MA8701 Advanced methods in statistical inference and learning

W6: Statistical inference for penalized GLM methods

L11 + L2

Mette Langaas

~~2/12/23~~

L11

Lectured 13.02.2023

Before we begin

Outline

- Prediction vs statistics inference: what are the aims?
- Sampling distributions
- Bayesian lasso
- Boostrapping
- Debiased lasso

L11

- Sample splitting
- Inference after selection (forward regression example, polyhedral result, PoSI)
- Reproducibility crisis and selective inference
- Conclusions

L12

Literature

L11

Main source:

- [HTW] Hastie, Tibshirani, Wainwright: "Statistical Learning with Sparsity: The Lasso and Generalizations". CRC press. Ebook. Chapter 6.0, 6.1, 6.2, 6.4, 6.5. (Results from 6.3 through Taylor and Tibshirani (2015))

Also: brush up on **bootstrap intervals from TMA4300**, where Givens and Hoeting (2013) is on the reading list. See specifically chapter 9 (9.2.1 and 9.3 will be used here). NTNU-access to the full book if you are on vpn.

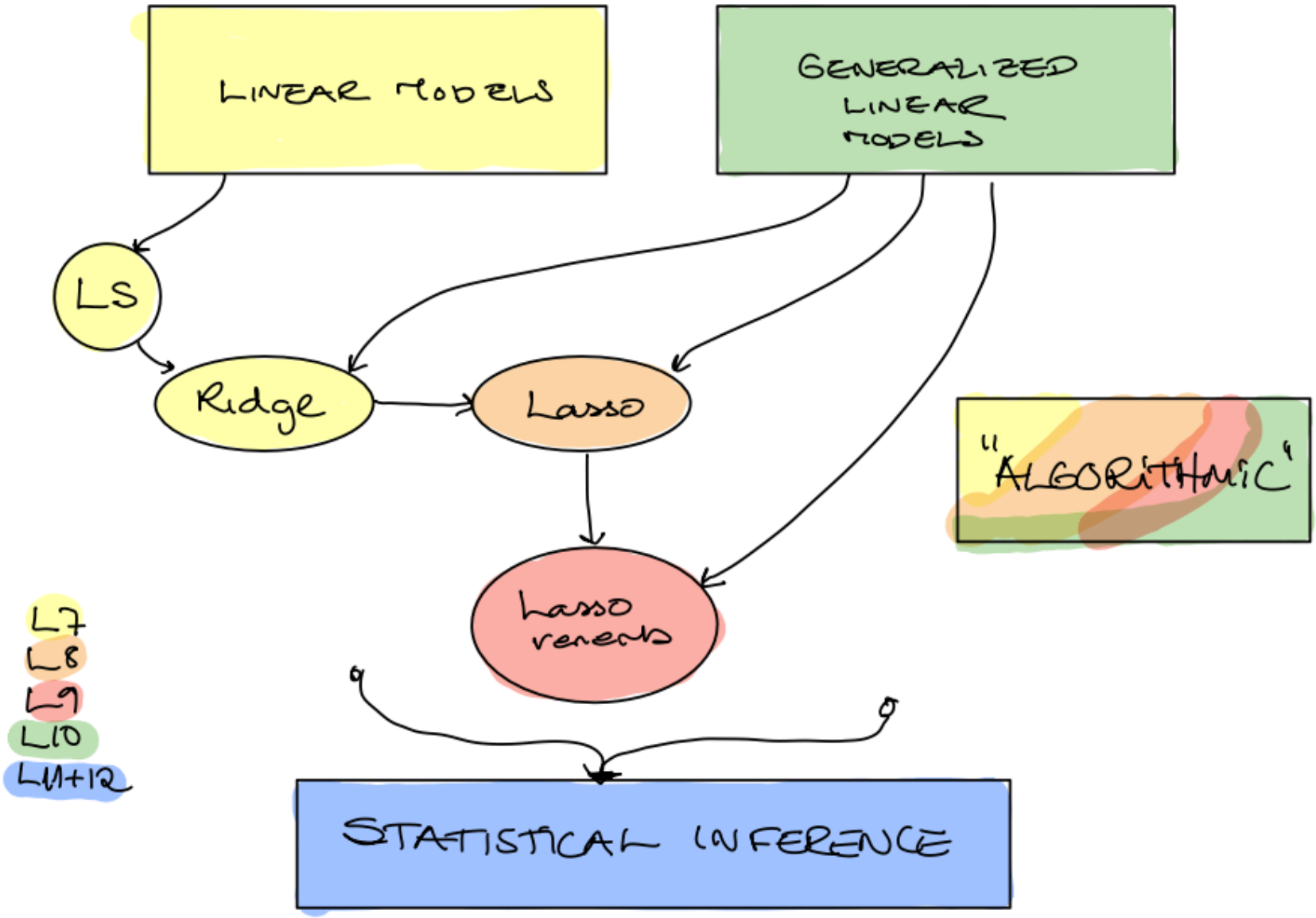
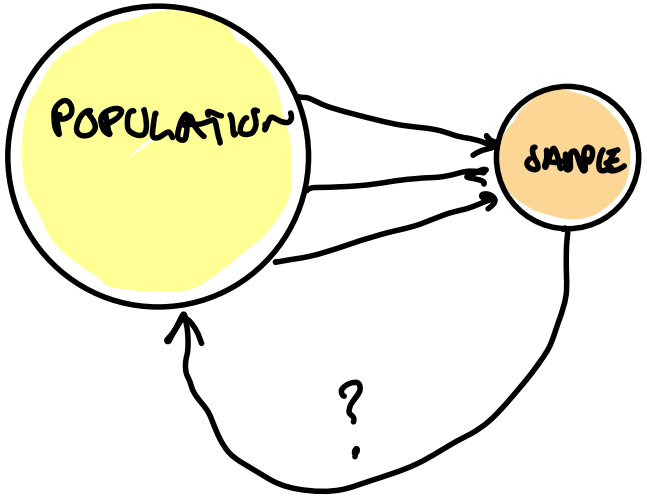


Figure 1: Overview of Part 2

What do we mean by statistical inference here?



- say something about the uncertainty in prediction
- test hypothesis: $\beta_j = 0$? Hypothesis test

point estimation $\hat{\beta}$

CI: $\beta \leftarrow \hat{\beta}$

Bayesian: credibility interval

Statistics vs Machine learning

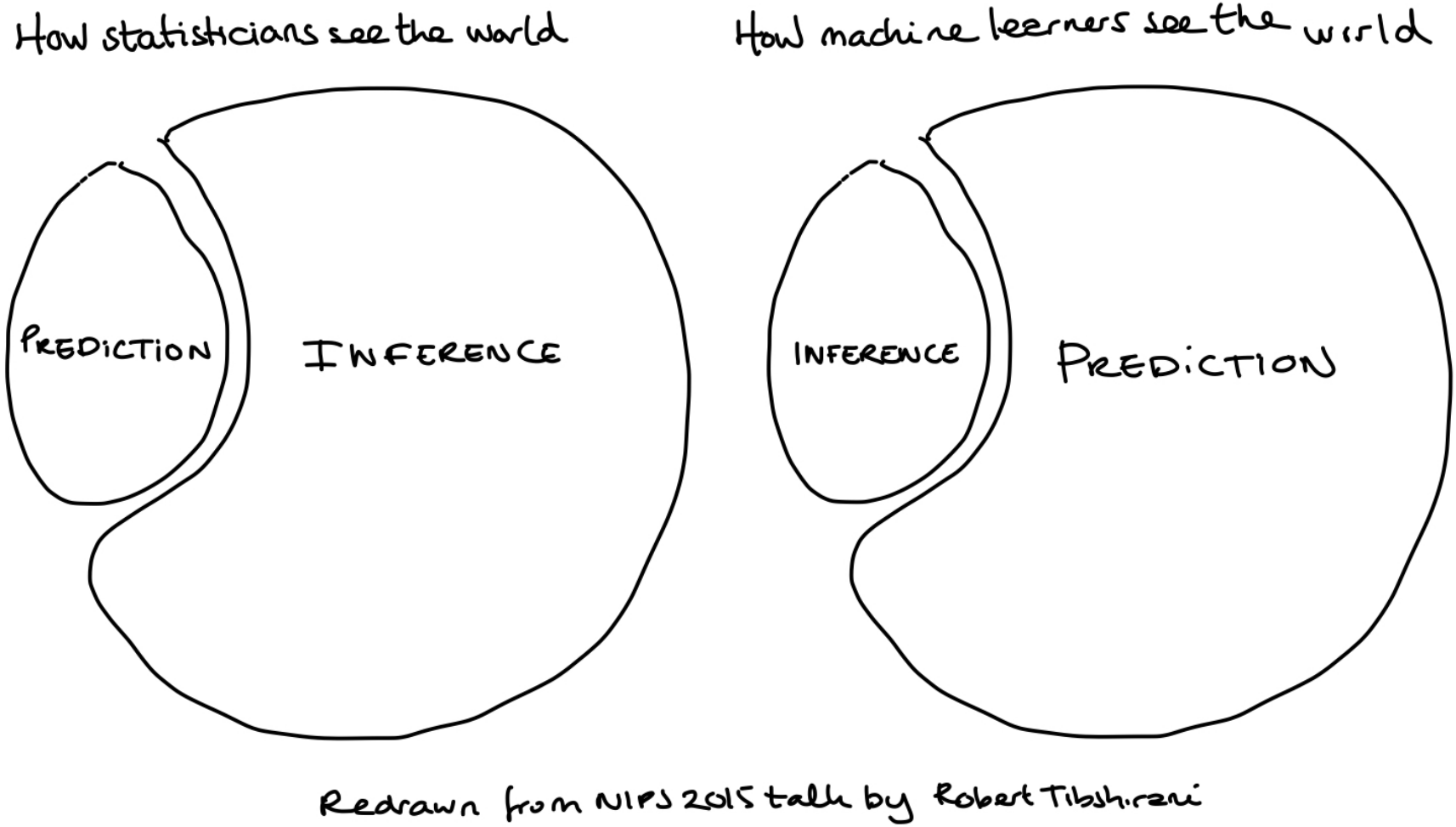


Figure 2: Figures redrawn from Robert Tibshiran´s Breiman lecture at the NIPS 2015
https://www.youtube.com/watch?v=RKQJEvc02hc&t=81s.
(Conference on Neural Information Processing System)

Prediction vs statistical inference

Prediction

- Predict the value of the progression variable for a person with diabetes.
- Predict the probability of heart disease for a person from the population in the South African heart disease example.

Inference

- Assess the goodness of the prediction (MSE, error rate, ROC-AUC) - with uncertainty.
- Interpret the GLM-model - which covariates are included? // selected
- Confidence interval for the model regression parameters.
- Testing hypotheses about the model regression parameters.

Challenge : all possible covariates or only the selected ? (Lasso)

Known sampling distributions

For the linear regression and logistic regression we know the sampling distribution of the regression coefficient estimators. Then it is easy to construct confidence intervals and perform hypothesis tests.

What are the known results?

Multiple linear regression

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim N_N(0, \sigma^2 I)$ (independent observation pairs).

$$\hat{\beta}_{\mathsf{LS}} = (X^T X)^{-1} X^T Y$$

with $\hat{\beta}_{\mathsf{LS}} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$.
Restricted maximum likelihood estimator for $ $\hat{\sigma}^2$$:

$$\hat{\sigma}^2 = \frac{1}{N-p}(Y - X\hat{\beta}_{\mathsf{LS}})^T (Y - X\hat{\beta}_{\mathsf{LS}}) = \frac{\mathsf{SSE}}{N-p}$$

with $\frac{(N-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-p}$.

Statistic for inference about β_j, c_{jj} is diagonal element j of $(X^T X)^{-1}$.

$$T_j = \frac{\hat{\beta}_{\mathsf{LS},j} - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{N-p}$$

or, inference can be done asymptotically and then replace the t with the normal distribution.

T_j is the starting point for constructing CIs for β_j and testing hypotheses about β_j.

