vector in the linear model

$$Y = X^\intercal \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Then

$$Y = \sum_{j=1}^{p} X_j \mathrm{cor}(X_j, Y | X_k, k \neq j) \frac{\sigma}{\sqrt{\mathbb{V}(X_j)}} + \varepsilon.$$

$\rightsquigarrow$ β_j is proportional to the partial correlation between X_j and Y
i.e. the effect of X_j on Y once removed the other effects.

$$\mathrm{cov}(\hat{\beta}_i^{\mathrm{ols}}, \hat{\beta}_j^{\mathrm{ols}}) \propto -\mathrm{cor}(X_i, X_j | X_k, k \neq i, j),$$

$\rightsquigarrow$ Predictors with strong relationships induce negative covariance on the associated coefficients...

Outline

Example: prostate cancer data set I

The data set: 97 patient with prostate cancer

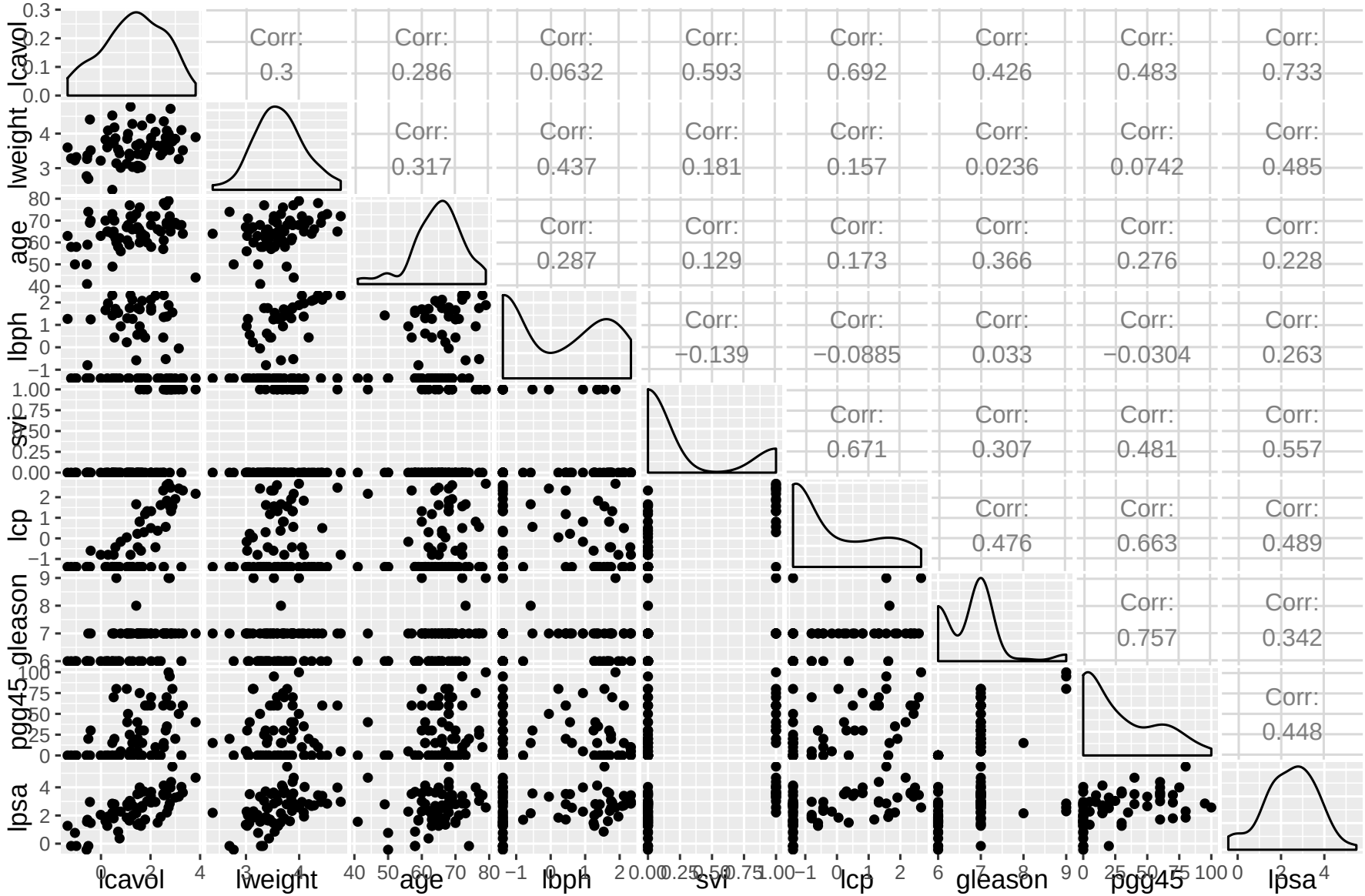
Examine the correlation between the level of cancer-specific antigen ($\mathbf{y}$) and various clinical measures.

```
load("prostate.rda")
dim(prostate)

## [1] 97 10

print(head(prostate), digits=3)

##   lcavol lweight age  lbph svi   lcp gleason pgg45   lpsa train
## 1 -0.580    2.77  50 -1.39   0 -1.39       6     0 -0.431  TRUE
## 2 -0.994    3.32  58 -1.39   0 -1.39       6     0 -0.163  TRUE
## 3 -0.511    2.69  74 -1.39   0 -1.39       7    20 -0.163  TRUE
## 4 -1.204    3.28  58 -1.39   0 -1.39       6     0 -0.163  TRUE
## 5  0.751    3.43  62 -1.39   0 -1.39       6     0  0.372  TRUE
## 6 -1.050    3.23  50 -1.39   0 -1.39       6     0  0.765  TRUE
```

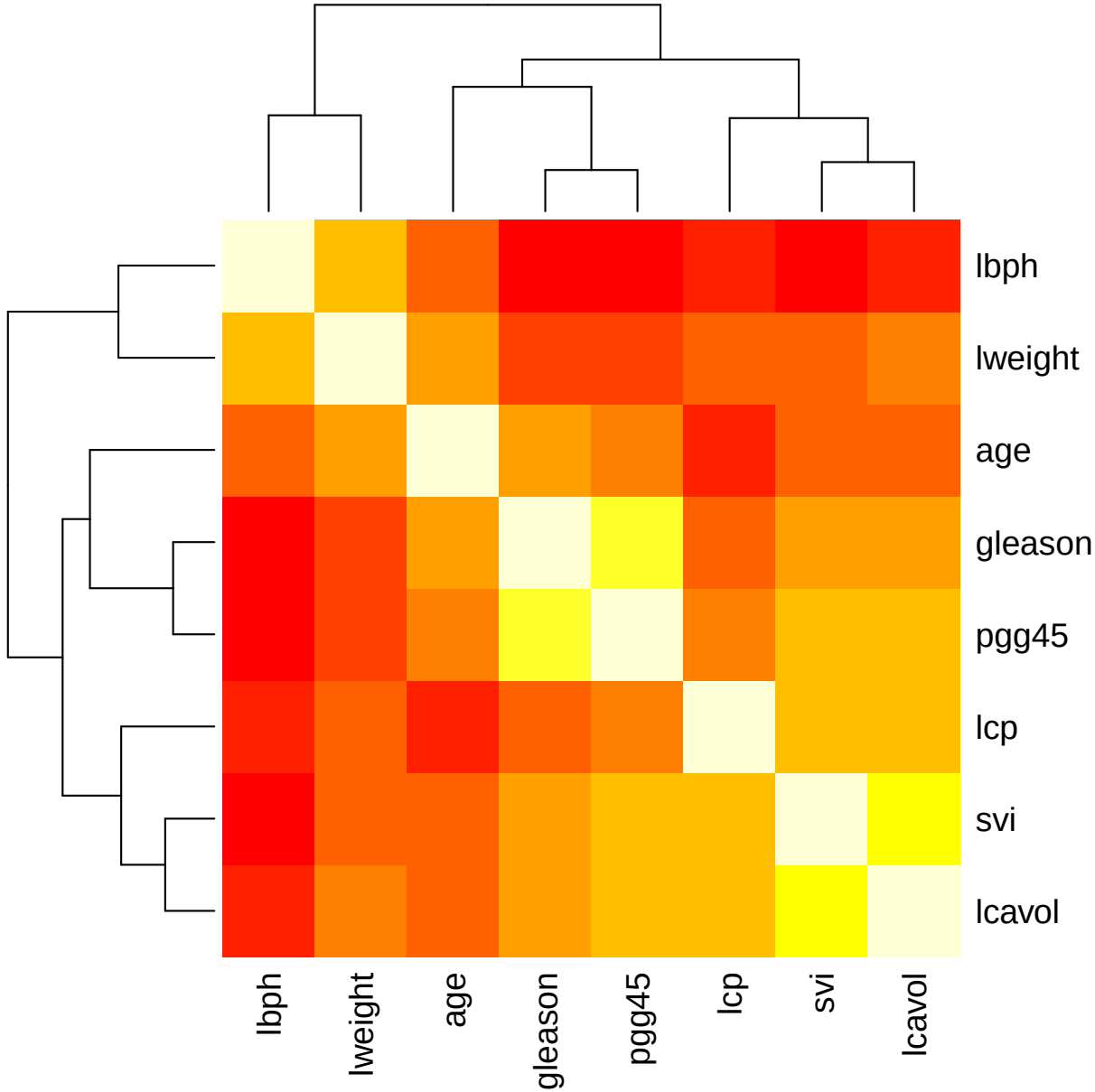
Correlations between predictors I

```
print(as.dist(var(prostate[prostate$train,1:8])),digits=1)

##         lcavol lweight    age   lbph    svi    lcp gleason
## lweight  0.178
## age      2.669   1.132
## lbph     0.115   0.305  3.155
## svi      0.309   0.036  0.406 -0.086
## lcp      1.205   0.105  1.817 -0.182  0.395
## gleason  0.376   0.008  1.946  0.034  0.091  0.473
## pgg45   17.592   1.036 60.630 -1.304  5.924 27.193  15.725

print(as.dist(cor(prostate[prostate$train,1:8])),digits=1)

##         lcavol lweight   age  lbph   svi   lcp gleason
## lweight   0.30
## age       0.29    0.32
## lbph      0.06    0.44  0.29
## svi       0.59    0.18  0.13 -0.14
## lcp       0.69    0.16  0.17 -0.09  0.67
## gleason   0.43    0.02  0.37  0.03  0.31  0.48
## pgg45     0.48    0.07  0.28 -0.03  0.48  0.66    0.76
```

Correlations between predictors II



Correlations between predictors III



OLS and limitations I

For studying the correlation effect, we normalize and create test and train sets
Pour étudier l'effet des corrélations, on ajuste un modèle avec des prédicteurs de variances comparables (normalisées).

```
prostate.train <- subset(prostate, train==TRUE, -train)
prostate.train[, 1:8] <- scale(prostate.train[, 1:8], FALSE, TRUE)
prostate.test <- subset(prostate, train==FALSE, -train)
prostate.test[, 1:8] <- scale(prostate.test[, 1:8], FALSE, TRUE)
model.full <- lm(lpsa~.,prostate.train)
```

Estiamting prediction error

```
y.hat <- predict(model.full, newdata=prostate.test)
y.test <- prostate.test$lpsa
err.ols <- mean((y.test-y.hat)^2)
print(err.ols)

## [1] 0.5221043
```

OLS and limitations II

```
summary(model.full)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate.train)
##
## 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)   0.4292     1.5536   0.276  0.78334
## lcavol        1.0466     0.1950   5.366 1.47e-06 ***
## lweight       2.2623     0.8224   2.751  0.00792 **
## age          -1.2477     0.8938  -1.396  0.16806
## lbph          0.2123     0.1032   2.056  0.04431 *
## svi           0.3515     0.1423   2.469  0.01651 *
## lcp          -0.2924     0.1566  -1.867  0.06697 .
## gleason      -0.2012     1.3716  -0.147  0.88389
## pgg45         0.3737     0.2151   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
```

Comments

Why do some coefficients in $\boldsymbol{\beta}$ are not well estimated/ have large variance? (pgg45, gleason)

Statistical issue

Correlated variables are not well estimated,

- they carry the same information regarding the response.
- Rember that $\text{cov}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j) \propto -\text{cor}(X_i, X_j | X_k, k \neq i, j)$.

Numerical issue

Correlated variables leads no bad conditioning of $\mathbf{X}^T\mathbf{X}$,

- Rember that $\mathbb{V}(\hat{\boldsymbol{\beta}}_p^{\text{ols}}) = \frac{\sigma^2}{\|\mathbf{z}_p\|}$ in the Gram-Schmidt procedure.
- OLS cannot be computed when they are redundant variables in $\mathbf{X}$ or when $n < p$.

$\rightsquigarrow$ interpretation becomes rather difficult

Solutions

Variable selection

If the underlying model is assumed to have only few predictors truly related to the outcome, we may want to select those with the highest effect. We are looking for both

- better interpretability.
- better predictive performances.

Regularization

If all the predictors have similar or close effects on the response, selection (and thus interpretability) is out of reach.
We may regularize the problem by constraining the parameters $\boldsymbol{\beta}$ to live in an appropriate set that will make the $\mathbf{X}^T\mathbf{X}$ invertible.

Outline

Variable Selection

Problematic

With many regressor,

- we integrate more and more information in the model ;
- we have more and more parameters to estimate and $\mathbb{V}(\hat{Y}_i) \nearrow$.

Idea

Look for a (small) set $\mathcal{S}$ with k variables among p such that

$$Y \approx X_{\mathcal{S}}^T \hat{\boldsymbol{\beta}}_{\mathcal{S}}.$$

Ingredients

To find this tradeoff, we need

1. a criterion to evaluate the performance ;
2. an algorithm to determine the subset of k variables optimising the criterion.

Variable Selection

Problematic

With many regressor,

- we integrate more and more information in the model ;
- we have more and more parameters to estimate and $\mathbb{V}(\hat{Y}_i) \nearrow$.

Idea

Look for a (small) set $\mathcal{S}$ with k variables among p such that

$$Y \approx X_{\mathcal{S}}^T \hat{\boldsymbol{\beta}}_{\mathcal{S}}.$$

Ingredients

To find this tradeoff, we need

1. a criterion to evaluate the performance ;
2. an algorithm to determine the subset of k variables optimising the criterion.

Outline

Estimation of the prediction error by cross-validation

For the regression: PRESS (*predicted residual sum of squares*)

Principe

1. Split the data into K subsets,
2. Successively use each subset as the test set,
3. Compute the test error for the K subsets,
4. Average the K error to get the final estiamte.

Formalism

Let $\kappa : \{1, \ldots, n\} \to \{1, \ldots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by randomization. Denote by $\hat{f}^{-\kappa(i)}$ the fitted model, computed with the kth part of the data removed. Then

$$\mathrm{CV}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \hat{\boldsymbol{\beta}}^{-\kappa(i)})^2$$

provides an estimate of the prediction error.

Penalized Criterion

Principle

Idea

Rather than estimating the prediciton error with the test error, we estimate how much the training error under estimate the true prediction error.

General form

Based on the available model fit, compute

$$\hat{\text{err}} = \text{err}_{\mathcal{D}} + \text{"optimism"}.$$

Remarks

- "penalize"to much complex models

Penalized Criterion

Principle

Idea

Rather than estimating the prediciton error with the test error, we estimate how much the training error under estimate the true prediction error.

General form

Based on the available model fit, compute

$$\hat{\text{err}} = \text{err}_{\mathcal{D}} + \text{"optimism"}.$$

Remarks

- "penalize"to much complex models

Penalized Criteria

The most Popular in linear regression

Let k be the size of the current model (i.e. the current number of predictors).

Criterion for the Linear regression model σ known

We choose the model with size k minimizing one of the following

- **Mallows** C_p

$$C_p = \frac{\text{err}_{\mathcal{D}}}{\sigma^2} - n + 2\frac{k}{n}$$

- **Akaïke Information Criteria** equivalent to C_p when σ is known

$$\text{AIC} = -2\text{loglik} + 2k = \frac{n}{\sigma^2}\text{err}_{\mathcal{D}} + 2k.$$

- **Bayesian Information Criterion**

$$\text{BIC} = -2\text{loglik} + k\log(n) = \frac{n}{\sigma^2}\text{err}_{\mathcal{D}} + k\log(n).$$

Penalized Criteria

The most Popular in linear regression

Let k be the size of the current model (i.e. the current number of predictors).

Criterion for the Linear regression model σ unknown

We choose the model with size k minimizing one of the following

- **Mallows** C_p σ estimated by the unbiased estimator $\hat{\sigma}$

$$C_p = \frac{\text{err}_{\mathcal{D}}}{\hat{\sigma}^2} - n + 2\frac{k}{n}$$

- **Akaïke Information Criteria** σ^2 estimated by $\text{err}_{\mathcal{D}}/n$

$$\text{AIC} = -2\text{loglik} + 2k = n\log(\text{err}_{\mathcal{D}}) + 2k.$$

- **Bayesian Information Criterion** σ^2 estimated by $\text{err}_{\mathcal{D}}/n$

$$\text{BIC} = -2\text{loglik} + k\log(n) = n\log(\text{err}_{\mathcal{D}}) + k\log(n).$$

C_p/AIC: proof

Ideally, we would like to minimize the error of the mean distance between the true model $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ and the OLS. This diustance splits as follows

$$\begin{aligned}\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}\|^2 =& \|\mathbf{y} - \varepsilon - \mathbf{P}_{\mathbf{X}}\mathbf{y}\|^2 \\ =& \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\varepsilon\|^2 - 2\varepsilon^\intercal(\mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y}) \\ =& n\text{err}_{\mathcal{D}} + \|\varepsilon\|^2 - 2\varepsilon^\intercal(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\boldsymbol{\mu} + \varepsilon) \\ =& n\text{err}_{\mathcal{D}} - \|\varepsilon\|^2 + 2\varepsilon^\intercal\mathbf{P}_{\mathbf{X}}\varepsilon - 2\varepsilon^\intercal(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\mu}\end{aligned}$$

On average we get

- $\mathbb{E}[\|\varepsilon\|^2] = n\sigma^2$
- $\mathbb{E}[\varepsilon^\intercal(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\mu}] = 0$
- $\mathbb{E}[2\varepsilon^\intercal\mathbf{P}_{\mathbf{X}}\varepsilon] = 2\mathbb{E}[\text{trace}(\varepsilon^\intercal\mathbf{P}_{\mathbf{X}}\varepsilon)] = 2\text{trace}(\mathbf{P}_{\mathbf{X}})\sigma^2$

If k is the dimension of the space of the projection, we find

$$\mathbb{E}\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}}\|^2 = n\text{err}_{\mathcal{D}} - n\sigma^2 + 2k\sigma^2$$

We then just have to divide by $n\sigma^2$.

Outline

Exhaustive search (best-subset)

Algorithm

For $k = 0, \ldots, p$, find the subset with k variables with the smallest SCR among 2^k models.

Properties

- Generalize to any criterion (R^2, AIC, BIC...)
- Efficient algorithm with pruning ("Leaps and Bound")
- impossible as soon as $p > 30$.

(Forward regression)

Algorithm

1. Begin with $\mathcal{S} = \emptyset$
2. at step k find the variable which, added to $\mathcal{S}$, gives the best model

2'. At step k find the best model by either adding or removing one variable.

3 etc. until p variables enter the model

Properties

- Best model is understood as SCR or R^2, AIC, BIC...
- useful when p is large
- large bias, but variance/complexity controlled.
- "greedy" algorithm

Forward-stepwise

Algorithm

1. Begin with $\mathcal{S} = \emptyset$
2. at step k find the variable which, added to $\mathcal{S}$, gives the best model

2'. At step k find the best model by either adding or removing one variable.

3 etc. until p variables enter the model

Properties

- Best model is understood as SCR or R^2, AIC, BIC...
- useful when p is large
- large bias, but variance/complexity controlled.
- "greedy" algorithm

Backward regression

Algorithm

1. Start with the full model $\mathcal{S} = \{1, \dots, p\}$
2. At step k, remove the less influent variable.
3. etc. until $\mathcal{S}$ is empty.

Properties

- Best model is understood as SCR or R^2, AIC, BIC...
- does not work when $n < p$
- large bias, but variance/complexity controlled.
- "greedy" algorithm

Outline

Exhaustive serach I

```
library(leaps)
```

Get all possible models

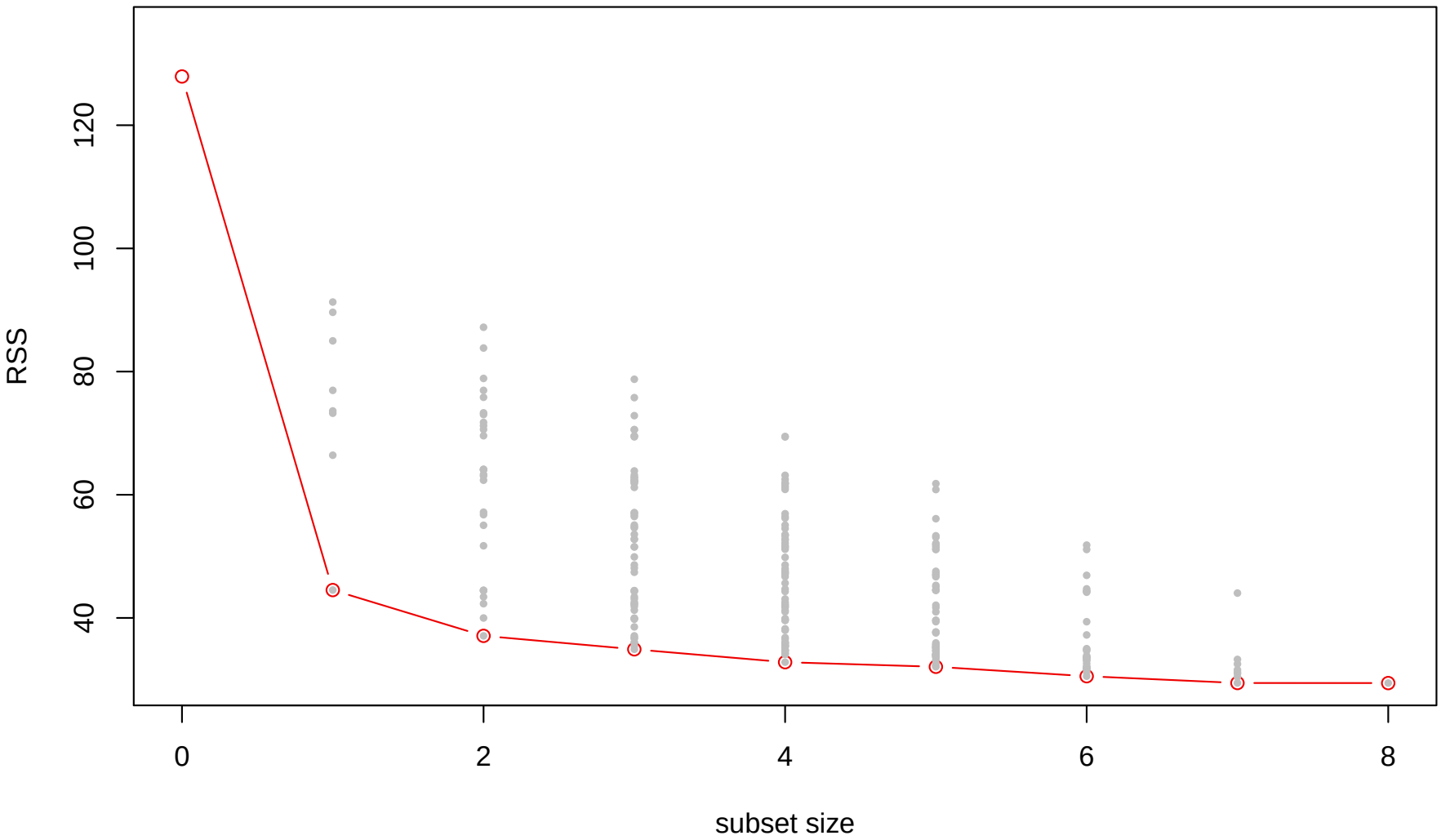
```
out <- regsubsets(lpsa ~ . , data=prostate.train,
                  nbest=100, really.big=TRUE)
bss <- summary(out)
```

Extract size and RSS. Add the null model (just l'intercept)

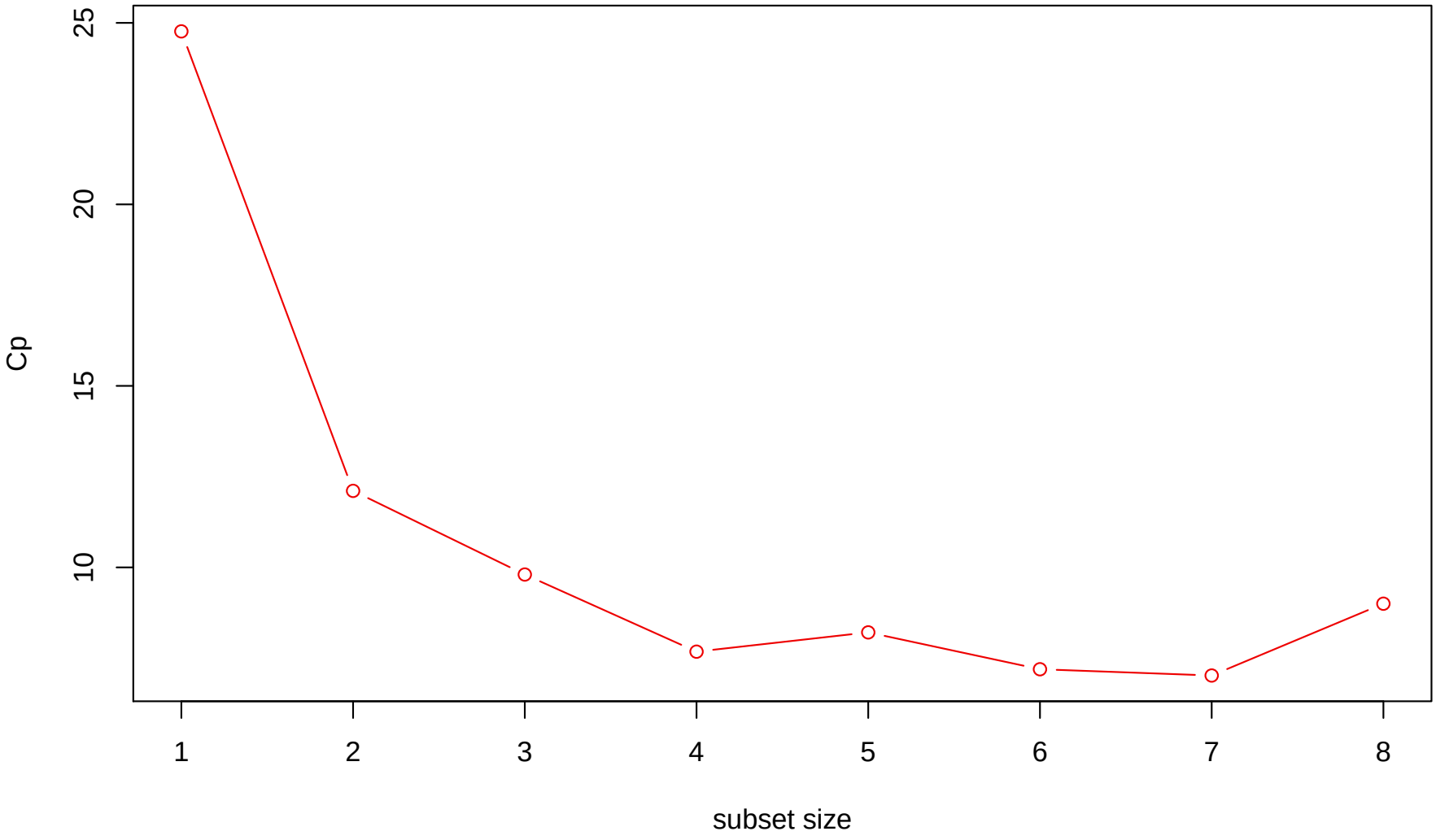
```
bss.size  <- as.numeric(rownames(bss$which))
intercept <- lm(lpsa ~ 1, data=prostate)
bss.best.rss  <- c(sum(resid(intercept)^2), tapply(bss$rss  , bss.size, min))
```

```
plot(0:8, bss.best.rss, ylim=c(30, 135), type="b",
     xlab="subset size", ylab="RSS", col="red2" )
points(bss.size, bss$rss, pch=20, col="gray", cex=0.7)
```