Observe: the least squares estimator is *unbiased*!

$$\Rightarrow \left(-t_{\frac{\alpha}{2},N-p} \leq \frac{\hat{\beta} - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \leq t_{\frac{\alpha}{2},N-p} \right) \text{ and solve}$$

Logistic regression

$\widehat{\beta}$ is now not on closed form, but asymptotically when $N \to \infty$

$$\widehat{\beta} \approx N_p(\beta, (X^T \hat{W} X)^{-1})$$

where $W = diag(\hat{\pi}_i(1 - \hat{\pi}_i))$, so that inference can be based on the asymptotic normality of each element of the regression estimate vector.

Observe: the logistic regression parameter estimator is *unbiased*.

Similar with $-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}$

Confidence interval — generic set-up

Set-up

- We have a random sample $Y_1, Y_2, \ldots, Y_N$ from
- some distribution F with some (unkonwn) parameter θ.
- Let $y_1, y_2, \ldots, y_N$ be the observed values for the random sample.

Statistics

- We have two statistics $\hat{\theta}_L(Y_1, Y_2, \ldots, Y_N)$ and $\hat{\theta}_U(Y_1, Y_2, \ldots, Y_N)$ so that

$$P(\hat{\theta}_L(Y_1, Y_2, \ldots, Y_N) \leq \theta \leq \hat{\theta}_U(Y_1, Y_2, \ldots, Y_N)) = 1 - \alpha$$

where $\alpha \in [0, 1]$

→ Coverage of CI : 95% of constructed CI contain true β_j

/ Generate data for model check

Confidence interval

The numerical interval

$$[\hat{\theta}_L(y_1, y_2, \ldots, y_N), \hat{\theta}_U(y_1, y_2, \ldots, y_N)]$$

is called a $(1 - \alpha)$ 100% confidence interval.

Sampling distribution for ridge and lasso?

Multippel linear ridge regression

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T Y$$

$$\hat{\beta}(\lambda)_{\text{ridge}} \sim N\{(X^T X + \lambda I_p)^{-1} X^T X \, \beta,$$

$$\sigma^2 (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1}\}.$$

how to estimate? $\frac{\text{residuals}^2}{df(\lambda)}$

$$\text{df}(\lambda) = \text{tr}(H_\lambda) = \text{tr}(X(X^T X + \lambda I)^{-1} X^T) = \cdots = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

- What can we do with that?
- What if the design matrix is orthogonal, does that help?

Orthogonal: $\hat{\beta}_R = \frac{1}{1+\lambda}\hat{\beta}_{LS}$, $E(\hat{\beta}_R) = \frac{1}{1+\lambda}\beta$

Element j

$$Cov(\hat{\beta}_R) = \frac{\sigma^2}{(1+\lambda)^2} I$$

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{\beta}_{Rj} - \frac{1}{1+\lambda}\beta_j}{\frac{\sigma}{1+\lambda}} \leq z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$\hat{\beta}_{Rj} - z_{\frac{\alpha}{2}}\frac{\sigma}{1+\lambda} \leq \frac{1}{1+\lambda}\beta_j \leq \hat{\beta}_{Rj} + z_{\frac{\alpha}{2}}\frac{\sigma}{1+\lambda}$$

$$(1+\lambda)\hat{\beta}_{Rj} - z_{\frac{\alpha}{2}}\sigma \leq \beta_j \leq (1+\lambda)\hat{\beta}_{Rj} + z_{\frac{\alpha}{2}}\sigma$$

$(1+\lambda)\hat{\beta}_{Rj} = \hat{\beta}_{LS}$

as LS, with $(X^TX)^{-1} = I$

$$\hat{\beta}_{ridge} \sim N(A\beta, \sigma^2 C)$$

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_p^T \end{bmatrix} \qquad [A\beta]_j = a_j^T \beta$$

$$\hat{\beta}_{ridge,j} \sim N(a_j^T \beta, \sigma^2 C_{jj})$$

↑ not only a function of β_j

Logistic ridge

For large sample sizes the ridge logistic regression estimator is approximately multivariate normal (Wieringen (2021) Section 5.3).

$$\hat{\beta}(\lambda) \approx N_p(\beta - \lambda(X^T \hat{W} X + \lambda I)^{-1}\beta),$$

$$(X^T \hat{W} X + \lambda I)^{-1} - (X^T \hat{W} X + \lambda I)^{-2})$$

This is based on the asymptotic normality of the score function (gradient of the loglikelihood - here the penalized loglikelihood).

- What can we do with that?
- What if the design matrix is orthogonal, does that help?

Same as above

Lasso

Some results using approximations to ridge (for mean and variance, see Wieringen (2021) p 97), but else *no parametric version of sampling distribution* known.

What is our aim?

Penalized estimation: reduce variance by introducing (strong) bias.
The squared bias then is a major part of the mean squared error, and the variance is thus a minor part.
But, do we need to use the ridge or lasso estimator to construct a confidence interval for β or test if $\beta_j = 0$?

- $p > N$ → ?
- $p < N$? → just use LS to make CI

Debiased (desparsified) lasso

(HTW Section 6.4)

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{N} M X^T (Y - X\hat{\beta}_\lambda)$$

the matrix M is some approximation to the inverse of

$$\hat{\Sigma} = \frac{1}{N} X^T X$$

Use the debiased estimator to form CI from:

$$\hat{\beta}^d \sim N(\beta, \frac{\sigma^2}{N} M \hat{\Sigma}^{-1} M^T)$$

Interpretation of debiasing: assume we want to minimize the residual sum of squares using an approximate Newton step starting at the lasso estimator.

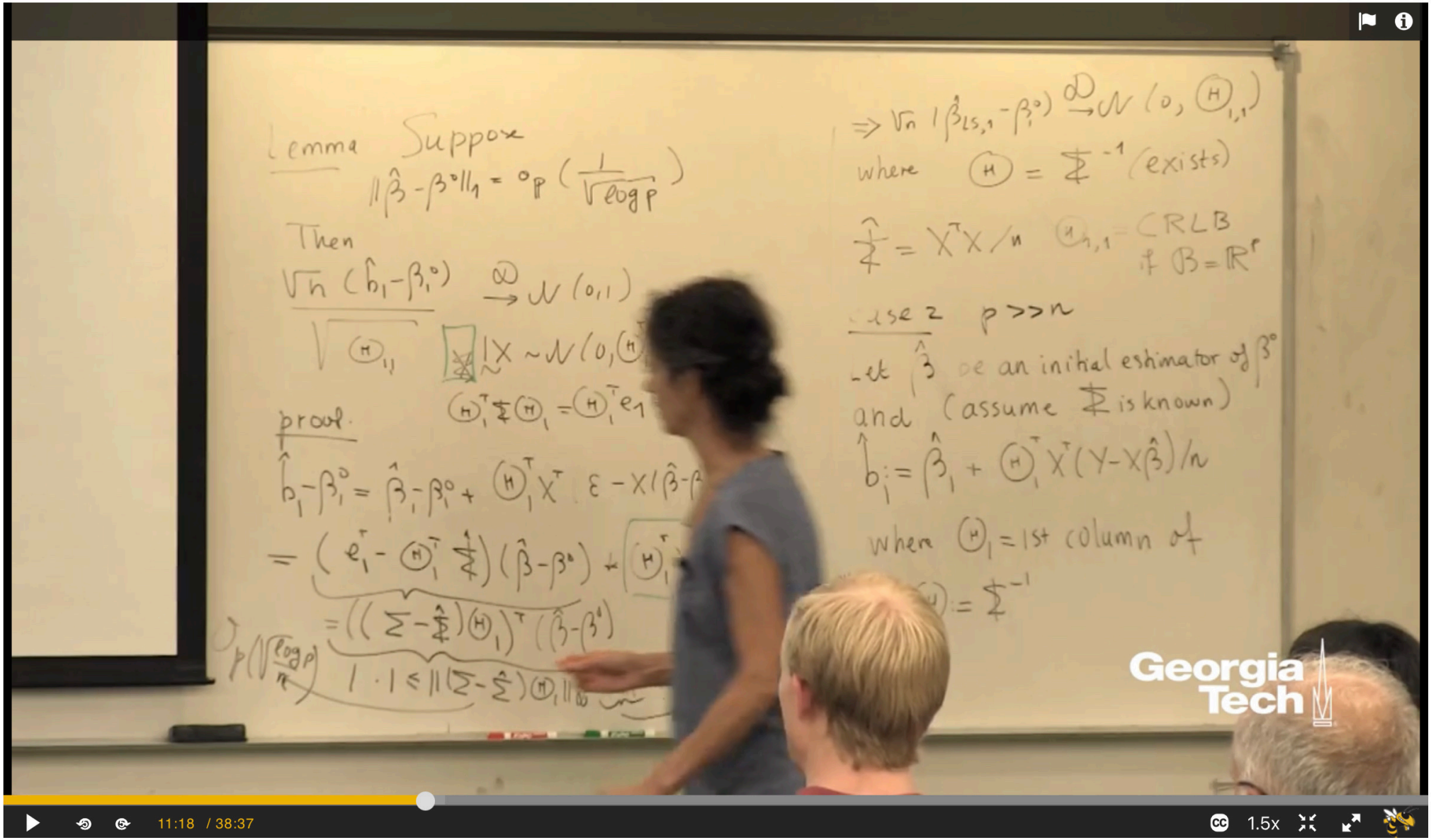
Why this equation:

$$\hat{\Sigma} = \frac{1}{N} X^T X \quad \text{if} \quad M = N \cdot (X^T X)^{-1} = \hat{\Sigma}^{-1}$$

$$\hat{\beta}^d = \hat{\beta}_2 + \frac{1}{N} N (X^T X)^{-1} X^T (y - X \hat{\beta}_2)$$

$$= \hat{\beta}_2 + (X^T X)^{-1} X^T y - \underbrace{(X^T X)^{-1} X^T X}_{I} \hat{\beta}_2$$

$$= (X^T X)^{-1} X^T y = \hat{\beta}_{LS}$$



https://mediaspace.gatech.edu/media/Sara+van+de+Geer+-+The+Debiased+Lasso/1_2d0tbi58

Two solutions are based on

- neighbour-based methods to impose sparsity
- optimzation problem to get $\hat{M}$ such that $\hat{\Sigma}\hat{M} \approx I$ while the variance of the debiased estimator is small.

See Hastie, Tibshirani, and Wainwright (2015) page 159 for references to these solutions.

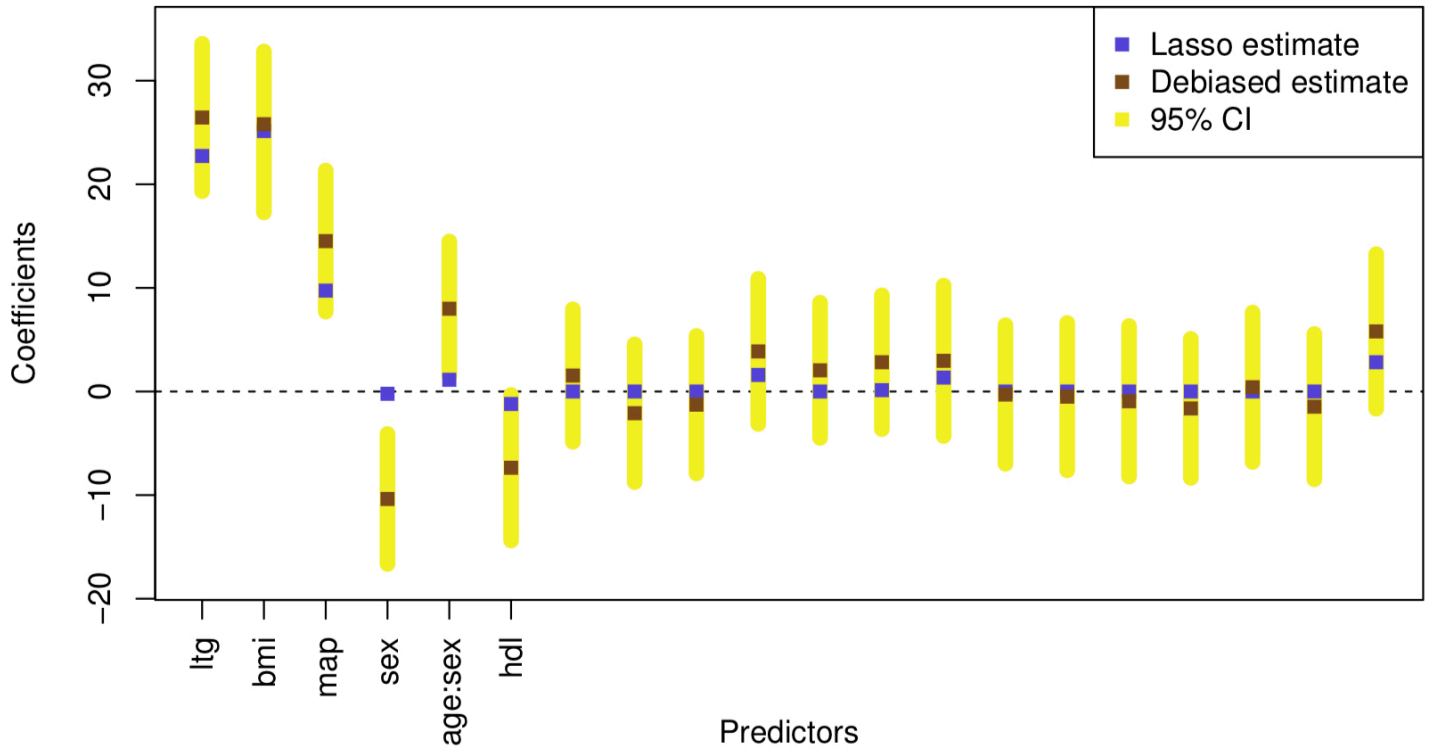


Figure 6.13 *Diabetes data: Lasso estimates, debiased lasso estimates, and confidence intervals from the debiasing approach. These intervals have not been adjusted for multiple comparisons. The first 6 predictors have intervals not containing zero: when Bonferroni-adjusted, this number drops to three.*

Figure 3: Figure 6.13 in Hastie, Tibshirani, and Wainwright (2015)

Conclusion

This is absolutely not straightforward.