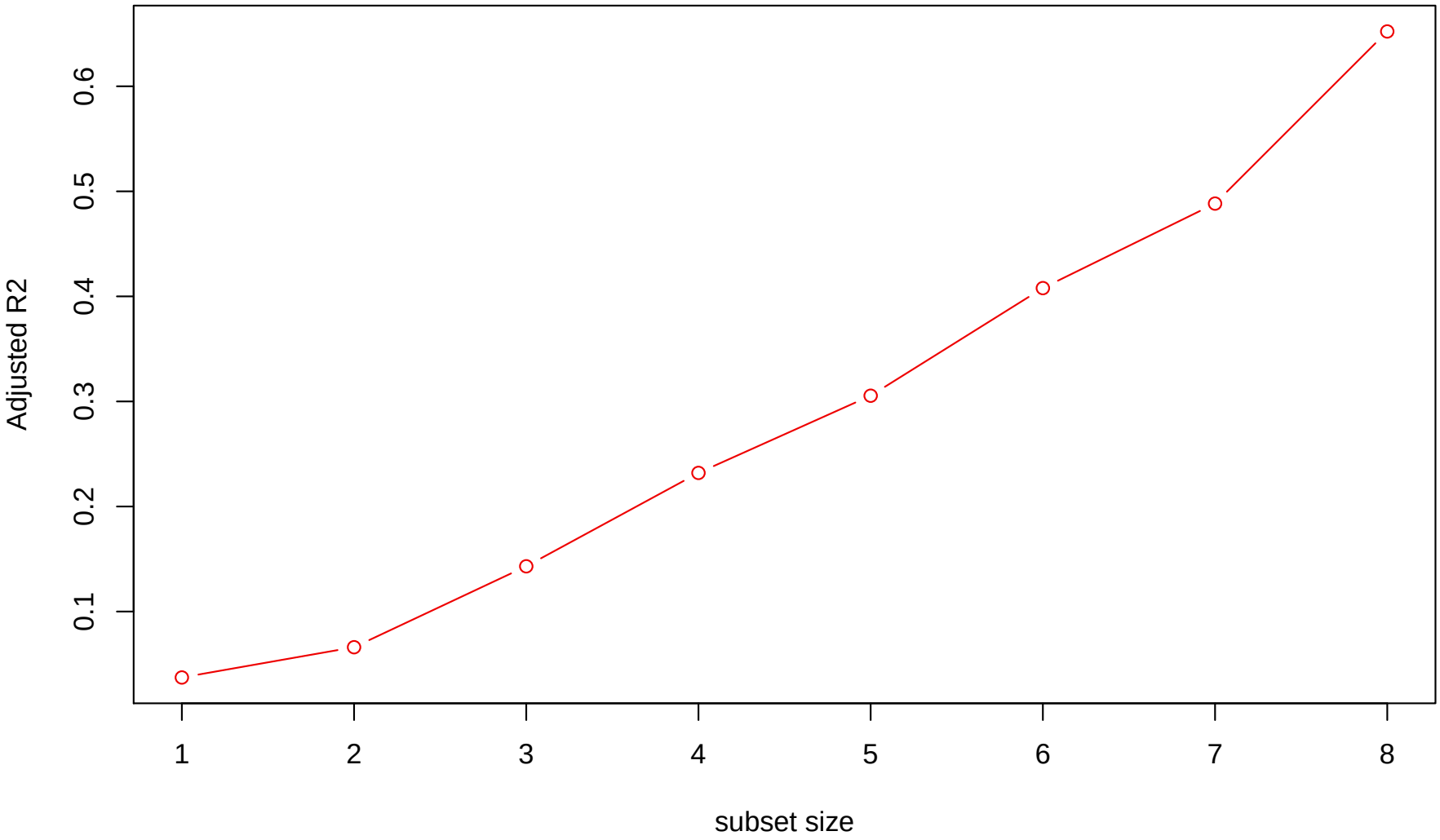
Exhaustive serach II



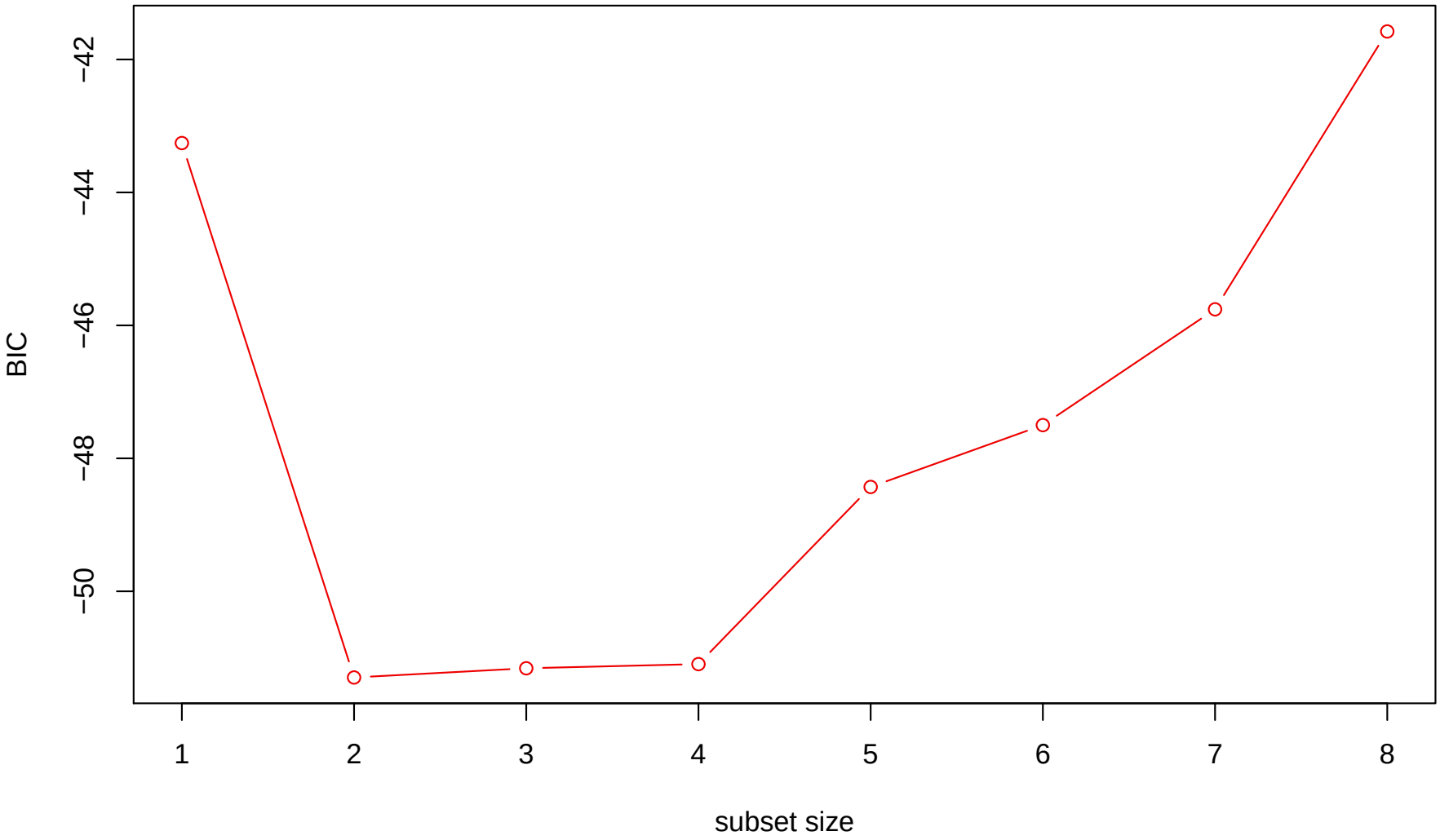
Exhaustive search III



Exhaustive search VI



Exhaustive search V



Forward-Stepwise (I)

Create the nul model and the full model

```
null  <- lm(lpsa ~ 1, data=prostate.train)
full  <- lm(lpsa ~ ., data=prostate.train)
```

Create the scope of models

```
lower <- ~1
upper <- ~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45
scope <- list(lower=lower,upper=upper)
```

Stepwise with AIC: forward, backward, both

```
fwd  <- step(null, scope, direction="forward" , trace=FALSE)
bwd  <- step(full, scope, direction="backward", trace=FALSE)
both <- step(null, scope, direction="both"    , trace=FALSE)
```

$\rightsquigarrow$ 3 equivalent models

Forward regression

```
fwd
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph, data = prostate.train)
##
## Coefficients:
## (Intercept)       lcavol      lweight          svi          lbph
##     -0.3259       0.9177       1.9853       0.3203       0.2052
```

```
fwd$anova
```

```
##        Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1           NA        NA        66   96.28145  26.29306
## 2  + lcavol -1 51.752862        65   44.52858 -23.37361
## 3 + lweight -1  7.436737        64   37.09185 -33.61680
## 4     + svi -1  2.184097        63   34.90775 -35.68291
## 5    + lbph -1  2.092754        62   32.81499 -37.82507
```

Backward regression

```
bwd
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     pgg45, data = prostate.train)
##
## Coefficients:
## (Intercept)       lcavol      lweight          age         lbph
##      0.2591       1.0419       2.2814      -1.2791       0.2116
##         svi          lcp        pgg45
##      0.3536      -0.2911       0.3532
```

```
bwd$anova
```

```
##        Step Df   Deviance Resid. Df Resid. Dev       AIC
## 1           NA         NA        58   29.42638 -37.12766
## 2 - gleason  1 0.01091586        59   29.43730 -39.10281
```

Stepwise regression

```
both
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph, data = prostate.train)
##
## Coefficients:
## (Intercept)       lcavol      lweight          svi          lbph
##     -0.3259       0.9177       1.9853       0.3203       0.2052
```

```
both$anova
```

```
##        Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1           NA        NA        66   96.28145  26.29306
## 2  + lcavol -1 51.752862        65   44.52858 -23.37361
## 3 + lweight -1  7.436737        64   37.09185 -33.61680
## 4     + svi -1  2.184097        63   34.90775 -35.68291
## 5    + lbph -1  2.092754        62   32.81499 -37.82507
```

Performance on test data

```
print(err.ols)

## [1] 0.5221043

print(err.AIC.fwd  <- mean((y.test-predict(fwd ,prostate.test))^2))

## [1] 0.4520967

print(err.AIC.bwd  <- mean((y.test-predict(bwd ,prostate.test))^2))

## [1] 0.517824

print(err.AIC <- mean((y.test-predict(both,prostate.test))^2))

## [1] 0.4520967
```

Stepwise: BIC modification

More sparse model

```
BIC <- step(null, scope, k=log(n <- nrow(prostate)), trace=FALSE)
BIC
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight, data = prostate.train)
##
## Coefficients:
## (Intercept)       lcavol      lweight
##      -1.049        1.139        2.720
```

```
print(err.BIC  <- mean((y.test-predict(BIC ,prostate.test))^2))
```

```
## [1] 0.4908699
```

Comments

Interpretability

1. If the true $\mathcal{S}$ only contains a **few variables linked to the response**,
   $\rightsquigarrow$ variable selection algorithms can retrieve relevent predictors.
2. If the true $\mathcal{S}$ contains **many correlated predictors**
   $\rightsquigarrow$ the selected variables will be hardly interpretable.

Stability issue

With strong correlation or when $n < p$, **small changes** in the data can induce **large discrepencies** between the sets of selected variables.

Outline

Outline

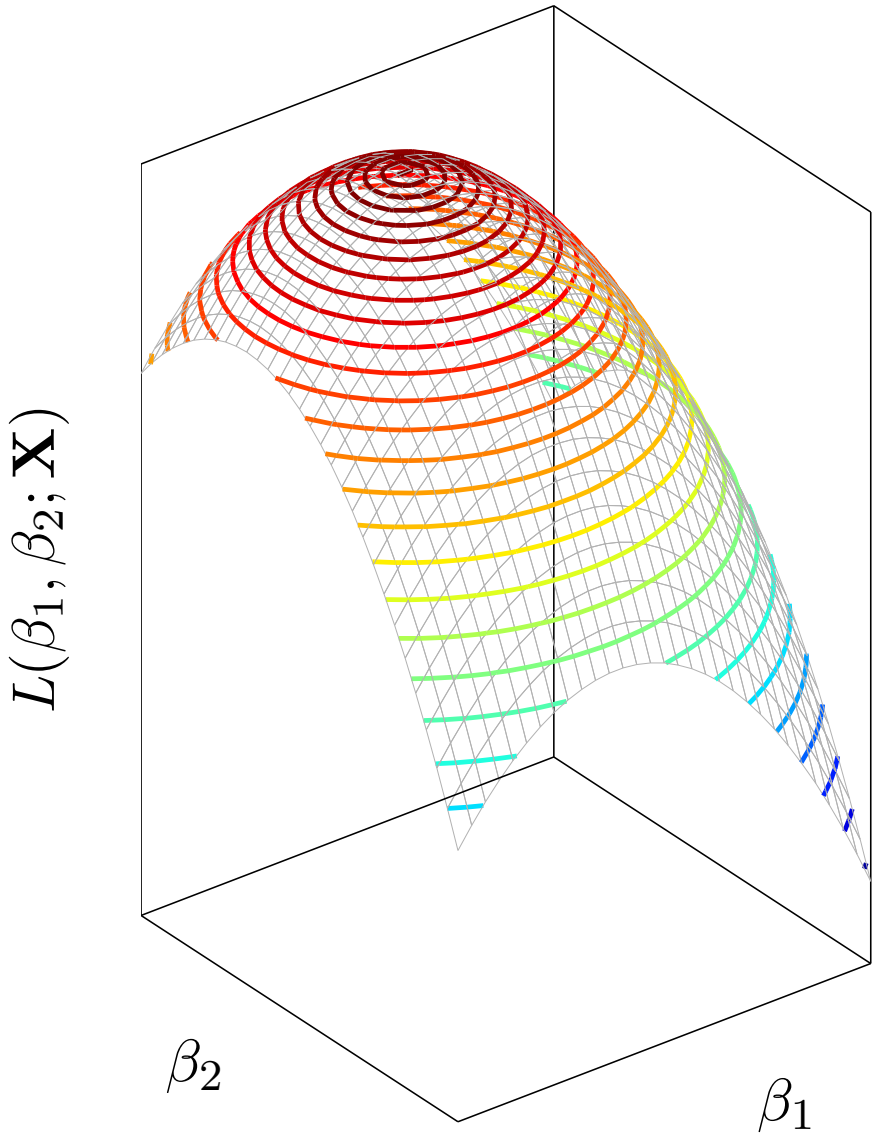
Several goals

Control the parameter $\hat{\beta}$ to

1. Regularize the problem
   - For numerical purpose, (conditioning of $\mathbf{X}^T\mathbf{X}$),
   - For stability purpose, (correlation between $(X_1, \ldots, X_p)$).
2. Enhance the prediction
   - By trading a little bias vs variance
   - By controlling irrelevant variables
3. Looking towards interpretability
   - By controlling model complexity,
   - By embedding the variable selection (Lasso).

A Geometric View of Shrinkage

Constrained Optimization



We basically want to solve a problem of the form

$$\underset{\beta_1,\beta_2}{\text{maximize}}\, L(\beta_1, \beta_2; \mathbf{X})$$

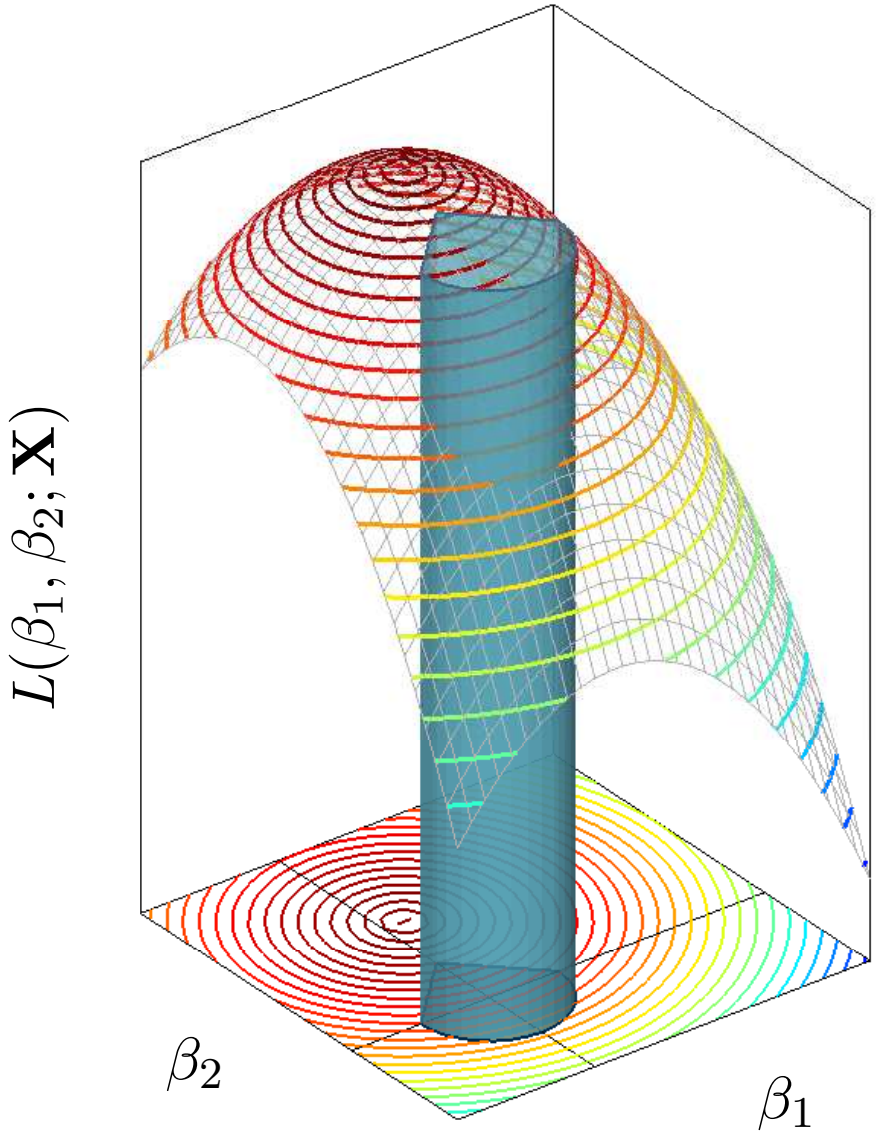
where L is typically a concave likelihood function.
This is strictly equivalent to solve

$$\underset{\beta_1,\beta_2}{\text{minimize}}\, L'(\beta_1, \beta_2; \mathbf{X})$$

where $L' = -L$ is convex ! For instance the squared error loss in the OLS.

A Geometric View of Shrinkage

Constrained Optimization

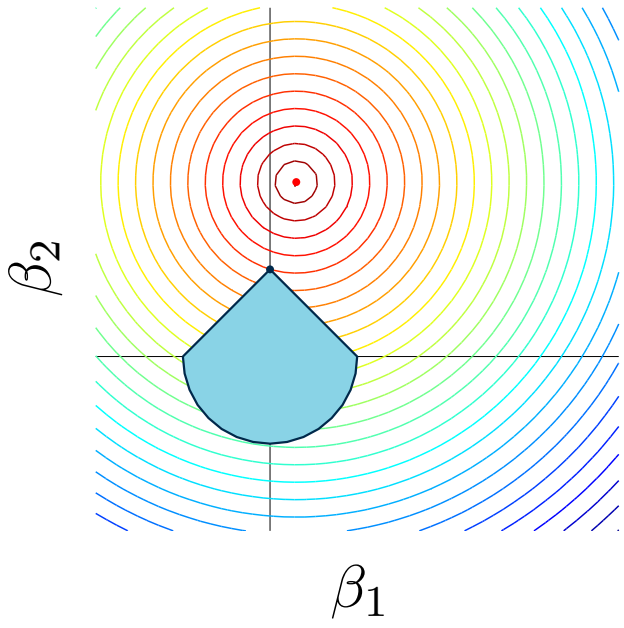


$$\begin{cases} \underset{\beta_1,\beta_2}{\text{maximize}} & L(\beta_1, \beta_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases},$$

where Ω defines a domain that constrains $\boldsymbol{\beta}$.

A Geometric View of Shrinkage

Constrained Optimization



$$\begin{cases} \underset{\beta_1,\beta_2}{\text{maximize}} & L(\beta_1, \beta_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases},$$

where Ω defines a domain that constrains $\boldsymbol{\beta}$.

$$\Updownarrow$$

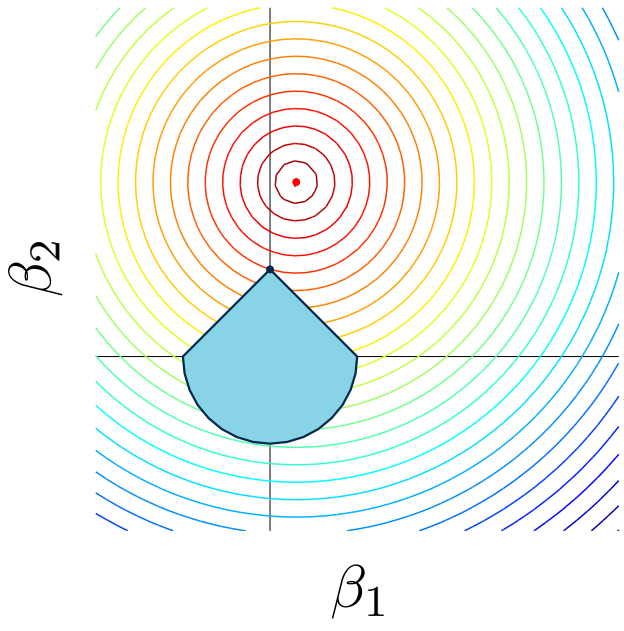
$$\underset{\beta_1,\beta_2}{\text{minimize}}\, J(\boldsymbol{\beta}),$$

with J the convex objective defined by

$$J(\boldsymbol{\beta}) = -L(\beta_1, \beta_2; \mathbf{X}) + \lambda\Omega(\beta_1, \beta_2)$$

A Geometric View of Shrinkage

Constrained Optimization



$$\begin{cases} \underset{\beta_1,\beta_2}{\text{maximize}} & L(\beta_1, \beta_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases},$$

where Ω defines a domain that constrains $\boldsymbol{\beta}$.

$$\Updownarrow$$

$$\underset{\beta_1,\beta_2}{\text{minimize}}\, J(\boldsymbol{\beta}),$$

with J the convex objective defined by

$$J(\boldsymbol{\beta}) = -L(\beta_1, \beta_2; \mathbf{X}) + \lambda\Omega(\beta_1, \beta_2)$$

How shall we define Ω ?

Outline

Outline

Definition

Fact

If the β_j are unconstrained, they can have very high magnitude and thus large variances.

Idea

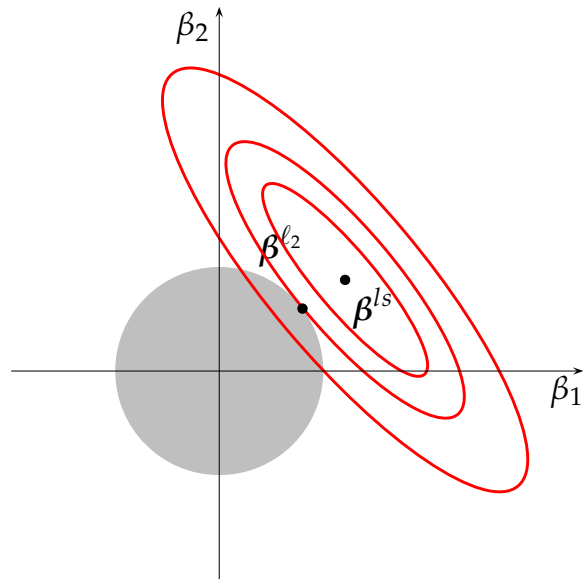
To control the variance, we should control the size of the coefficients in $\boldsymbol{\beta}$. This could induce a large decrease of the prediction error.

Ridge as a regularization problem

The ridge estimate of $\boldsymbol{\beta}$ is the solution to

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta} \in \mathrm{R}^{p+1}}{\arg\min}\, \mathrm{RSS}(\boldsymbol{\beta}), \quad \text{s.t.} \sum_{j=1}^{p} \beta_j^2 \leq s,$$

where s is a shrinkage factor.



A 2-dimensional toy example

Consider that the true relationship is $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ If X_1 and X_2 are strongly correlated, then $X_1 \approx X_2$ and for any $\gamma \geq 0$

$$\begin{aligned} Y &= X_1(\beta_1 + \gamma) + X_2(\beta_2 - \gamma) + \gamma(X_1 - X_2) + \varepsilon \\ &\approx X_1(\beta_1 + \gamma) + X_2(\beta_2 - \gamma) + \varepsilon. \end{aligned}$$

A large panel of fit with estimated $\boldsymbol{\beta}$ varying according to γ will produce the same prediction error.

For small s (or large λ in the Lagrangian form), the ridge controls

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2$$

which is minimal for $\gamma = (\beta_2 - \beta_1)/2$, and in this case $\beta_j = (\beta_1 + \beta_2)/2$.

A 2-dimensional toy example

Consider that the true relationship is $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ If X_1 and X_2 are strongly correlated, then $X_1 \approx X_2$ and for any $\gamma \geq 0$

$$\begin{aligned} Y &= X_1(\beta_1 + \gamma) + X_2(\beta_2 - \gamma) + \gamma(X_1 - X_2) + \varepsilon \\ &\approx X_1(\beta_1 + \gamma) + X_2(\beta_2 - \gamma) + \varepsilon. \end{aligned}$$

A large panel of fit with estimated $\boldsymbol{\beta}$ varying according to γ will produce the same prediction error.

For small s (or large λ in the Lagrangian form), the ridge controls

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2$$

which is minimal for $\gamma = (\beta_2 - \beta_1)/2$, and in this case $\beta_j = (\beta_1 + \beta_2)/2$.

A 2-dimensional toy example (in R) I

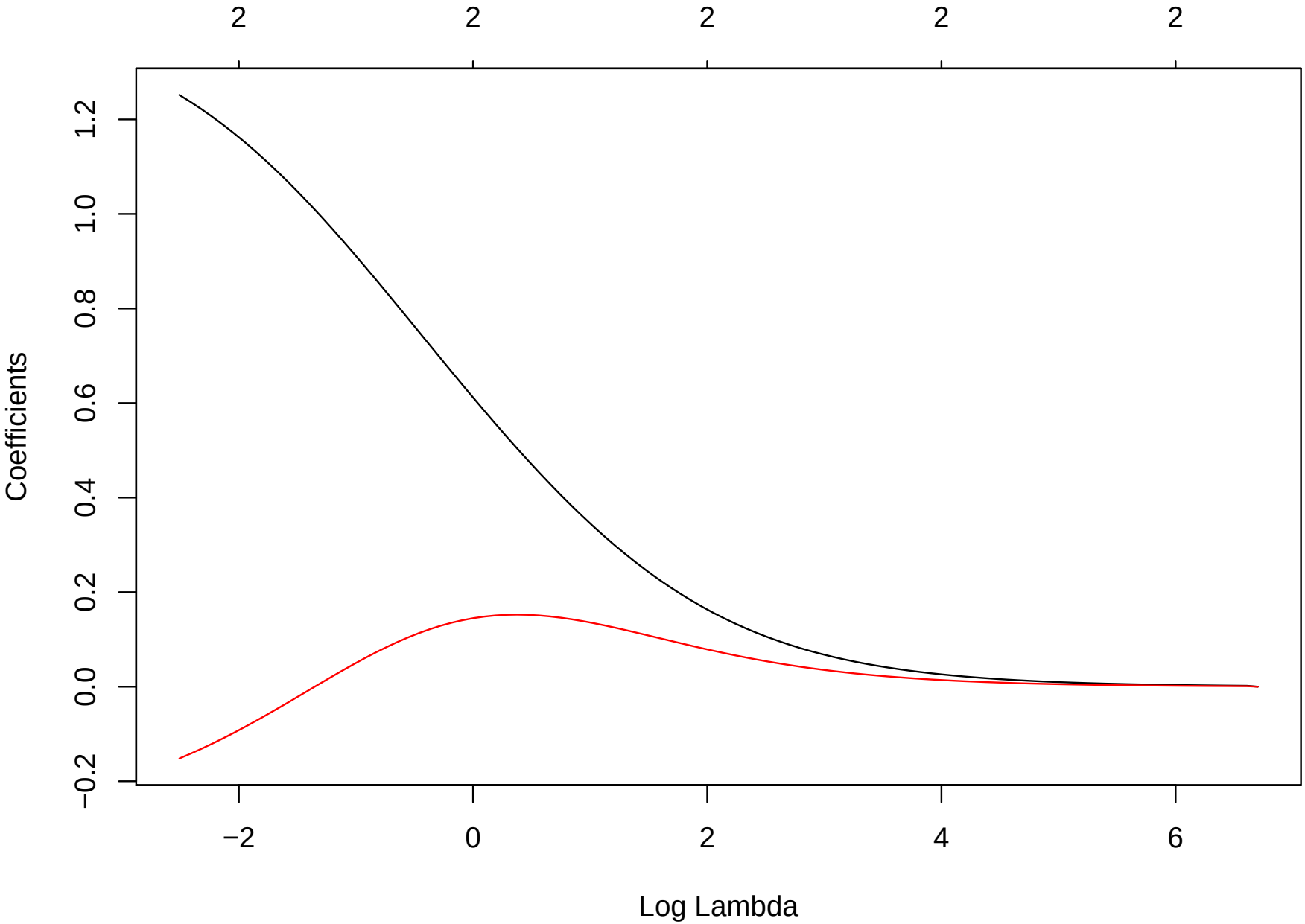
Generate two correlated predictors

```
suppressMessages(library(quadrupen))
x1 <- rnorm(5)
x2 <- x1 + rnorm(5,0, 0.5)
cor(x1,x2)

## [1] 0.6947718
```

Draw Y and plot the **ridge regularisation path**

```
library(glmnet)
y <- x1 + x2 +rnorm(5)
plot(glmnet(cbind(x1,x2),y, alpha=0), xvar="lambda")
```

A 2-dimensional toy example (in R) II



Ridge as penalized regression

Dont penalize the intercept thus consider $\boldsymbol{\beta} = (\beta_1, \dots \beta_p)$ and set

- $\hat{\beta}_0 = \bar{\mathbf{y}} - \bar{x}\hat{\boldsymbol{\beta}}$
- center $\mathbf{y}$ and $\mathbf{x}_j$, $j = 1, \dots, p$.

Standardize the $\mathbf{x}_j$ for the fit and send back $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ to the orginal scale.

Convex Langrangian form

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{ridge}} &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \\ &= (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{H}_\lambda\mathbf{y}.\end{aligned}$$

Strong convexity

Oppositely to the least squares, a non-singular solution always exists when $\lambda > 0$ whatever the conditioning of $\mathbf{X}^\intercal\mathbf{X}$ (original proposal).