The *adaptive and biased nature* of the estimation procedures makes it challenging to perform inference.

- We will *discuss* other (in addition to the debiasing) solutions to finding confidence intervals for regression parameters for lasso, and for constructing p-values for testing hypotheses about the regression parameters.
- We will address some *philosophical principles behind inference*
- and mention topics that can be studied further for the interested student!

Warning: there seems not to be consensus, but many interesting approaches and ideas that we may consider.

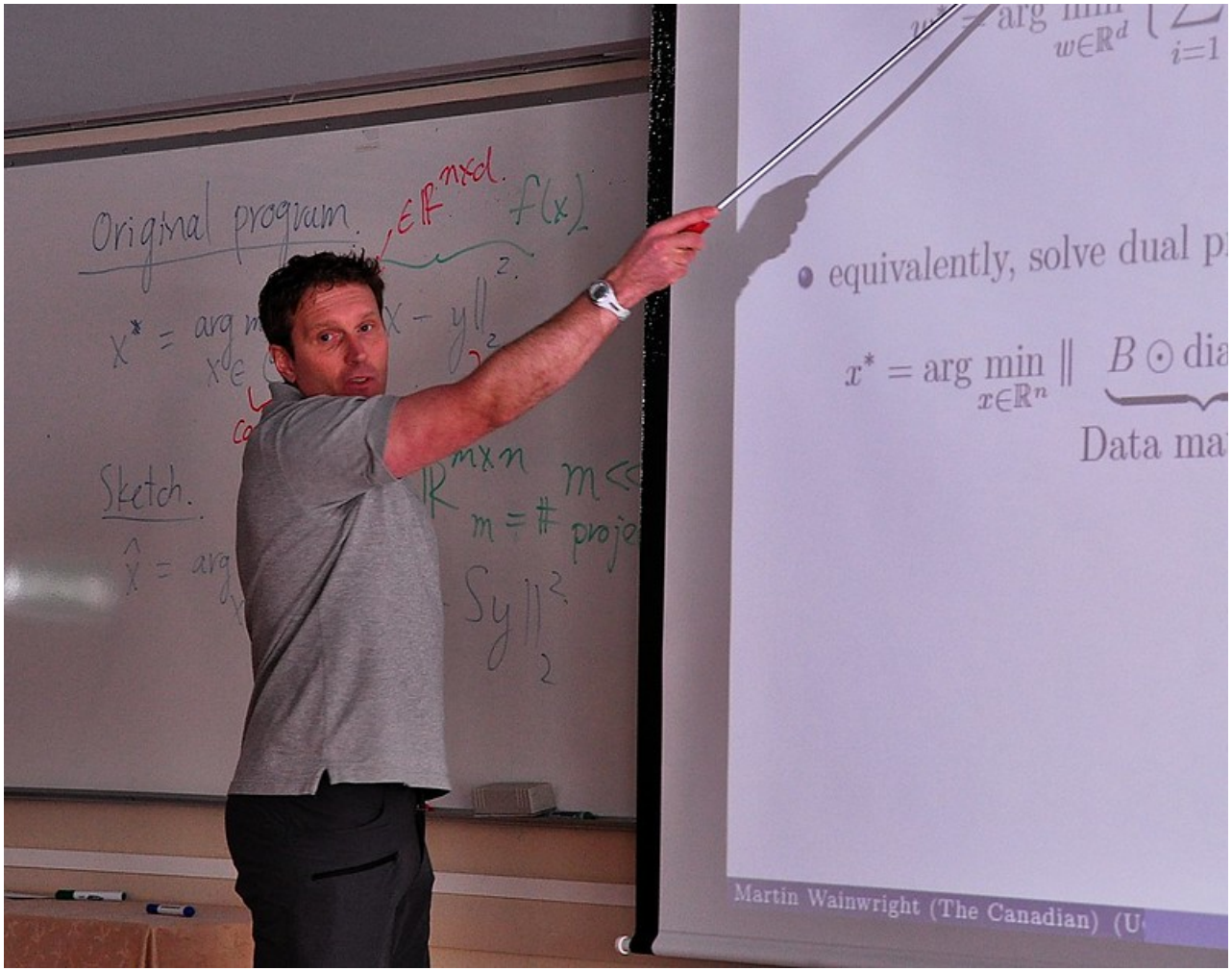
Statistical Science
2021, Vol. 36, No. 4, 562–577
https://doi.org/10.1214/20-STS815

In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference

Sen Zhao, Daniela Witten and Ali Shojaie

Abstract. A great deal of interest has recently focused on conducting inference on the parameters in a high-dimensional linear model. In this paper, we consider a simple and very naïve two-step procedure for this task, in which we (i) fit a lasso model in order to obtain a subset of the variables, and (ii) fit a least squares model on the lasso-selected set. Conventional statistical wisdom tells us that we cannot make use of the standard statistical inference tools for the resulting least squares model (such as confidence intervals and p-values), since we peeked at the data twice: once in running the lasso, and again in fitting the least squares model. However, in this paper, we show that under a certain set of assumptions, with high probability, the set of variables selected by the lasso is identical to the one selected by the noiseless lasso and is hence deterministic. Consequently, the naïve two-step approach can yield asymptotically valid inference. We utilize this finding to develop the *naïve confidence interval*, which can be used to draw inference on the regression coefficients of the model selected by the lasso, as well as the *naïve score test*, which can be used to test the hypotheses regarding the full-model regression coefficients.

Key words and phrases: Confidence interval, lasso, p-value, post-selection inference, significance testing.



2014 Abel symposium on high dimensional data

Martin Wainwright

Trevor Hastie

Diabetes data

In a medical study the aim was to explain the ethiology of diabetes progression. Data was collected from $n = 442$ diabetes patients, and from each patient the following measurements are available:

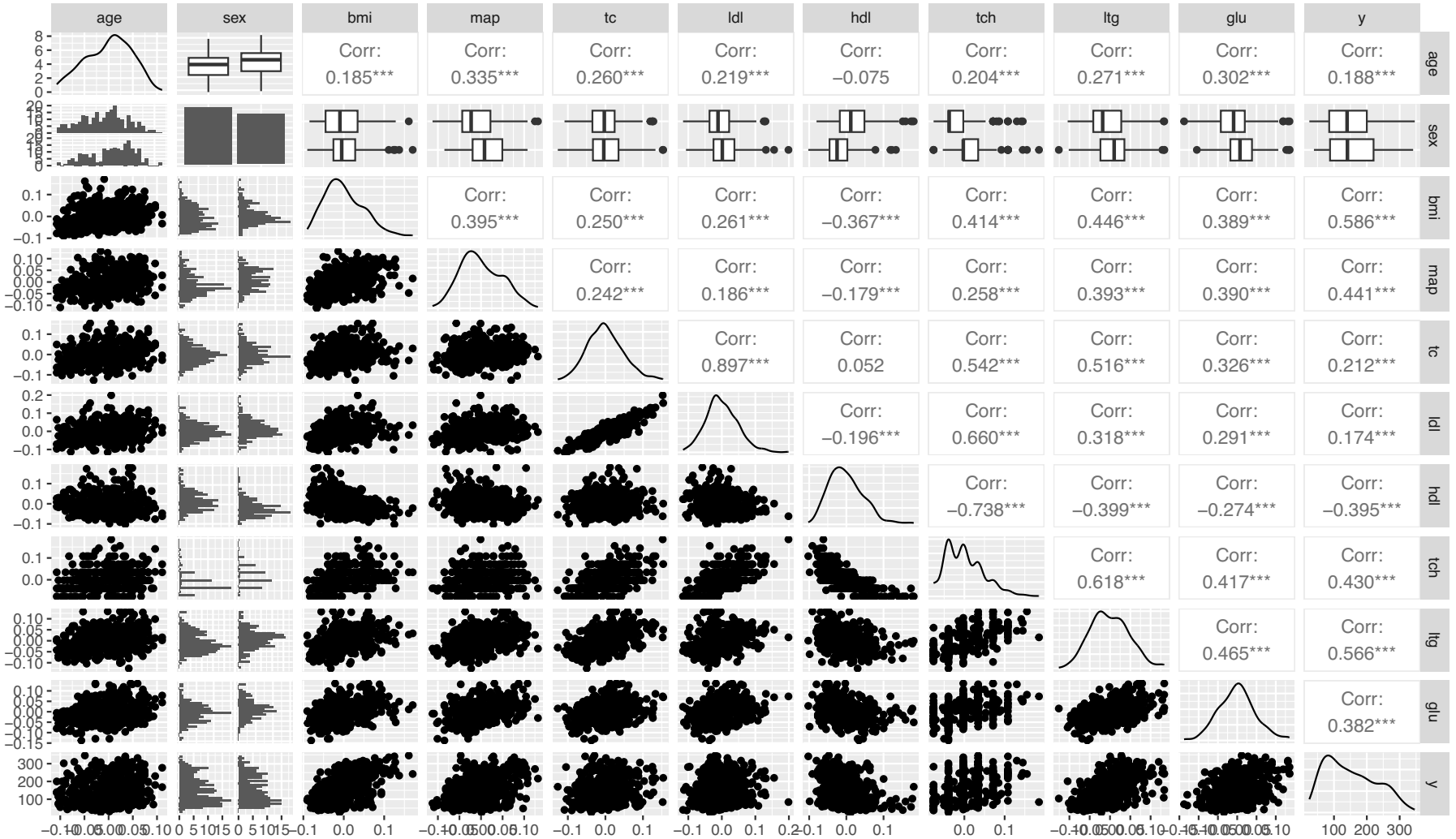
- age (in years) at baseline
- `sex` (0=female and 1=male) at baseline
- body mass index (`bmi`) at baseline
- mean arterial blood pressure (`map`) at baseline
- six blood serum measurements: total cholesterol (`tc`), ldl cholesterol (`ldl`), hdl cholesterol (`hdl`), `tch`, `ltg`, glucose `glu`, all at baseline,
- a quantitative measurement of disease progression one year after baseline (`prog`)

All measurements except `sex` are continuous. There are 10 covariates.

The response is the disease progression `prog` - thus a regression problem.