Ridge fit for the prostate cancer data

Compute the ridge path

```
ridge.path <- glmnet(x.train,y.train, alpha=0)
```

Compute the prediction error on the test set for all λ

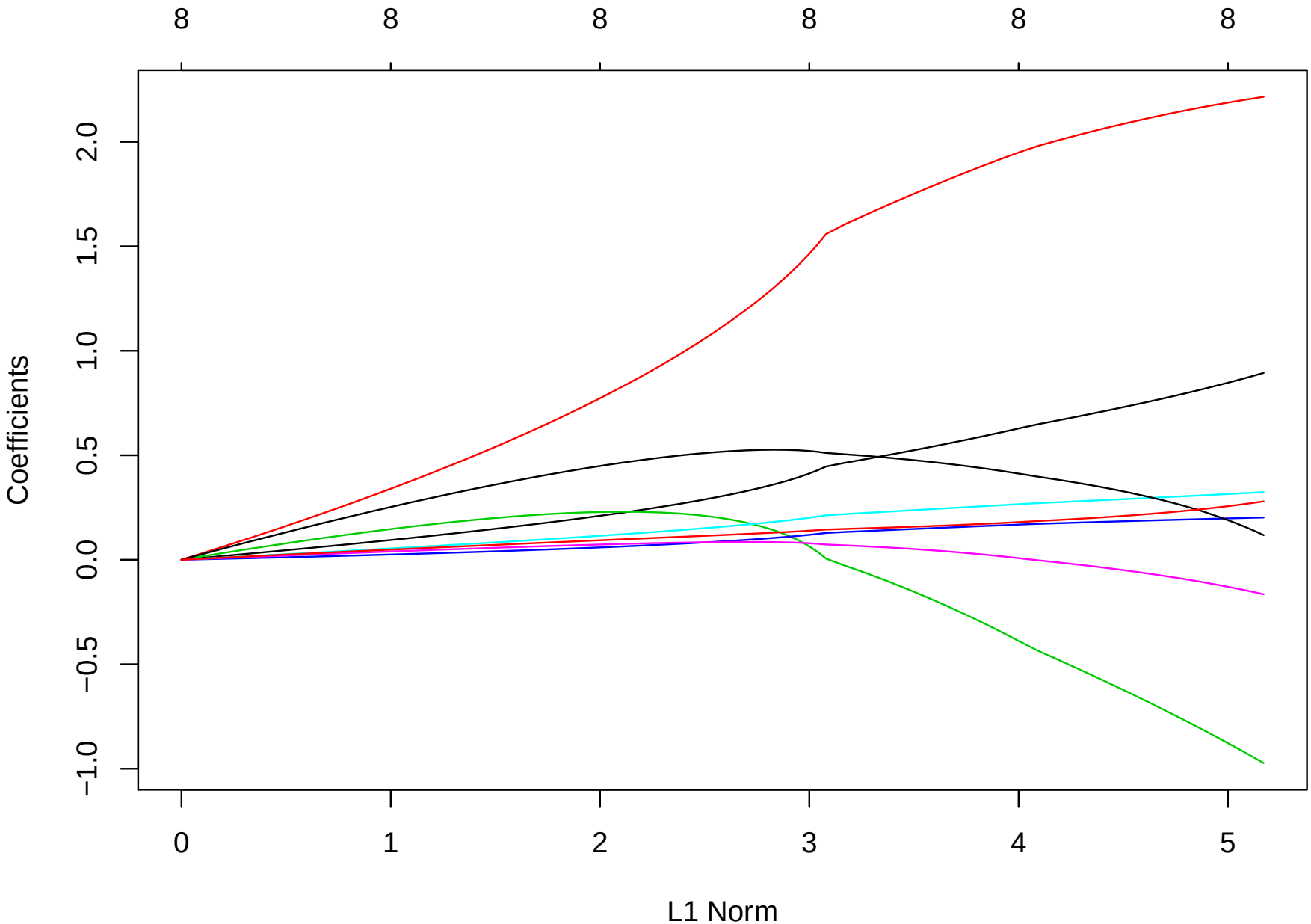
```
err <- colMeans((y.test-predict(ridge.path, x.test, type="response"))^2)
```

Then, $\lambda^\star$ that minimizes this error

```
ridge.path$lambda[which.min(err)]

## [1] 0.2228282
```

The prediction error is smaller than with the OLS

```
err.ridge <- err[which.min(err)]; err.ridge; err.ols

##       s89
## 0.4866302
## [1] 0.5221043
```


8
8
8
8
8
8
Coefficients
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5
L1 Norm

Outline

Classical options

CRoss-validation

We compute $CV(\lambda)$, the CV error along the λ path

1. if $K = n$, this is the LOOCV,
2. if $K = 2$, this is the hold out estimation,
3. in a high dimensional setup, we must choose K "carefully",

We choose λ minimising the CV

Penalized criteria

We choose λ minimizing a criterion with the form

$$\text{crit}(\lambda) = \text{err}_{\mathcal{D}}(\lambda) + \text{pen}(\text{df}_\lambda)$$

$\rightsquigarrow$ What sens give to the degrees of freedom for ridge regression?

Effective degrees of freedom

- Degrees of freedom of a model describes its complexity level.
- For the least squares, $\mathrm{df} = p$ (plus 1 for the intercept).
- Need a definition adapted to shrinkage methods.

Definition (Efron and others)

Consider a fitted vector $\hat{\mathbf{y}}$ from an observation $\mathbf{y}$. We define its degrees of freedom as

$$\mathrm{df}(\hat{\mathbf{y}}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(\hat{y}_i, y_i).$$

$\rightsquigarrow$The harder the fit to the data, the higher the covariance.

Effective degrees of freedom: the ridge case

Proposition

Consider a linear fitting method that predicts $\hat{\mathbf{y}}$ for entry $\mathbf{y}$ through the smoother matrix $\mathbf{H}$:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

The effective degrees of freedom of the model $\hat{\mathbf{y}}$ verifies

$$\mathrm{df}(\hat{\mathbf{y}}) = \mathrm{Tr}(\mathbf{H}).$$

Ridge: effective degrees of freedom

For ridge regression, df is a decreassig function of λ which tends to 0 (or 1 when considering the intercept):

$$\mathrm{df}(\hat{\mathbf{y}}_\lambda) = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda}.$$

Cross-Validation

Cross-validation is easily parallelized and is fast on small data sets

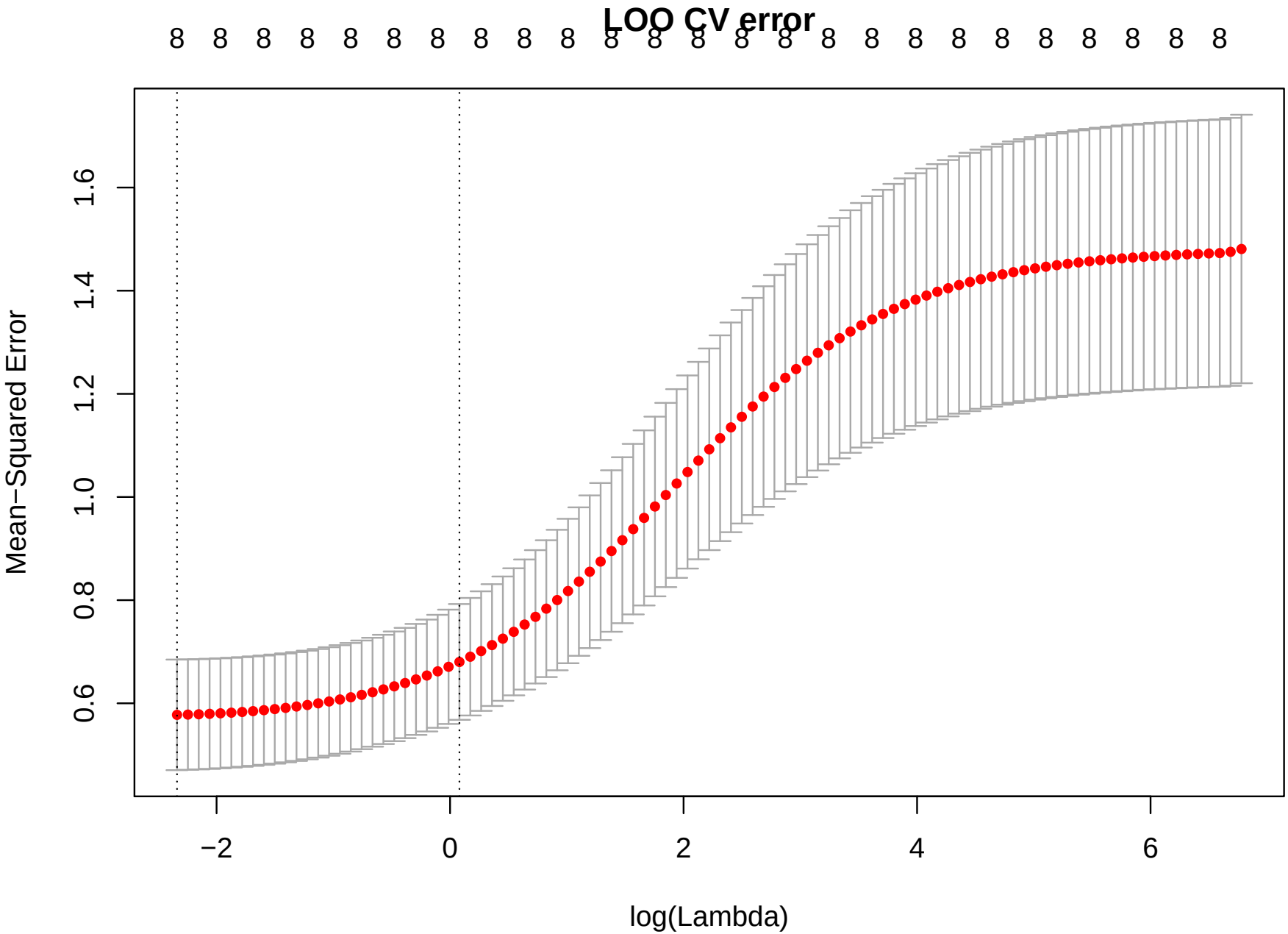
```
system.time(loo <- cv.glmnet(x.train,y.train,alpha=0,nfolds=n))
```

```
##    user  system elapsed
##   0.308   0.000   0.308
```

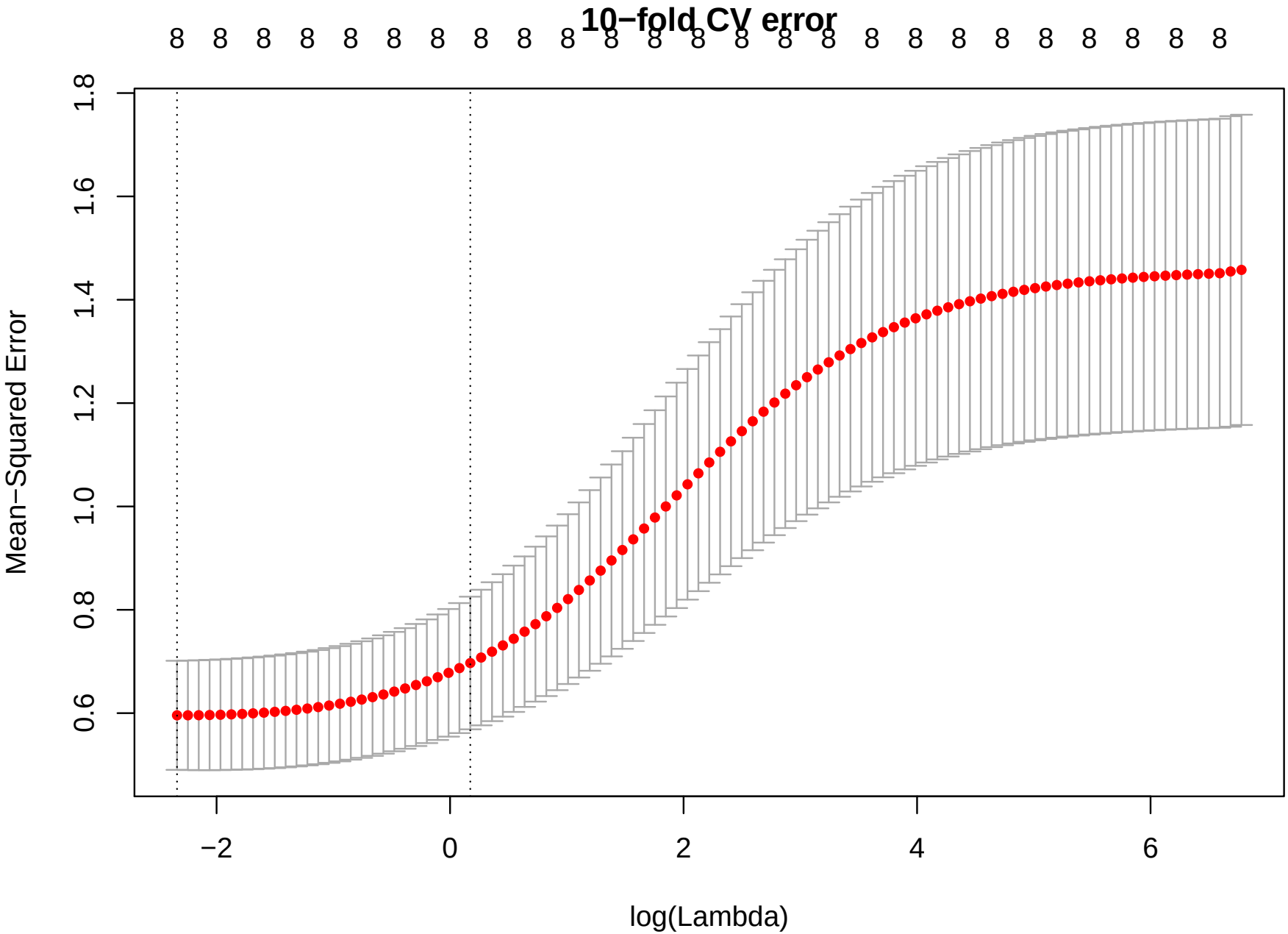
```
system.time(CV10 <- cv.glmnet(x.train,y.train,alpha=0,nfolds=10))
```

```
##    user  system elapsed
##   0.056   0.000   0.056
```

Leave one out



Ten fold


10−fold CV error
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
Mean−Squared Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
−2
0
2
4
6
log(Lambda)

Outline

Outline

The Lasso

Least Absolute Shrinkage and Selection Operator

Fact

Ridge performs regularization. . . but we also would like to select the most significant variables.

Idea

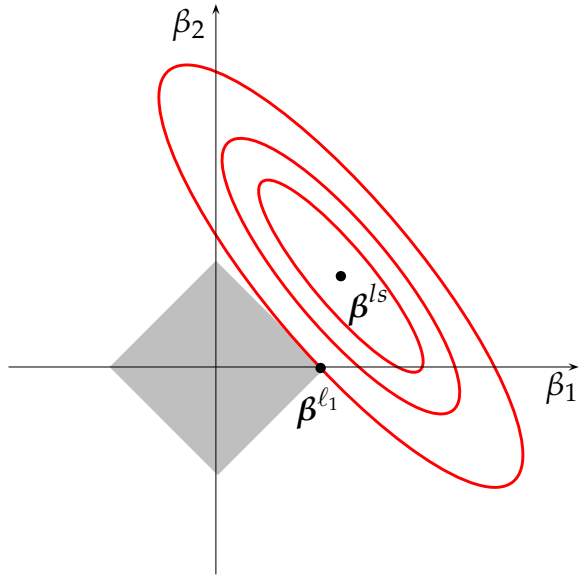
Suggest an admissible set that induces sparsity (force several entries to exactly zero in $\hat{\boldsymbol{\beta}}$).

Lasso as a convex optimization problem

The Lasso estimate $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ solves

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\text{minimize}}\, \text{RSS}(\boldsymbol{\beta}), \quad \text{s.t.} \sum_{j=1}^{p} |\beta_j| \leq s,$$

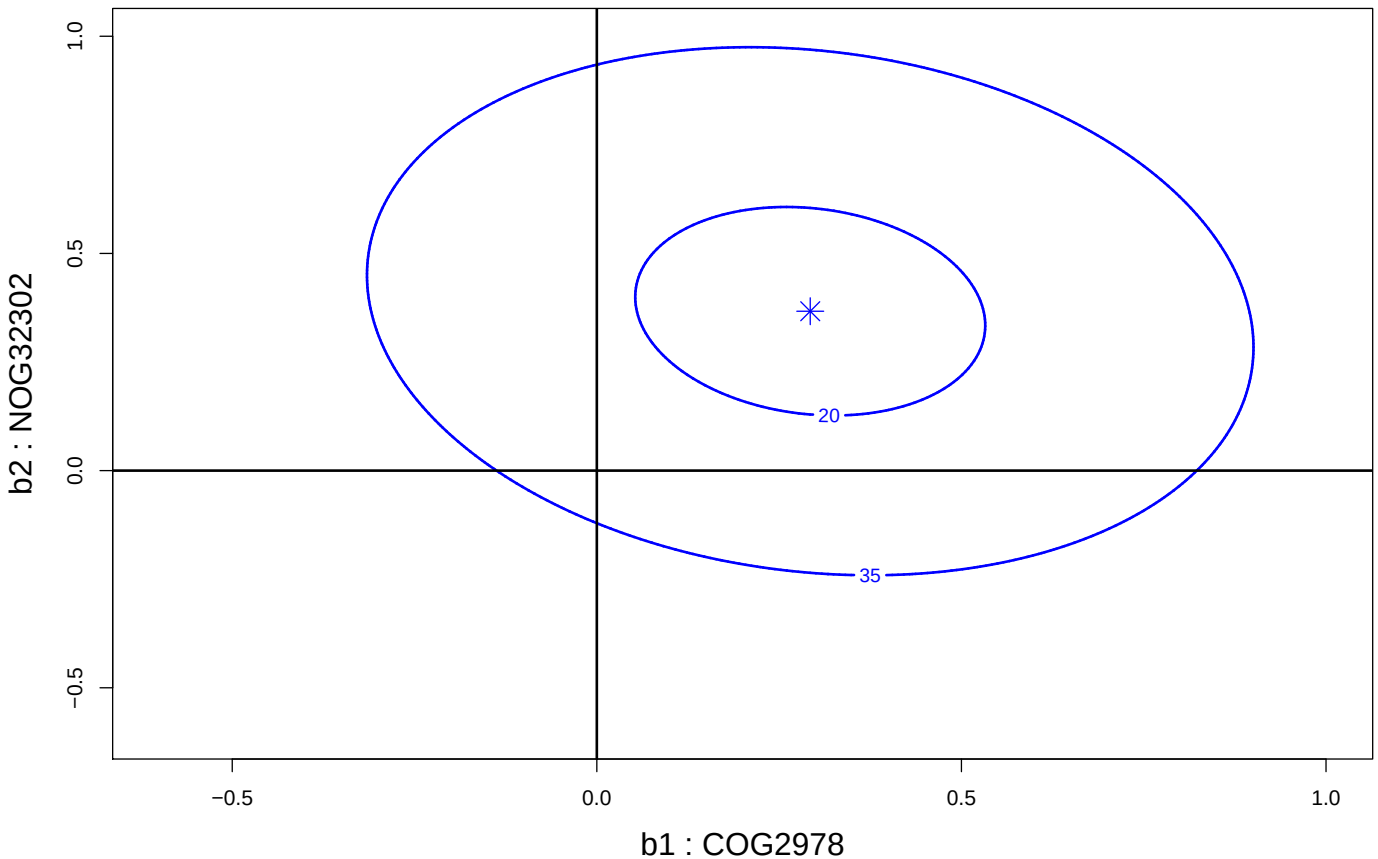
where s is a shrinkage factor.



Some more insights: 2-dimensional example

Thanks to Sylvie Huet

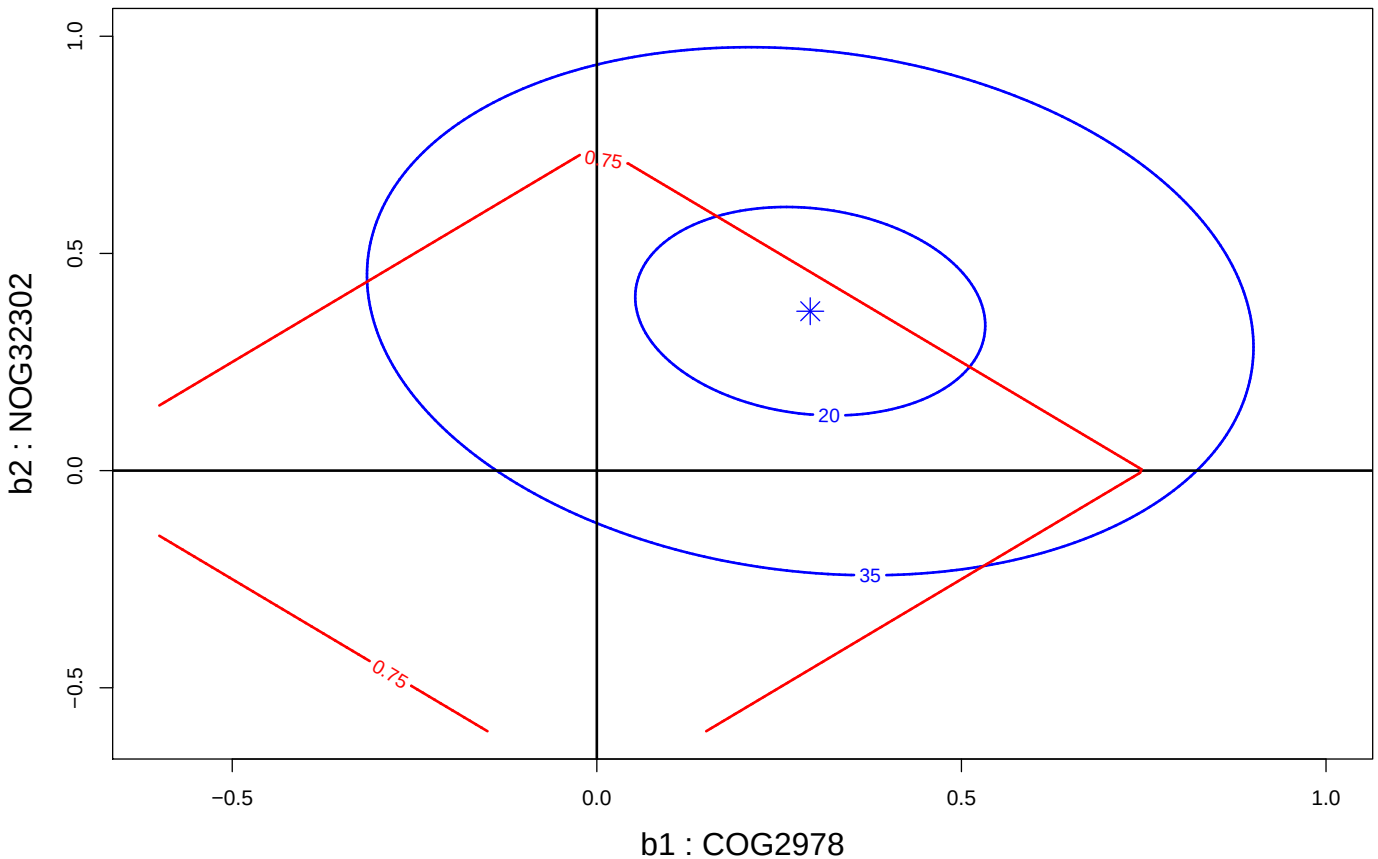
$$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2, \qquad \text{no constraints}$$



Some more insights: 2-dimensional example

Thanks to Sylvie Huet

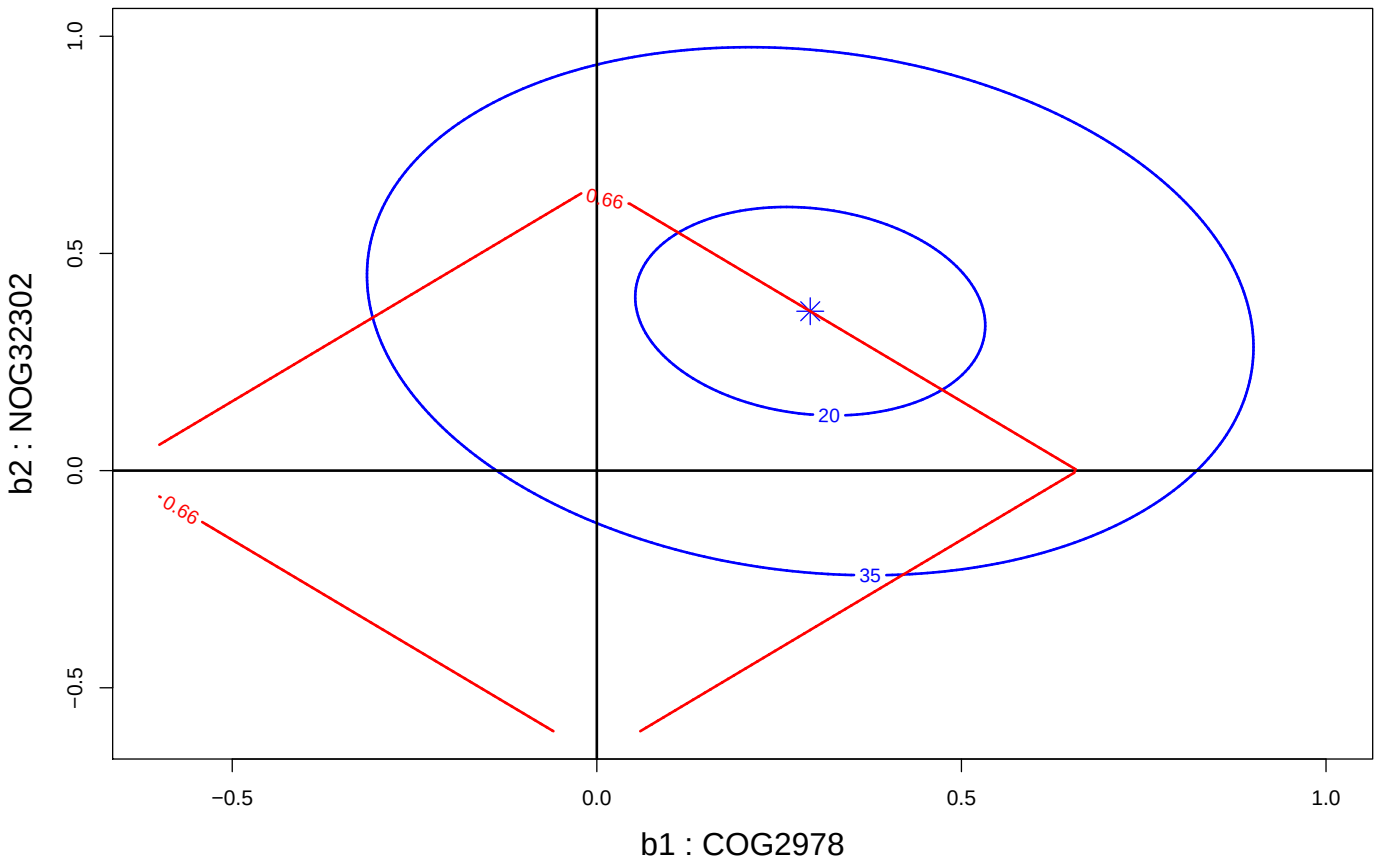
$$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2, \qquad \text{s.c. } |\beta_1| + |\beta_2| < 0.75$$



Some more insights: 2-dimensional example

Thanks to Sylvie Huet

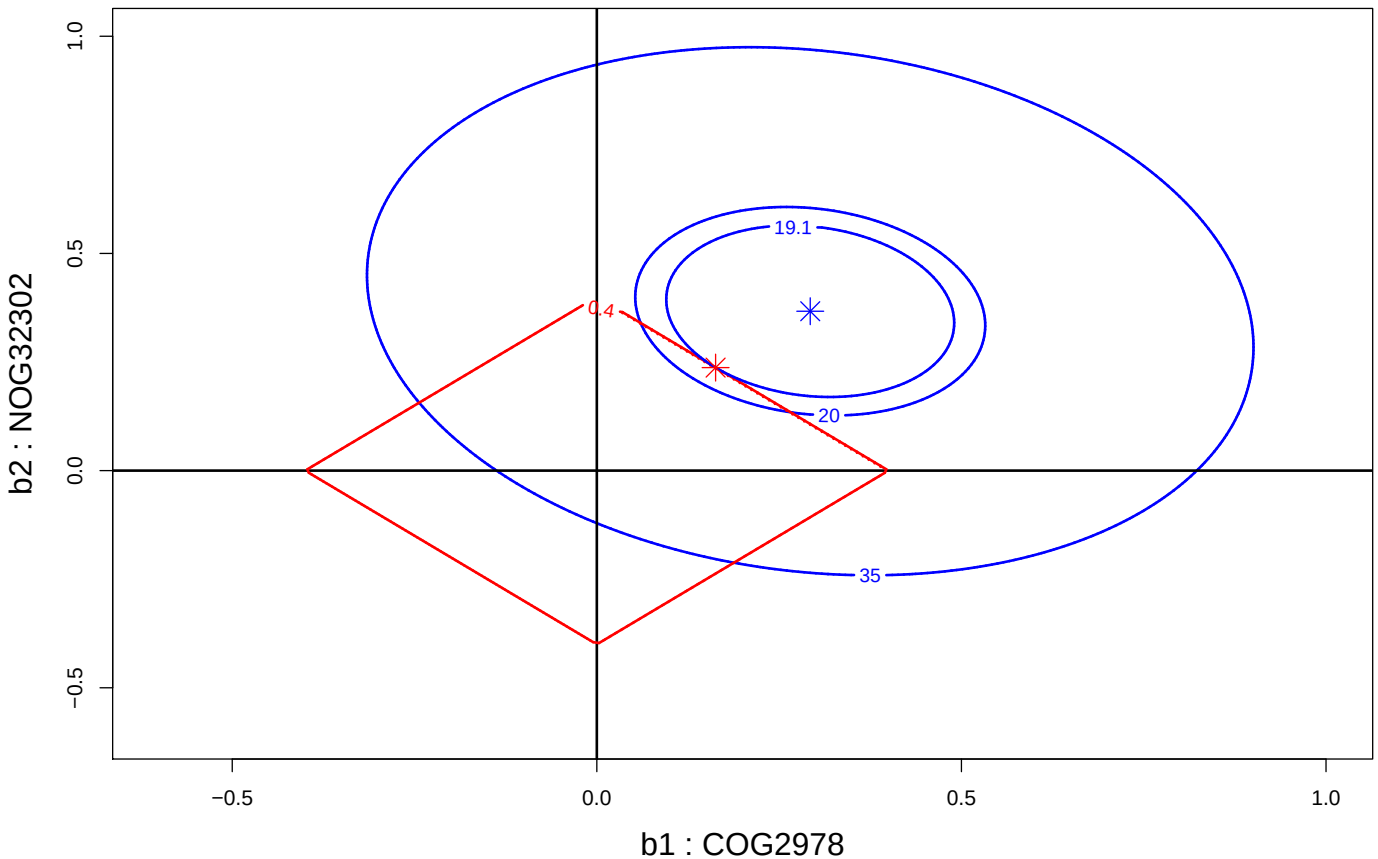
$$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2, \qquad \text{s.c. } |\beta_1| + |\beta_2| < 0.66$$