Data can be

- downloaded from https://web.stanford.edu/~hastie/StatLearnSparsity_files/DATA/diabetes.html in three variants: raw, standardized and 442×64 matrix with quadratic terms (not used here).
- Or, loaded from the `lars` package, that is automatically loaded in the `monomvn` package (where `blasso` is found).


age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu
y
Corr: 0.185***
Corr: 0.335***
Corr: 0.260***
Corr: 0.219***
Corr: −0.075
Corr: 0.204***
Corr: 0.271***
Corr: 0.302***
Corr: 0.188***
Corr: 0.395***
Corr: 0.250***
Corr: 0.261***
Corr: −0.367***
Corr: 0.414***
Corr: 0.446***
Corr: 0.389***
Corr: 0.586***
Corr: 0.242***
Corr: 0.186***
Corr: −0.179***
Corr: 0.258***
Corr: 0.393***
Corr: 0.390***
Corr: 0.441***
Corr: 0.897***
Corr: 0.052
Corr: 0.542***
Corr: 0.516***
Corr: 0.326***
Corr: 0.212***
Corr: −0.196***
Corr: 0.660***
Corr: 0.318***
Corr: 0.291***
Corr: 0.174***
Corr: −0.738***
Corr: −0.399***
Corr: −0.274***
Corr: −0.395***
Corr: 0.618***
Corr: 0.417***
Corr: 0.430***
Corr: 0.465***
Corr: 0.566***
Corr: 0.382***

Bayesian ridge and lasso

→ may lead to insight + another solution

(HTW 6.1 for lasso, WNvW Section 5.5 and 6.6)

For penalized models there exists Bayesian equivalents. We will here focus on the multiple linear regression model.

Bayes $$f(\beta \mid y, X) = \frac{f(y \mid X, \beta)\, f(\beta)}{f(y \mid X)}$$

data, likelihood, prior

$f(y \mid X)$: norm const not dep on β

Bayesian set-up

In the Bayesian statistics the regression parameters β are random quantities, and in addition to the likelihood also a prior for the regression parameters (and other parameters) are needed. When a *conjugate prior* the posterior distribution may be derived analytically.

Multiple linear regression: distribution of response - where we for simplicity assume that we have centred covariates and centred response (so no intercept term)

$$y \mid \beta, \sigma \sim N(X\beta, \sigma^2 I)$$

This gives the likelihood:

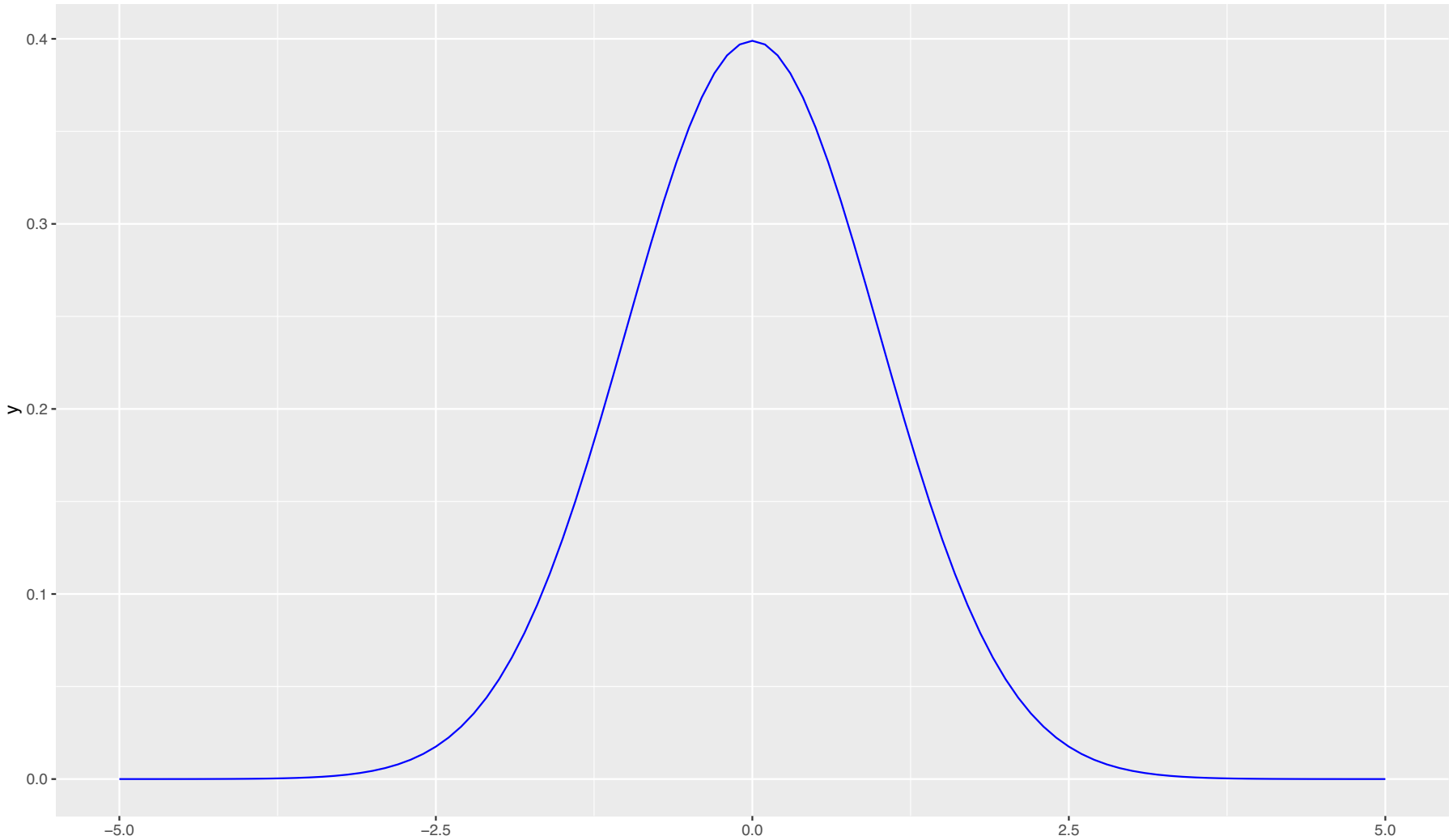
$$L(\beta \mid y, X, \sigma) \propto (\sigma^{-N/2}) \exp[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)]$$

Prior for regression parameters (ridge)

In Part 1 we worked with multiple imputation and one method for drawing observations was the Bayesian linear regression.
We will use the same priors here, for the σ^2 we use a inverse Gamma prior. For the regression coefficents a normal prior is used.

$$\beta \mid \sigma \sim \prod_{j=1}^{p} \sqrt{\frac{\lambda}{2\sigma^2}} \exp(-\frac{\lambda}{2\sigma^2}\beta_j^2)$$

and inverse gamma for σ


0.4
0.3
0.2
0.1
0.0
y
−5.0
−2.5
0.0
2.5
5.0

Posterior for regression parameters (ridge)

$$\beta, \mid X, Y, \sigma^2 \propto \exp[-\frac{1}{2\sigma^2}](\beta - \hat{\beta}(\lambda))^T(X^TX + \lambda I)(\beta - \hat{\beta}(\lambda))]$$

The posterior mean is the ridge estimator $\hat{\beta}(\lambda)$.

↓ we did.

Bayesian (multiple) imputation in MLR

Likelihood: $p(y \mid X, \phi, \sigma^2) \propto (\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}(y-X\phi)^T(y-X\phi)\right)$

"then posterior derived analytically"

Conjugate prior: $p(\phi, \sigma^2) = p(\phi \mid \sigma^2) \cdot p(\sigma^2)$

Bayesian like this!

$p(\phi \mid \sigma^2) \sim N(\mu_0, \sigma^2 \Lambda_0^{-1})$ — Λ_0 is precision matrix = inverse of covariance matrix

$p(\sigma^2) \sim$ inverse-gamma

Posterior: $p(\phi, \sigma^2 \mid y, X) \propto \underbrace{p(y \mid X, \phi, \sigma^2)}_{\text{likelihood}} \cdot \overbrace{p(\phi \mid \sigma^2)\, p(\sigma^2)}^{\text{prior}}$

Some rearrangement:

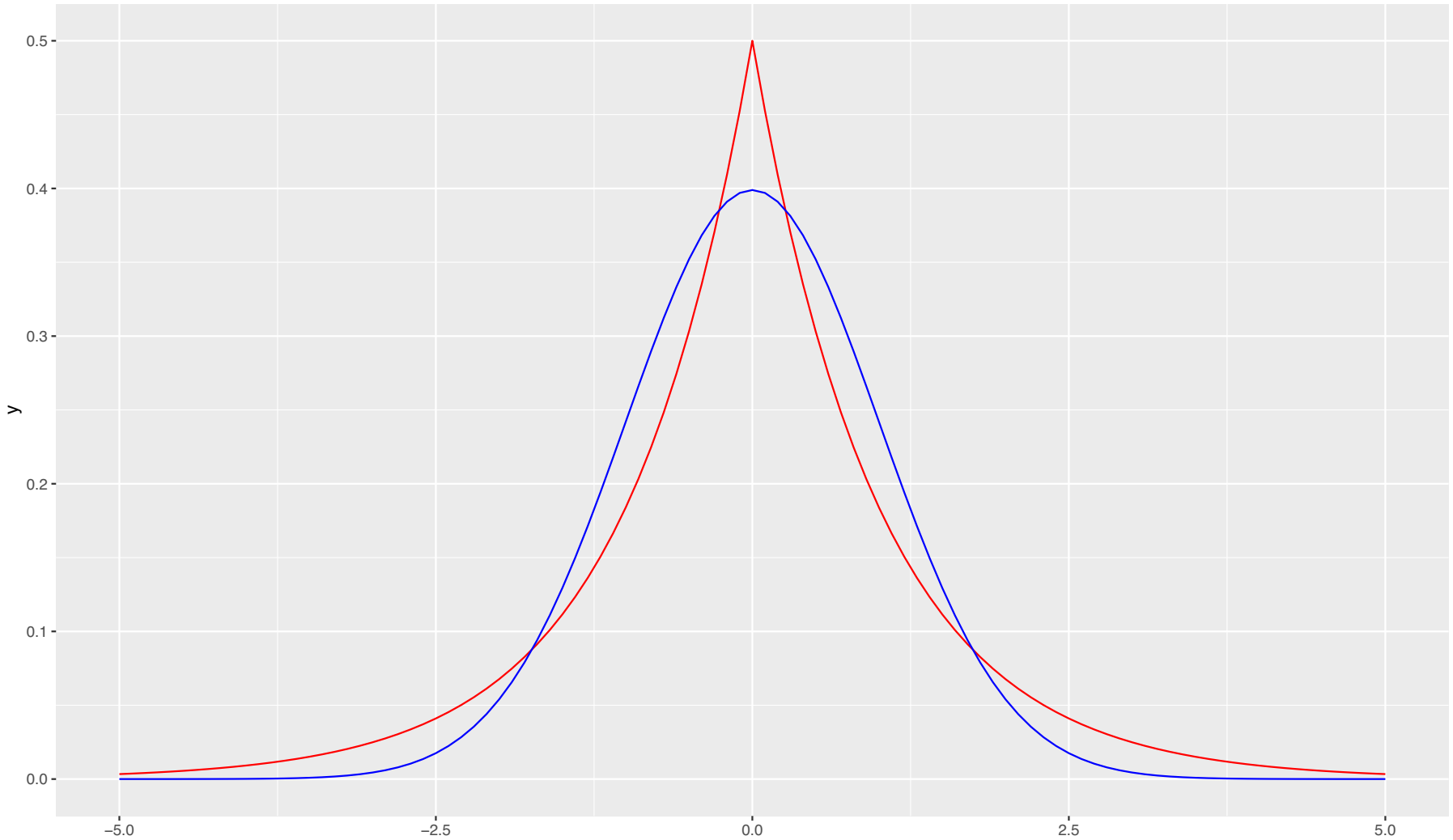
$E(\phi, \sigma^2 \mid y, X)$: the ϕ part: $\mu_N = (\underbrace{X^TX + \Lambda_0}_{\Lambda_N})^{-1}(X^TX\hat{\phi} + \Lambda_0\mu_0)$ — $\hat{\phi}$: LS est.

$\begin{bmatrix} \mu_N \\ E\sigma^2 \end{bmatrix}$ | $\mathrm{Var}(\phi, \sigma^2 \mid y, X) = \sigma^2 \Lambda_n^{-1}$

Prior for regression parameters (lasso)

$$\beta \mid \lambda, \sigma \sim \prod_{j=1}^{p} \frac{\lambda}{2\sigma} \exp(-\frac{\lambda}{\sigma}|\beta_j|)$$

This prior is called an i.i.d. *Laplacian* (or double exponential) prior.