Some more insights: 2-dimensional example

Thanks to Sylvie Huet

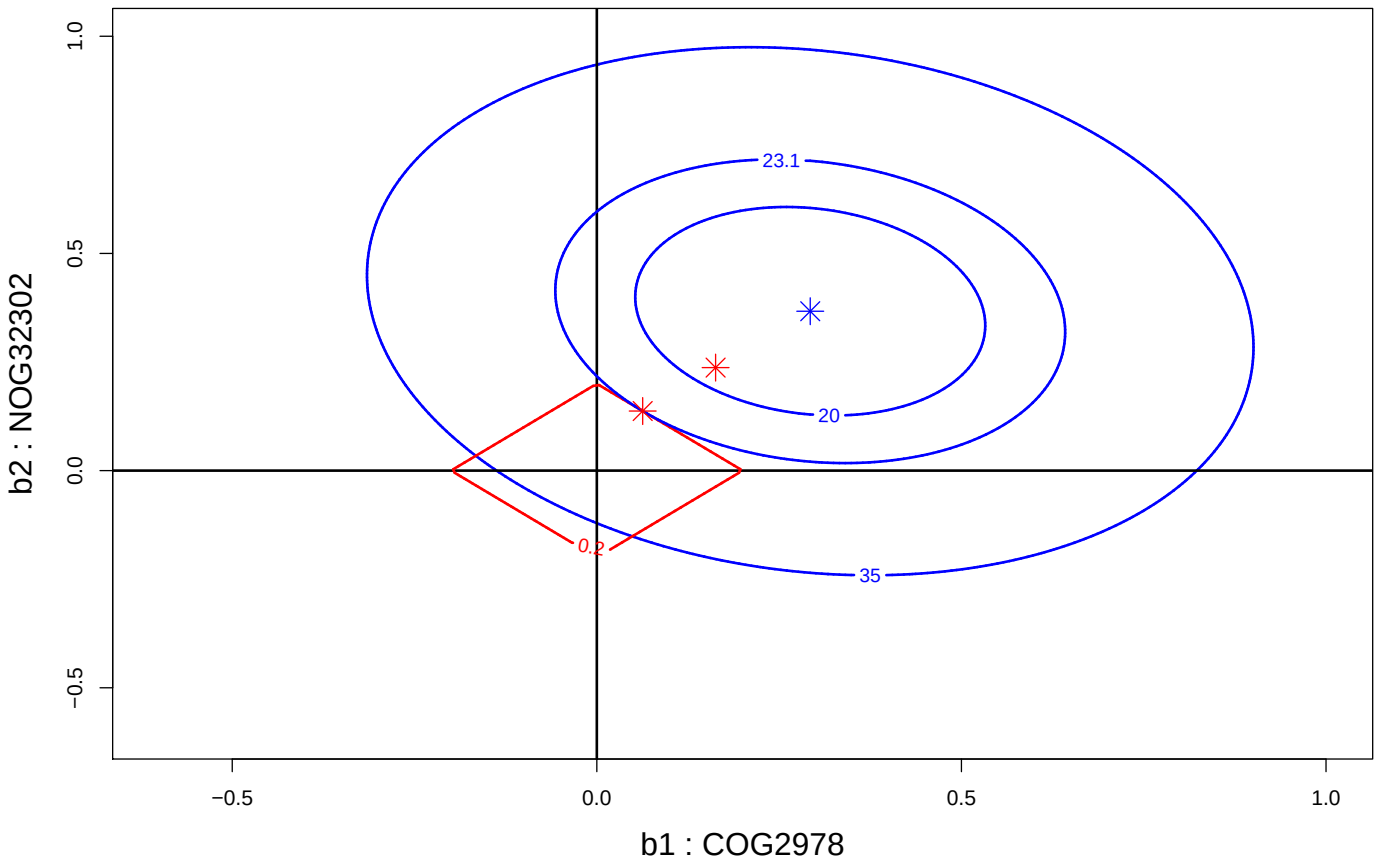
$$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2, \qquad \text{s.c. } |\beta_1| + |\beta_2| < 0.4$$



Some more insights: 2-dimensional example

Thanks to Sylvie Huet

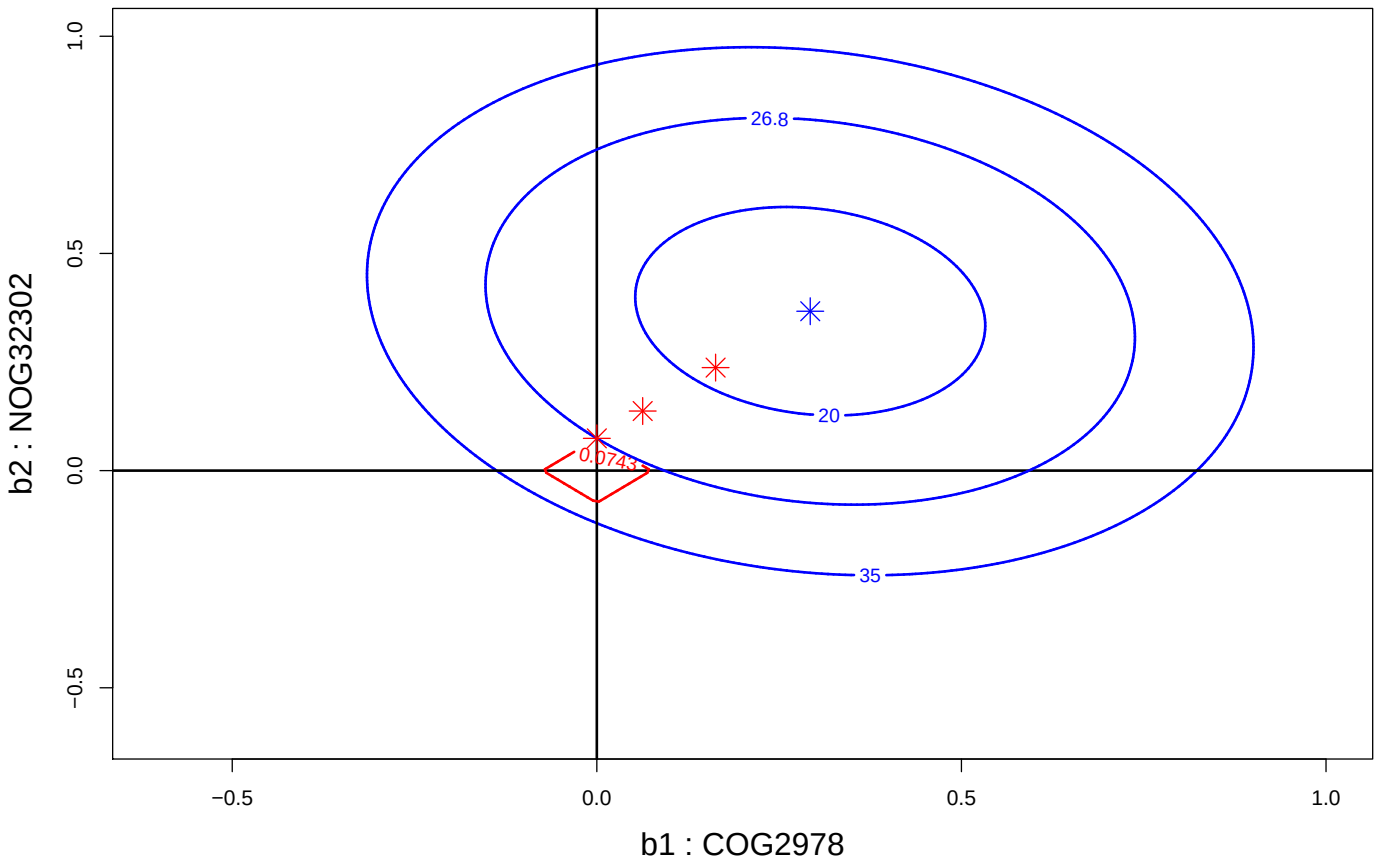
$$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2, \qquad \text{s.c. } |\beta_1| + |\beta_2| < 0.2$$



Some more insights: 2-dimensional example

Thanks to Sylvie Huet

$$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2, \qquad \text{s.c. } |\beta_1| + |\beta_2| < 0.0743$$



Lasso as penalized regression

Get rid of the intercept

We should not penalize the intercept term, thus

- $\hat{\beta}_0 = \bar{\mathbf{y}}$,
- center $\mathbf{y}$ and $\mathbf{x}_j$, $j = 1, \ldots, p$,
- scale the predictor before the fit,
- send $\hat{\boldsymbol{\beta}}$ back to the original scale.

Solve the convex, ℓ_1-penalized problem

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta} \in \mathrm{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

whose solution has no close form, but always exists and is unique as soon as $\mathbf{X}^\intercal \mathbf{X}$ has full rank.
$\rightsquigarrow$ Lasso performs regularization and variable selection but has no analytical solution.

Lasso fit on the prostate cancer data I

Compute the LASSO path

```
library(glmnet)
lasso.path <- glmnet(x.train,y.train)
```

Compute the prediction error on the test set for all λ

```
err <- colMeans((y.test-predict(lasso.path,x.test,type="response"))^2)
```

Then, $\lambda^\star$ that minimizes this error

```
lasso.path$lambda[which.min(err)]

## [1] 0.1135118
```

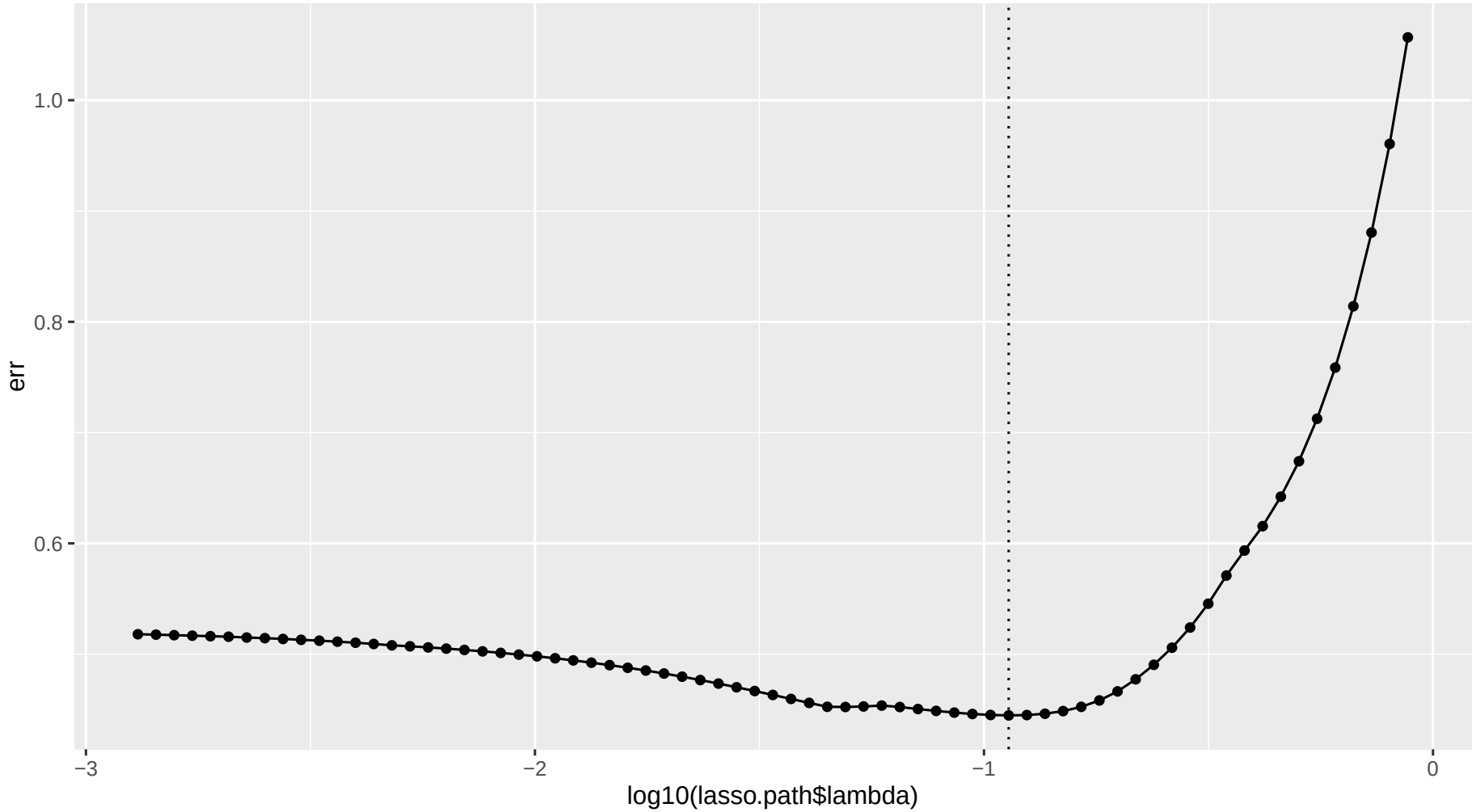
Lasso fit on the prostate cancer data II

The prediction error is smaller than with the OLS with only 5 coefficients

```
err[which.min(err)]

##       s22
## 0.4447306

lasso.path$beta[,which.min(err)]

##     lcavol    lweight        age       lbph        svi        lcp
## 0.83762993 1.74080154 0.00000000 0.09308703 0.18473598 0.00000000
##    gleason      pgg45
## 0.00000000 0.07755339
```