0.5
0.4
0.3
0.2
0.1
0.0
y
−5.0
−2.5
0.0
2.5
5.0

Posterior for regression parameters (lasso)

It can be shown that the negative log of the posterior density for $\beta \mid y, \lambda, \sigma$ is (up to an additive constant)

$$\frac{1}{2\sigma^2}\|y - X\beta\|_2^2 + \frac{\lambda}{\sigma}\|\beta\|_1 \qquad \cdot \sigma^2$$

Does this look familiar?

$\downarrow$

$\sigma\lambda$ is and λ

For a fixed value of σ and λ - the β giving the minimum of the negative log posterior is the *lasso* estimate where the regularization parameter is $\sigma\lambda$.

The minimum negative log posterior will then be the same as the maximum log posterior - and the maximum of a distribution is called the *mode* of the distribution.

The lasso estimate is the *posterior mode* in the Bayesian model.

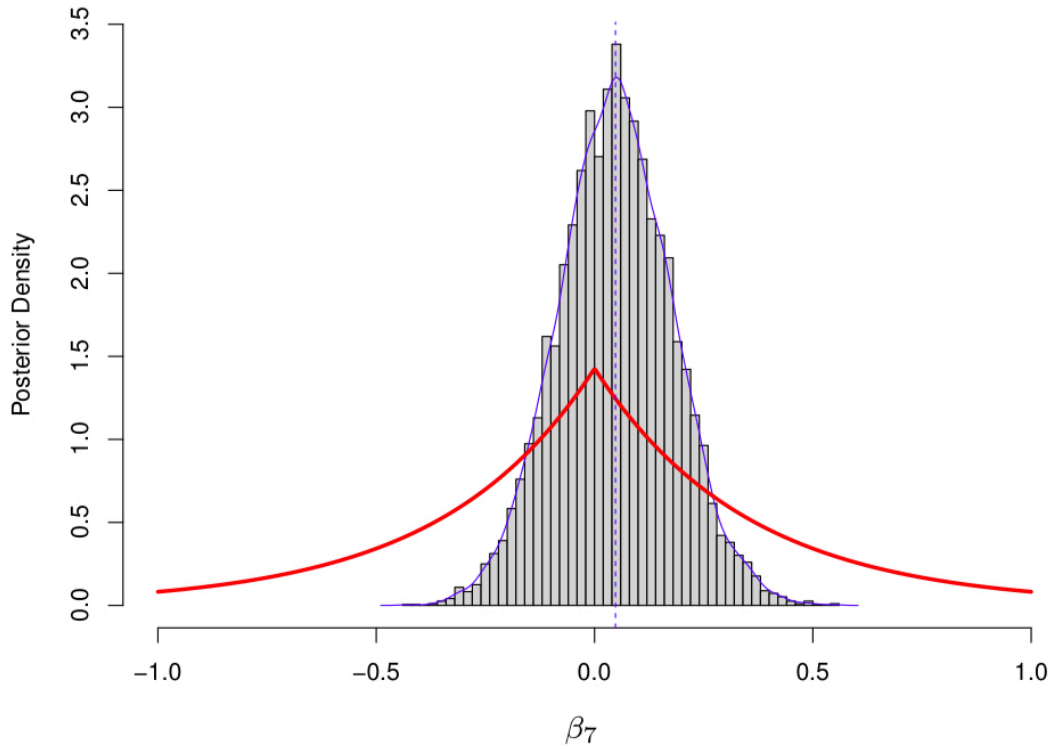


Figure 6.1 *Prior and posterior distribution for the seventh variable in the diabetes example, with λ held fixed. The prior in the figure is a double exponential (Laplace) distribution with density proportional to $\exp(-.0065|\beta_7|)$. The prior rate .0065 is a representative value just for illustration.*

Figure 4: Figure 6.1 in Hastie, Tibshirani, and Wainwright (2015)

From Hastie, Tibshirani, and Wainwright (2015): a 95% posterior credibility interval covers zero.

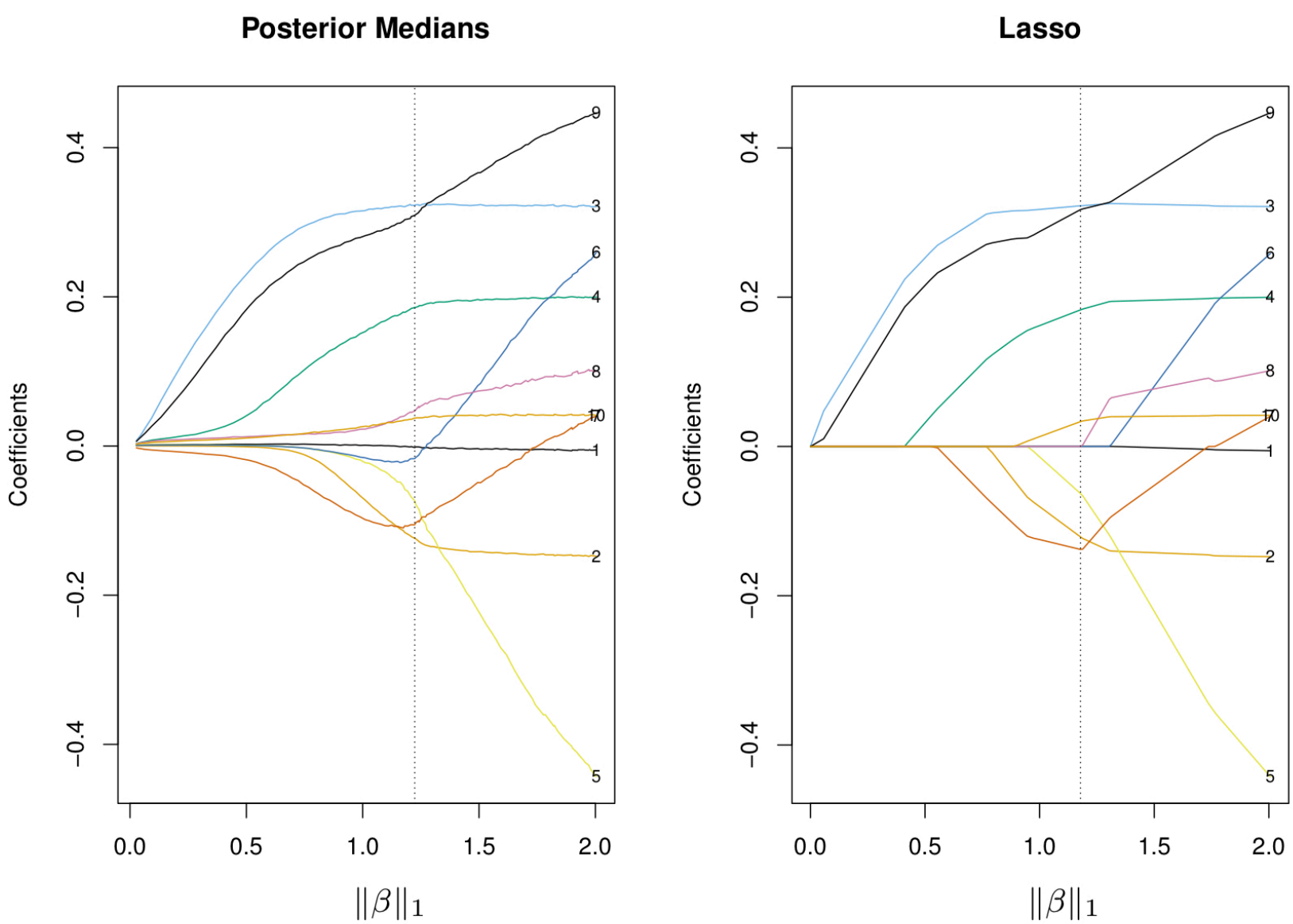


Figure 6.2 *Bayesian lasso on the diabetes data. The left plot shows the posterior medians from MCMC runs (conditional on λ). The right plot shows the lasso profile. In the left plot, the vertical line is at the posterior median of $\|\beta\|_1$ (from an unconditional model), while for the right plot the vertical line was found by N-fold cross-validation.*

Look similar

A full Bayesian approach requires priors for λ and σ, in addition to priors on the regression coefficient.

Markov Chain Monte Carlo MCMC is used efficiently sample realizations form the posterior distribution.

See Wieringen (2021) Chapter 2 for more on Bayesian regression and the connection to the ridge and Section 6.6 for connection to lasso.

Not only the point estimate

The posterior distribution gives the

- point estimates for the lasso (the mode of the distribution)

but

- also the *entire joint distribution*.

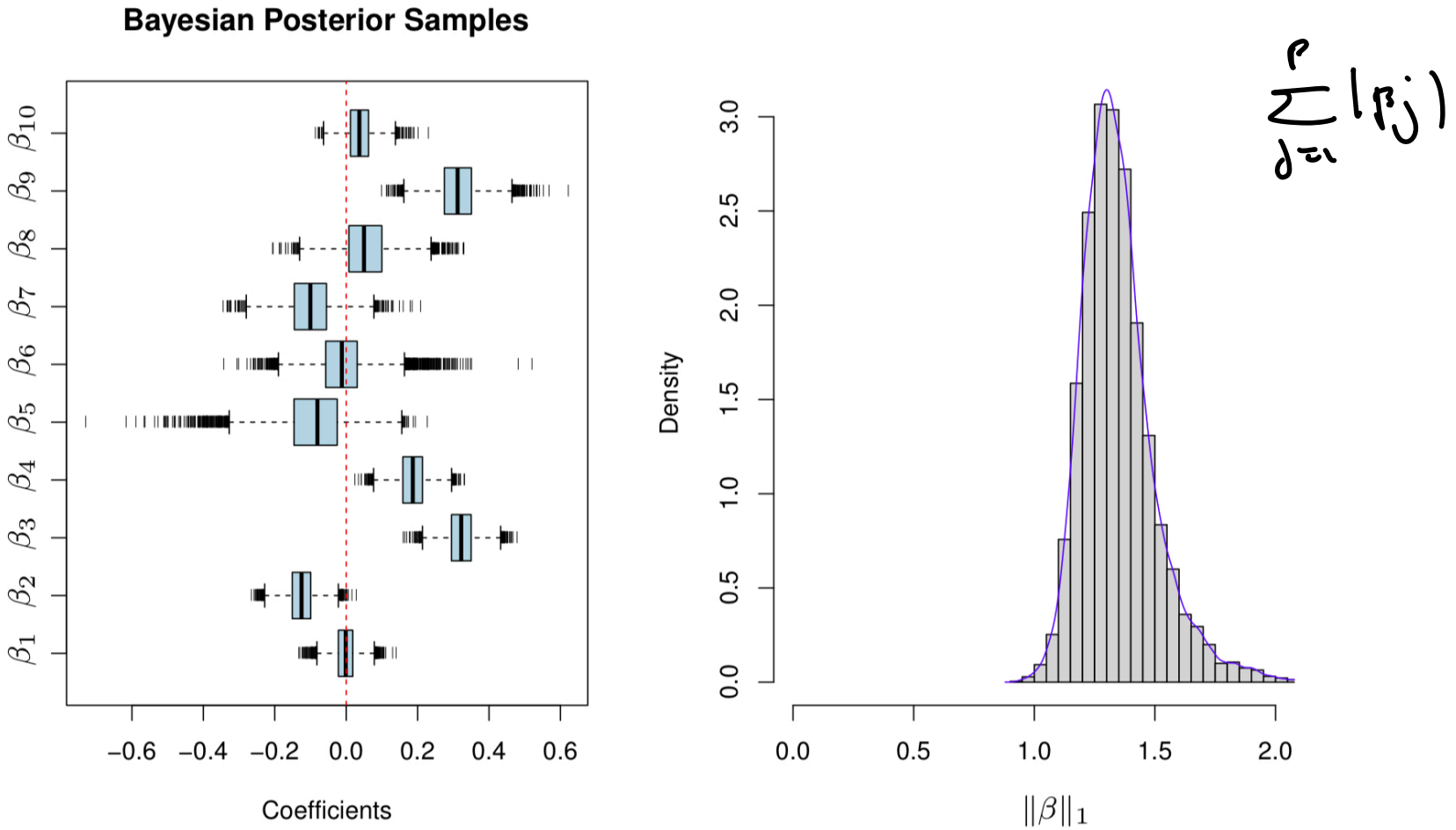
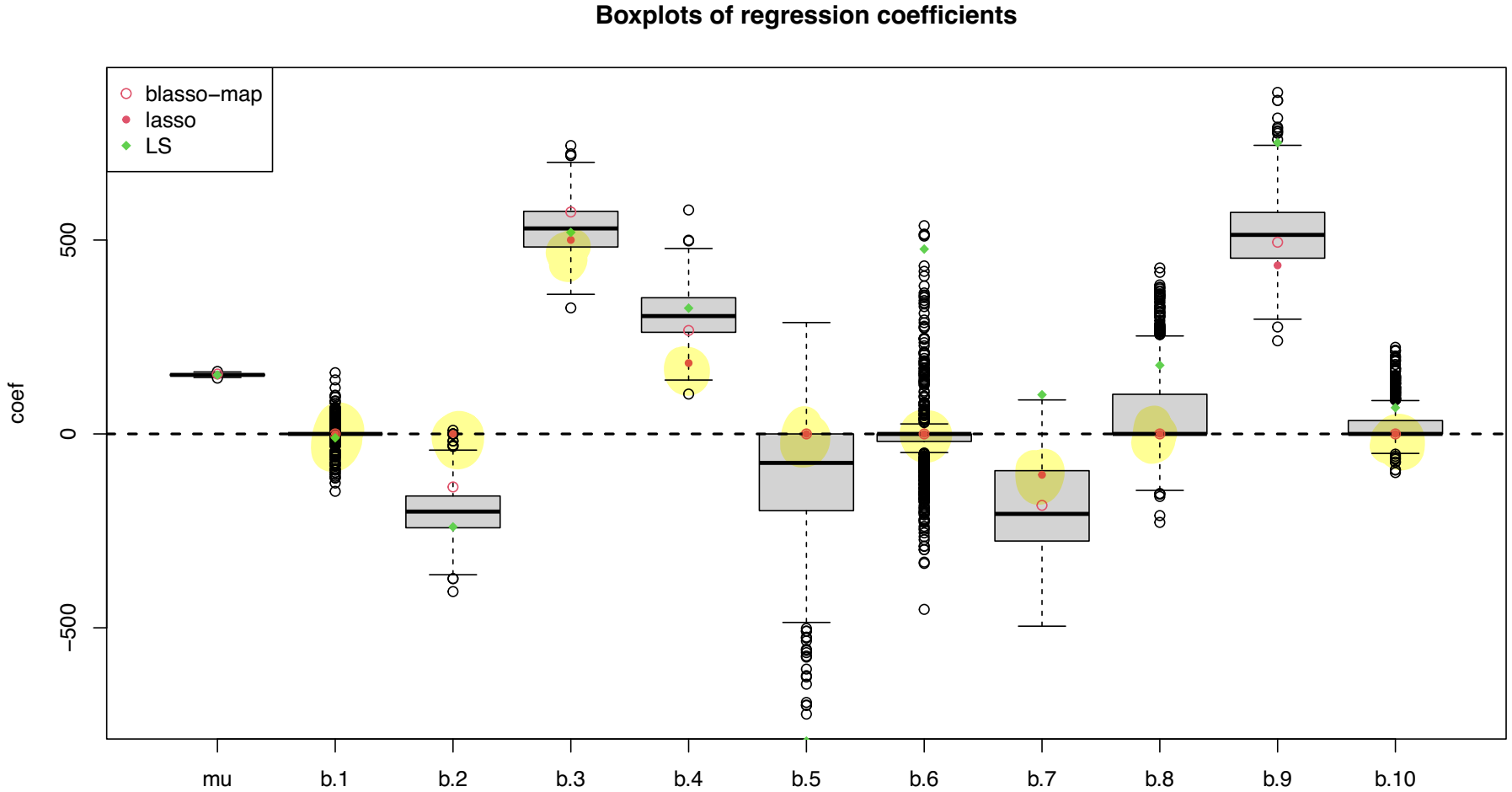
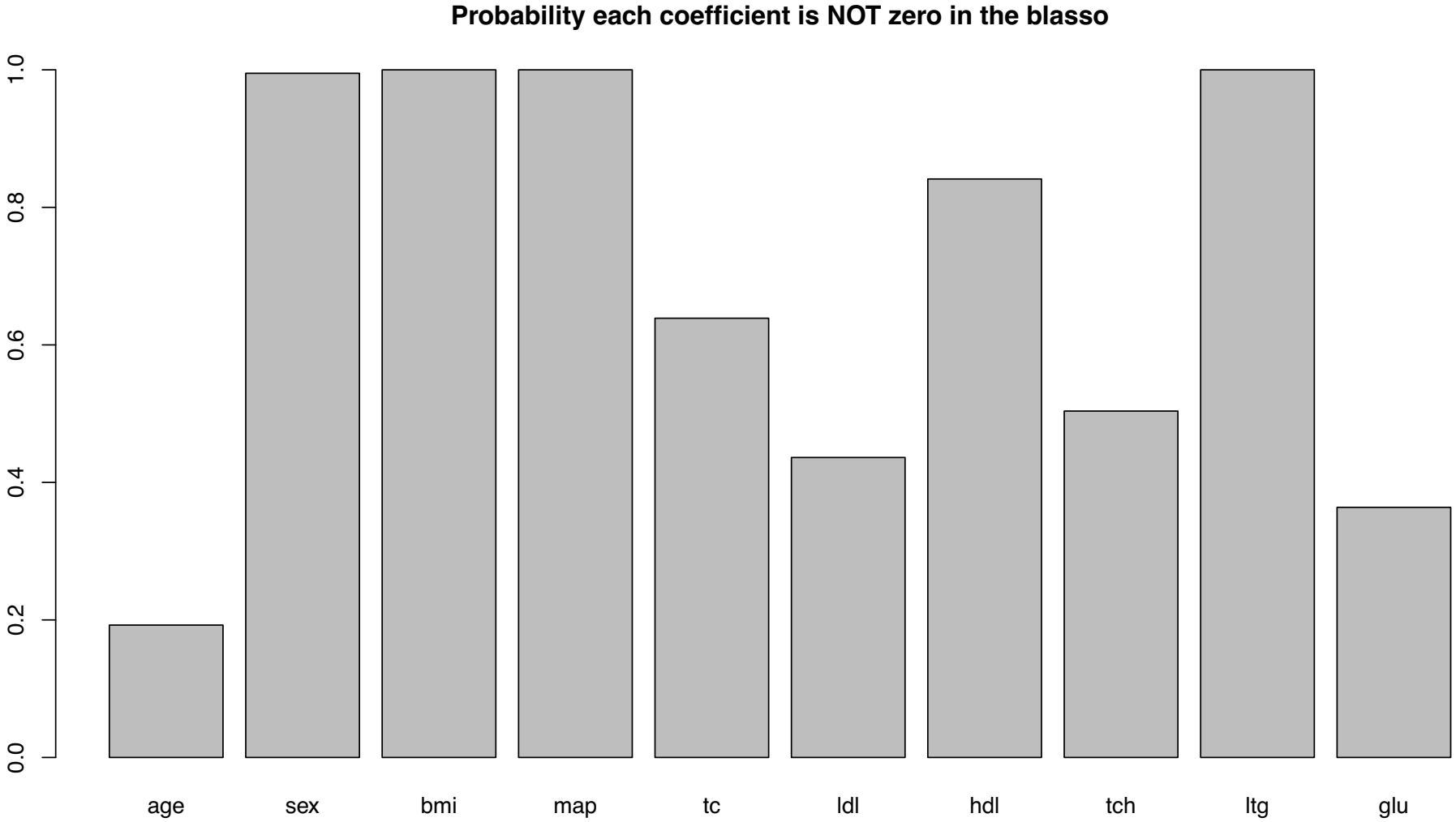


Figure 6.3 *Posterior distributions for the β_j and $\|\beta\|_1$ for the diabetes data. Summary of 10,000 MCMC samples, with the first 1000 "burn-in" samples discarded.*

Figure 6: Figure 6.3 in Hastie, Tibshirani, and Wainwright (2015)

Hastie, Tibshirani, and Wainwright (2015): 10 000 samples from the posterior.


Boxplots of regression coefficients
blasso−map
lasso
LS
coef
500
0
−500
mu
b.1
b.2
b.3
b.4
b.5
b.6
b.7
b.8
b.9
b.10


Probability each coefficient is NOT zero in the blasso
0.0
0.2
0.4
0.6
0.8
1.0
age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu

Bootstrap

(HTW 6.2)

Procedure to find lasso estimate $\hat{\beta}(\hat{\lambda}_{CV})$

(Copied word by word from HTW page 142)

Refer to these 6 steps as $\hat{\beta}(\hat{\lambda}_{CV})$-loop

1. Fit a lasso path to (X, y) over a dense grid of values $\Lambda = \{\lambda_l\}_{l=1}^{L}$.
2. Divide the training samples into 10 groups at random.
3. With the kth group left out, fit a lasso path to the remaining $9/10$ths, using the same grid Λ.
4. For each $\lambda \in \Lambda$ compute the mean-squared prediction error for the left-out group.
5. Average these errors to obtain a prediction error curve over the grid Λ.
6. Find the value $\hat{\beta}(\hat{\lambda}_{CV})$ that minimizes this curve, and then return the coefficient vector from our original fit in step (1) at that value of λ.

Observe:

- λ-path is the same for each run of the lasso
- the chosen λ is then used on the orginal data

Q: Is it possible to use resampling to estimate the distribution of the lasso $\widehat{\beta}$ estimator including the model selection (choosing λ)?

Non-parametric (paired) bootstrap

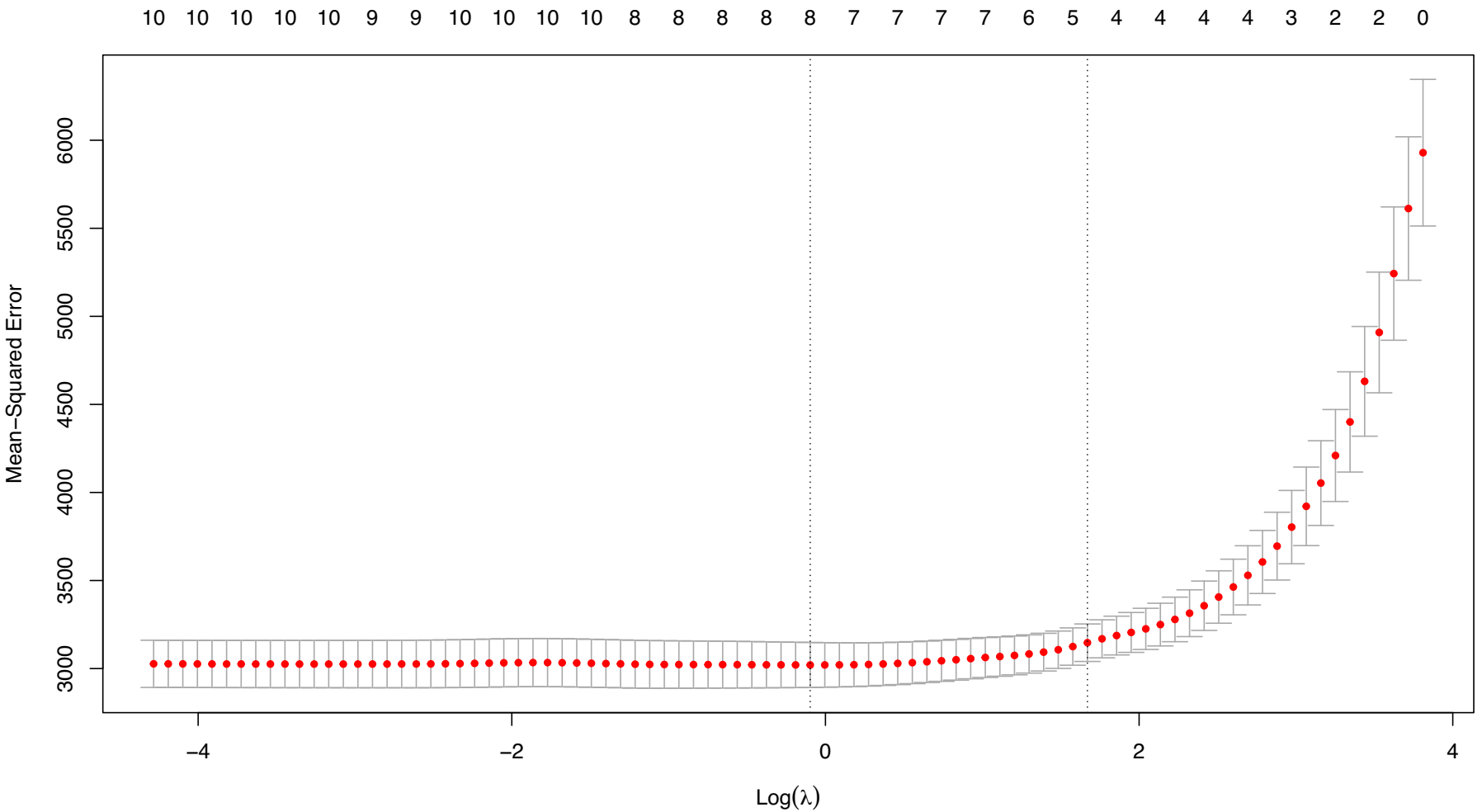
- Let F denote the joint distribution of (X, Y).
- The empirical $\hat{F}$ is $\frac{1}{N}$ for each observation (X, Y) in our training data (X_i, Y_i), $i = 1, \ldots, N$.
- Drawing from $\hat{F}$ is the same as drawing from the N observations in the training data with replacement.