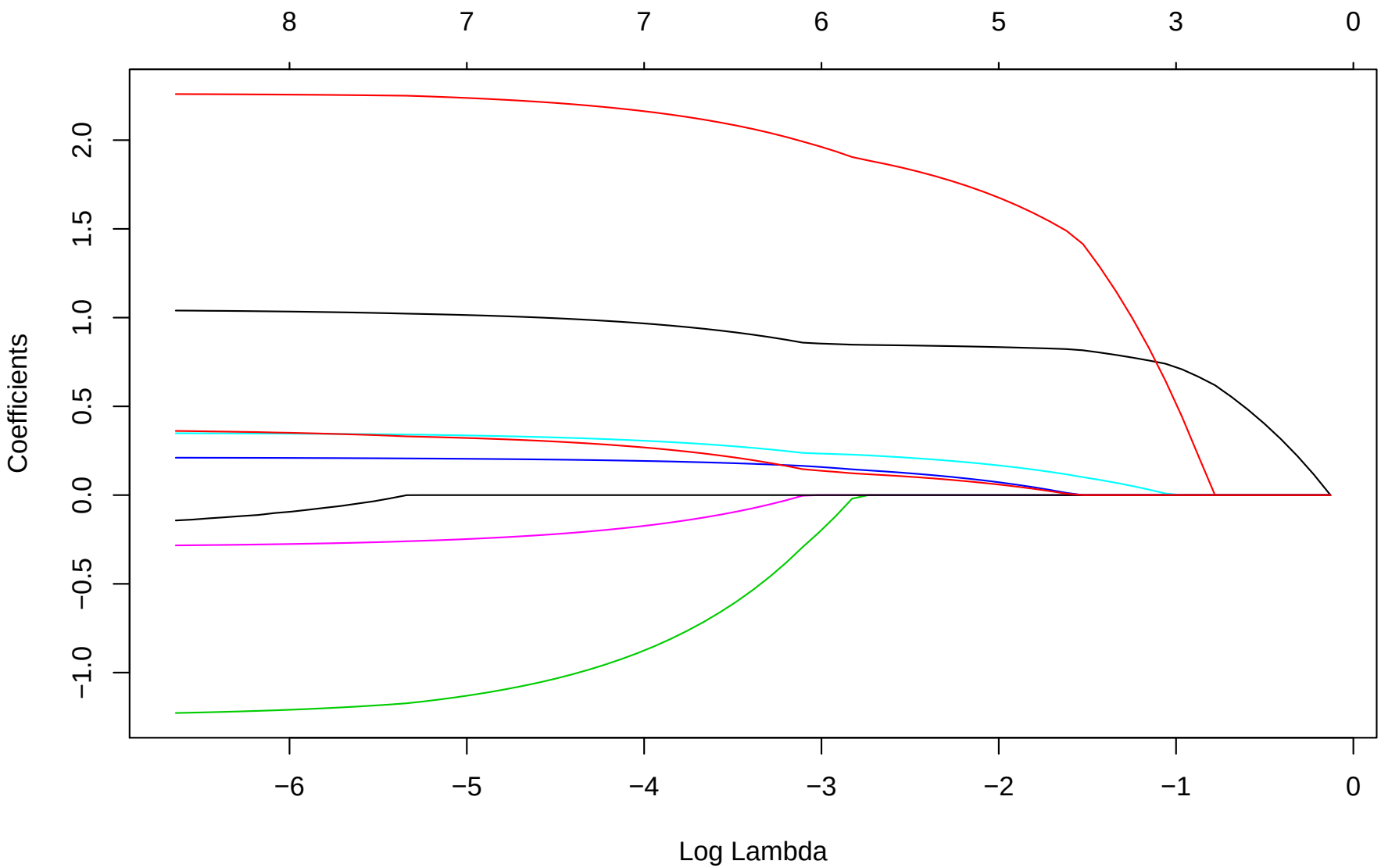
Prediction error on the test set

```
qplot(log10(lasso.path$lambda), err) + geom_line() +
geom_vline(xintercept=log10(lasso.path$lambda[which.min(err)]), lty=3)
```



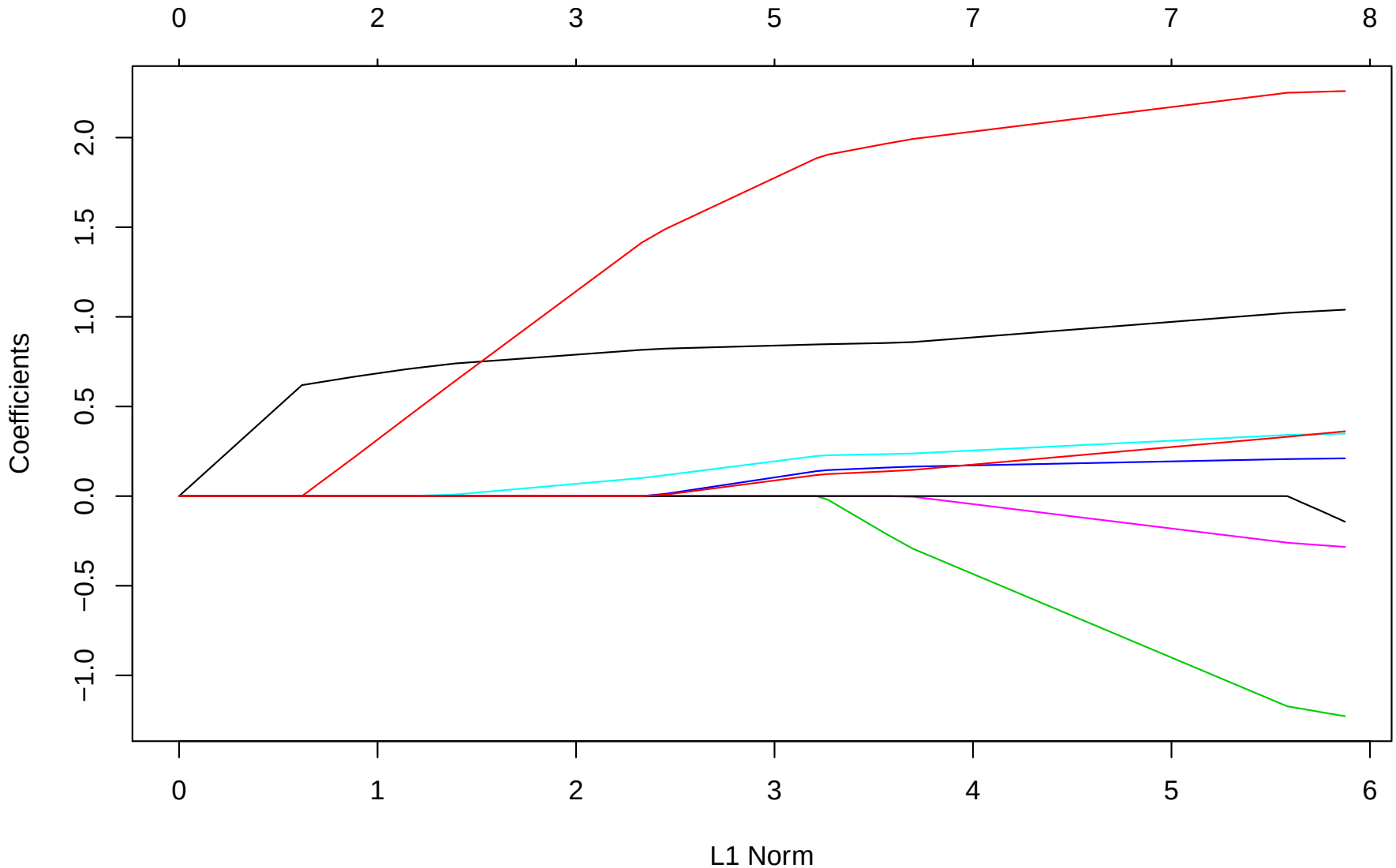
Path of solution (λ)

```
plot(lasso.path, xvar="lambda")
```



Path of solution (amount of shrinkage s)

```
plot(lasso.path, xvar="norm")
```



Outline

Critères pénalisés

LASSO degrees of freedom

It simply equals the number of active (non-null) coefficients)

$$\mathrm{df}(\hat{\mathbf{y}}_\lambda^{\mathsf{lasso}}) = \mathrm{card}(\{j : \beta_j(\lambda) \neq 0\}) = |\mathcal{A}|.$$

- Akaike Information Criterion

$$\mathrm{AIC} = -2\mathrm{loglik} + 2\frac{|\mathcal{A}|}{n},$$

- Bayesian Information Criterion

$$\mathrm{BIC} = -2\mathrm{loglik} + |\mathcal{A}| \log(n),$$

- modified BIC (when $n < p$)

$$\mathrm{mBIC} = -2\mathrm{loglik} + |\mathcal{A}| \log(p),$$

- Extended BIC add a prior on the number of model with size $|\mathcal{A}|$

$$\mathrm{eBIC} = -2\mathrm{loglik} + |\mathcal{A}|(\log(n) + 2\log(p)).$$

Cross-validation

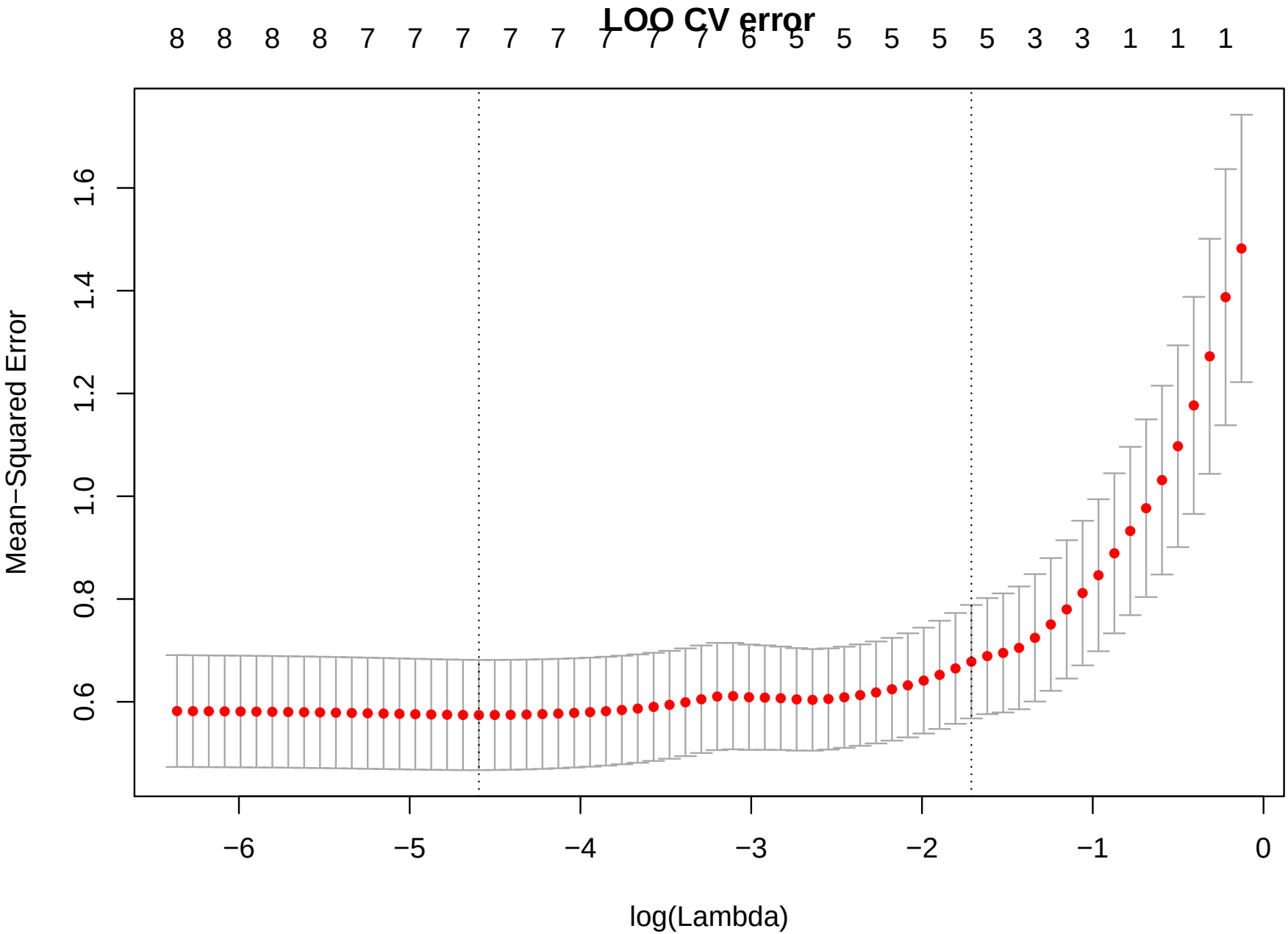
```
system.time(loo <- cv.glmnet(x.train,y.train,nfolds=n))
```

```
##    user  system elapsed
##   0.304   0.000   0.301
```

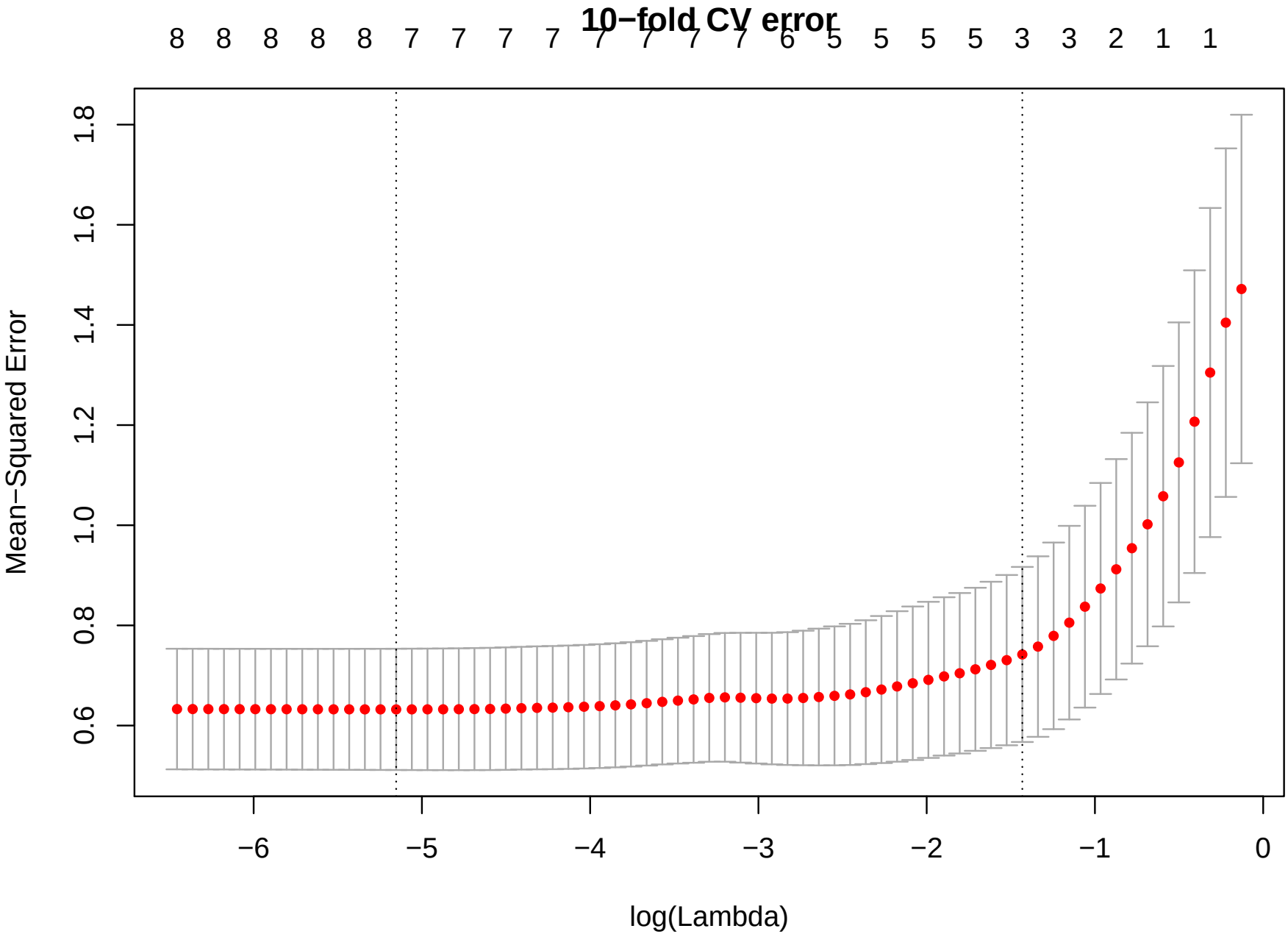
```
system.time(CV10 <- cv.glmnet(x.train,y.train,nfolds=10))
```

```
##    user  system elapsed
##   0.048   0.000   0.049
```

Leave one out



Ten fold


10−fold CV error
8 8 8 8 8 7 7 7 7 7 7 7 7 6 5 5 5 5 3 3 2 1 1
Mean−Squared Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
−6
−5
−4
−3
−2
−1
0
log(Lambda)