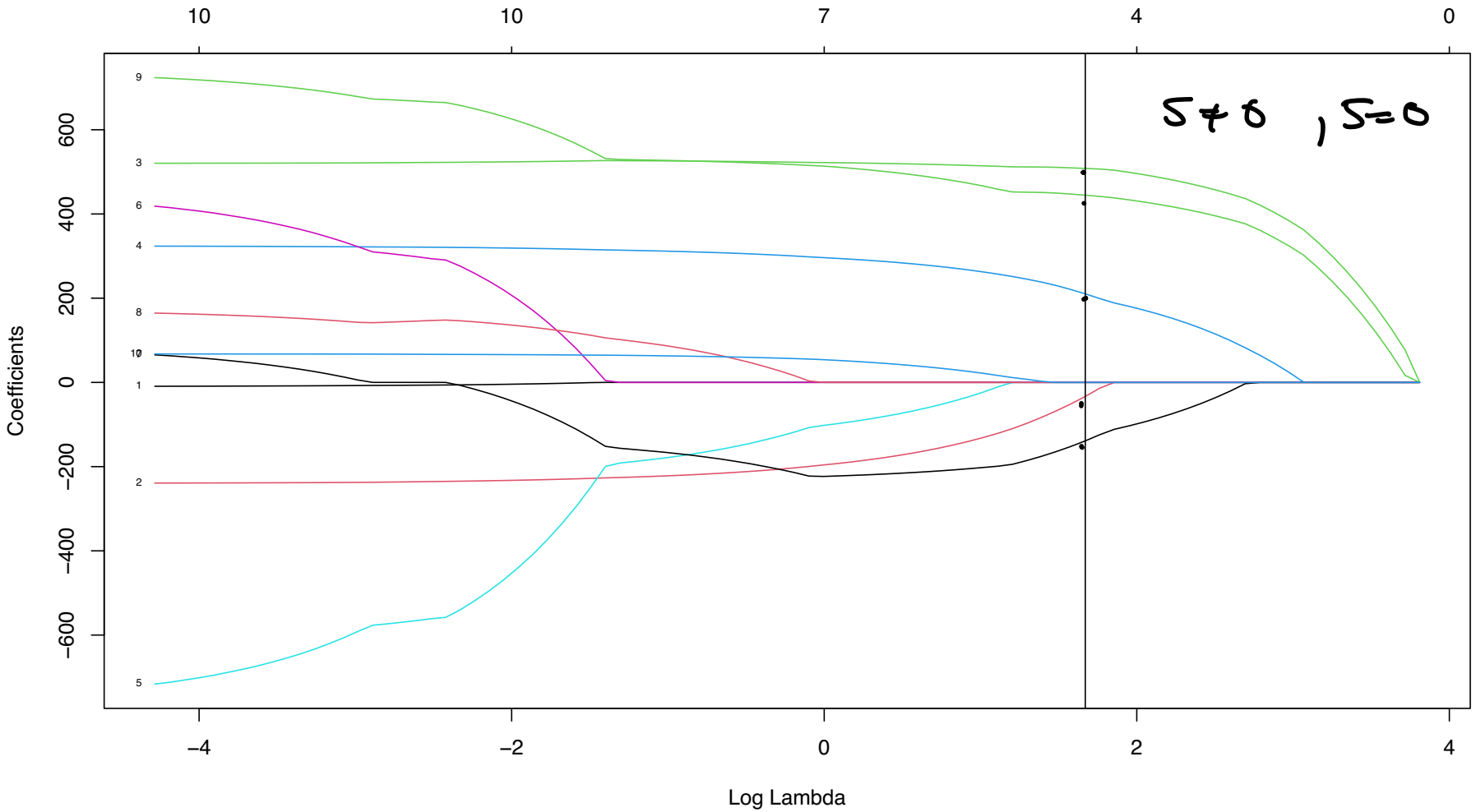
Now, we draw B bootstrap samples from the training data, and for each new bootstrap sample we run through the 6 steps in the $\hat{\beta}(\hat{\lambda}_{CV})$-loop.

- The result is B vectors $\hat{\beta}(\hat{\lambda}_{CV})$.
- We plot the result as
  - boxplots,
  - proportion of times each element of $\hat{\beta}(\hat{\lambda}_{CV})$ is equal 0.

Diabetes example

First : the original data set! — the 6 steps

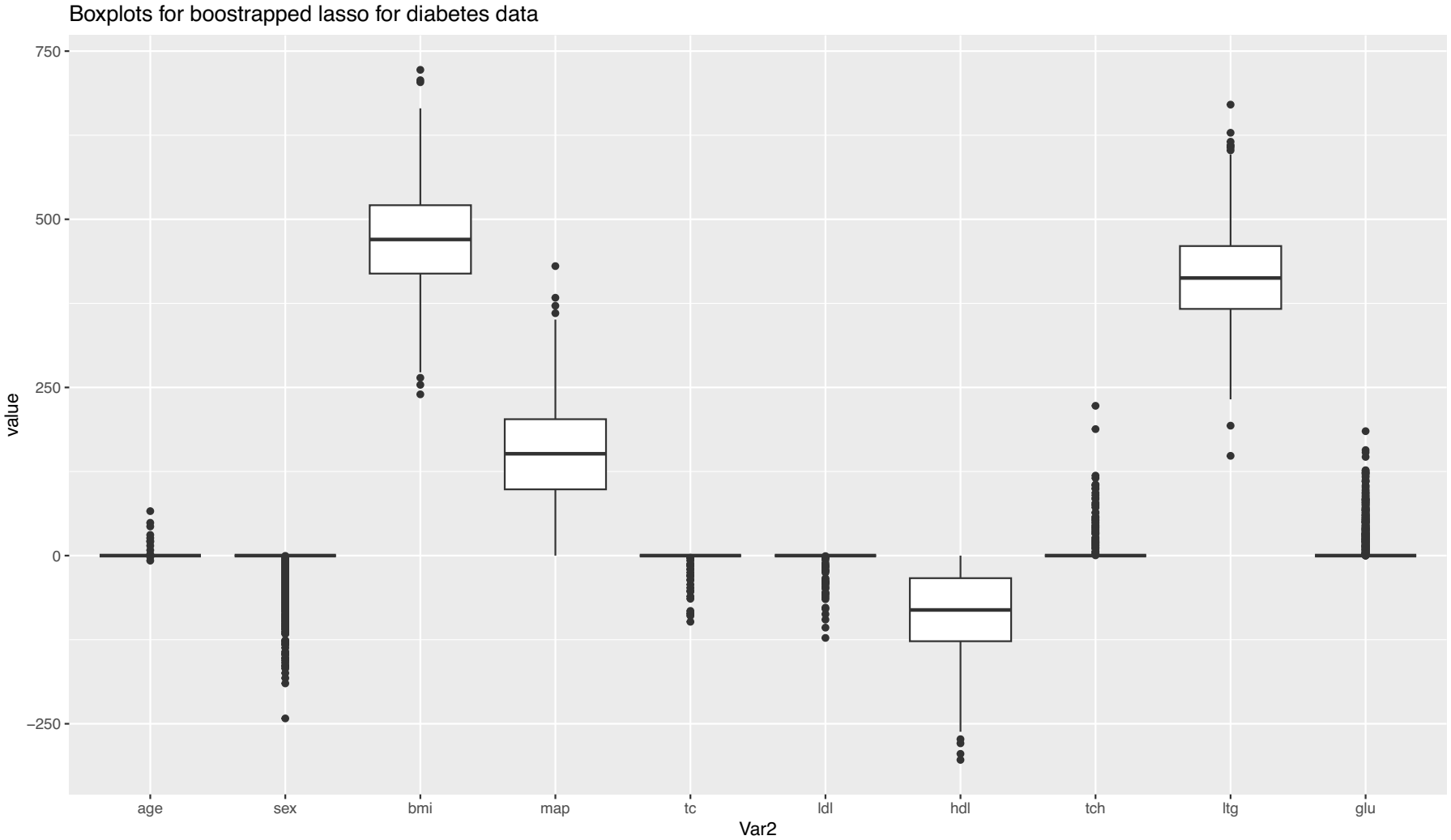

10 10 10 10 10 9 9 10 10 10 10 8 8 8 8 8 7 7 7 7 6 5 4 4 4 4 3 2 2 0
Mean−Squared Error
3000
3500
4000
4500
5000
5500
6000
−4
−2
0
2
4
Log(λ)

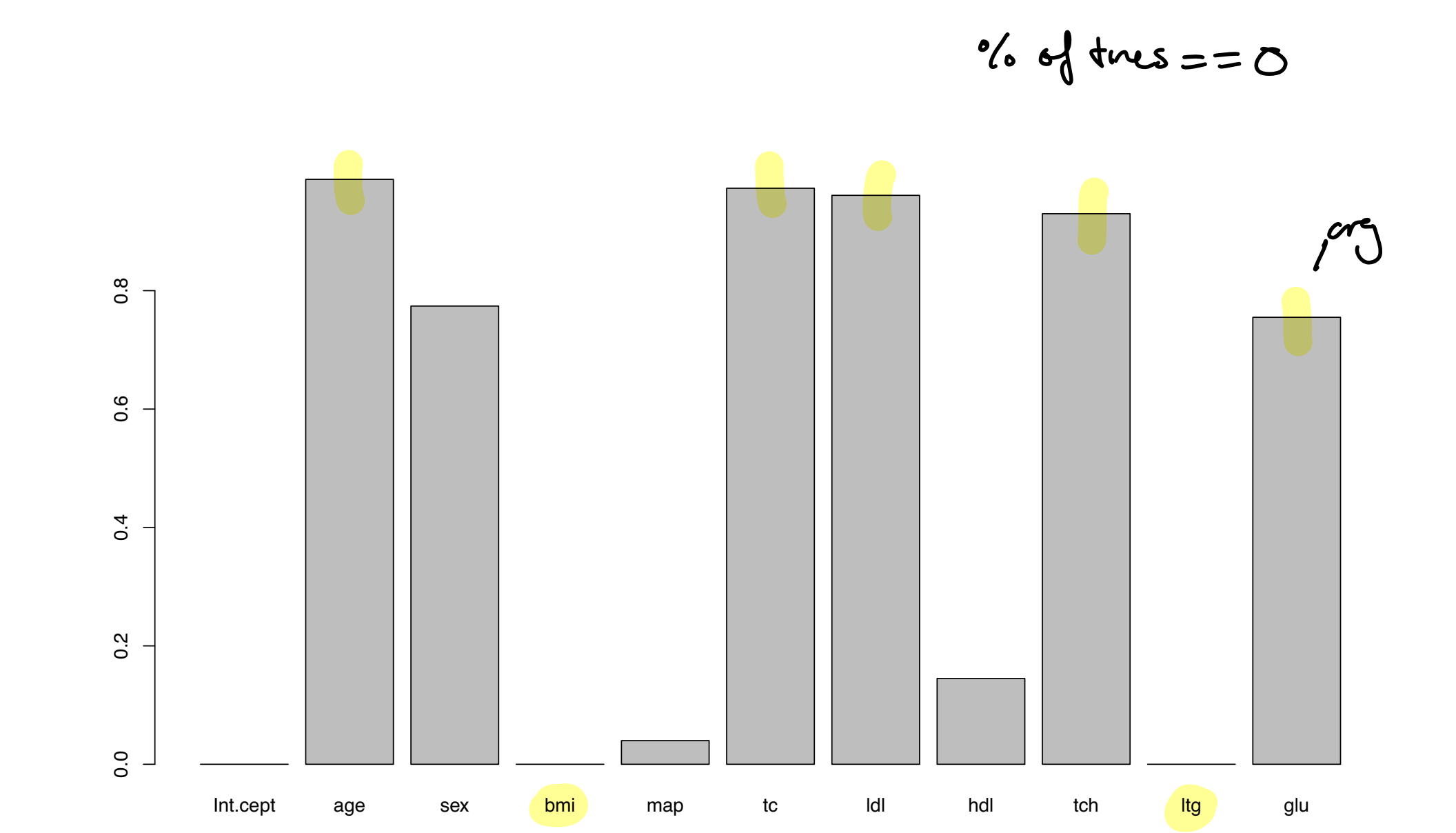

10
10
7
4
0
9
3
6
4
8
10
1
2
5
600
400
200
0
−200
−400
−600
Coefficients
−4
−2
0
2
4
Log Lambda

```
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  152.13348
age                .
sex            -33.35229
bmi            508.13935
map            210.34606
tc                 .
ldl                .
hdl           -138.84433
tch                .
ltg            444.59064
glu                .
```

$S: \hat{\beta}_j = 0$

$S: \hat{\beta}_j \neq 0$


Boxplots for boostrapped lasso for diabetes data
750
500
250
0
−250
value
age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu
Var2


% of trues == 0
org
0.0
0.2
0.4
0.6
0.8
Int.cept
age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu

Bootstrapping vs Bayesian lasso

The results from the Bayesian lasso on the proportion of times a coefficient is 0 and the boxplots are very similar to the results from the bootstrapping. The bootstrap seems to be doing the "same" as a Bayesian analysis with the Laplacian prior.

When the model is not so complex and the number of covariates is not too large ($p \sim 100$) the Bayesian lasso might be as fast as the bootstrapping, but for larger problems the bootstrap "scales better".

For GLMs the Bayesian solution is more demanding, but the bootstrap is as easy as for the linear model.

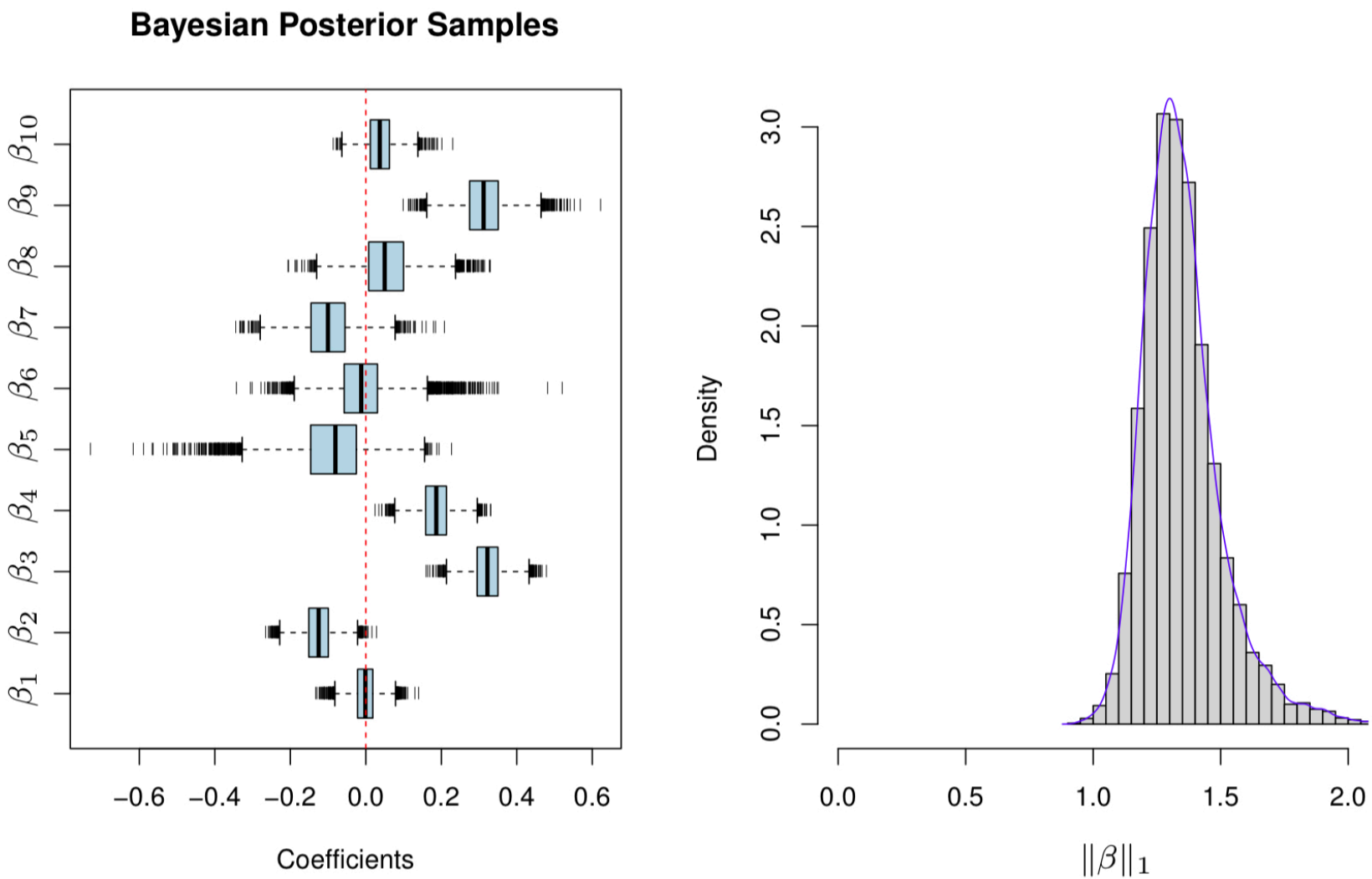


Figure 6.3 *Posterior distributions for the β_j and $\|\beta\|_1$ for the diabetes data. Summary of 10,000 MCMC samples, with the first 1000 "burn-in" samples discarded.*

Figure 7: Figure 6.3 in Hastie, Tibshirani, and Wainwright (2015)

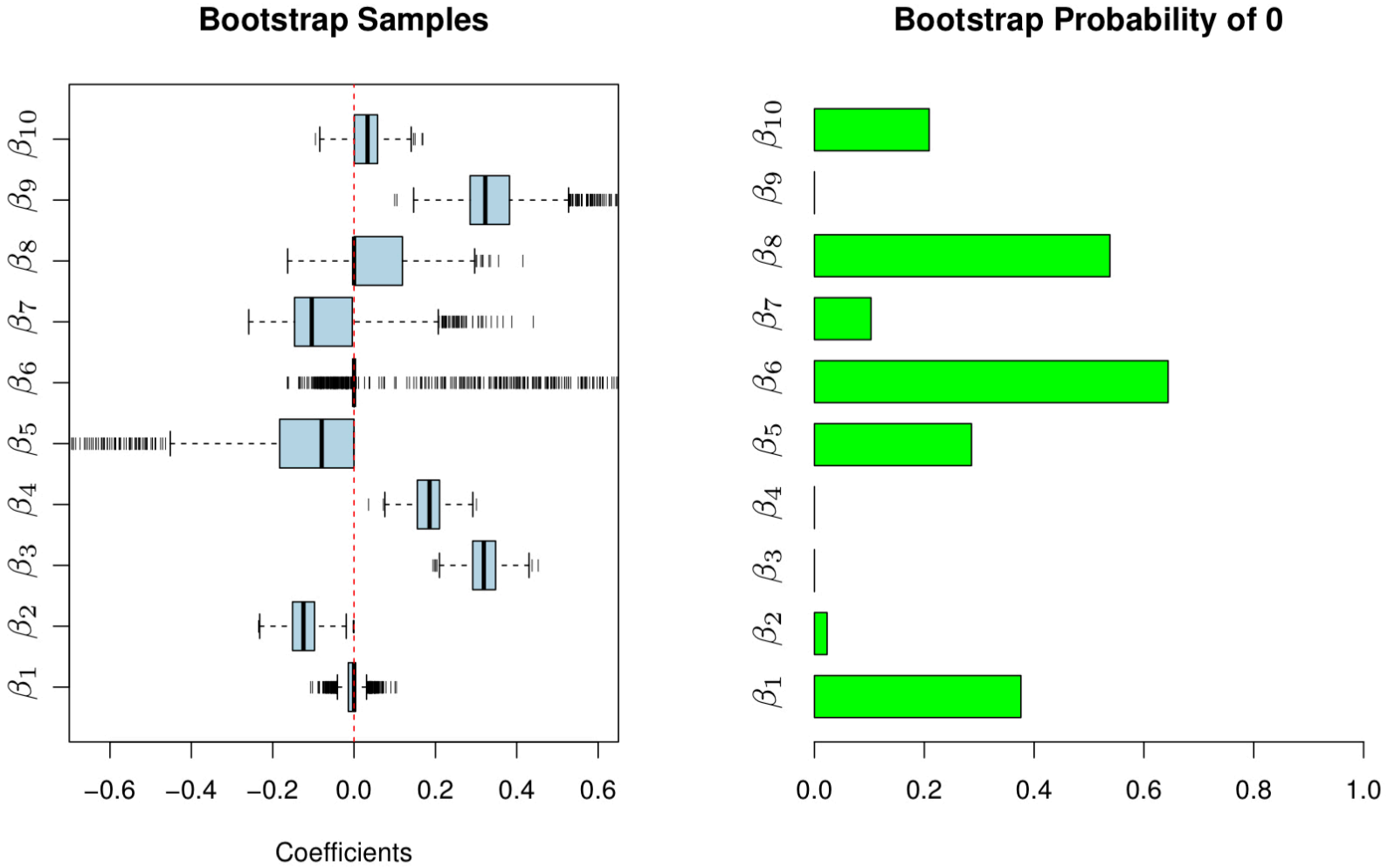


Figure 6.4 *[Left] Boxplots of 1000 bootstrap realizations of $\widehat{\beta}^*(\hat{\lambda}_{CV})$ obtained by the nonparametric bootstrap, which corresponds to re-sampling from the empirical CDF $\widehat{F}_N$. Comparing with the corresponding Bayesian posterior distribution in Figure 6.3, we see a close correspondence in this case. [Right] Proportion of times each coefficient is zero in the bootstrap distribution.*

Figure 8: Figure 6.4 in Hastie, Tibshirani, and Wainwright (2015)

Bootstrap percentile CI

To construct a $(1-\alpha)\cdot 100\%$ CI:

- order the bootstrap sample for the estimate of interest
- read off the $(1-\alpha/2)\cdot 100$ percetile
- $(\alpha/2)\cdot 100$ percentil

These are now the lower and upper limit of the CI.

Bootstrap BCa CI

See page 34 of Bootstrap confidence intervals in the master thesis of Lene Tillerli Omdal Section 3.6.2 and teaching material from TMA4300: Givens and Hoeting (2013) chapter 9.3. NTNU-access to the full book if you are on vpn.

Diabetes example

What if we calculated percentile bootstrap intervals - could we use that to say anything about the true underlying regression coefficients?

Results

Medical example

(See Figures from study in class notes.)

not included only shown and discussed in class

Bootstrap CIs for β_j

Sadly, there are two main challenges:

- The percentile interval is not a good choice for biased estimators, and it is not clear if the bias-corrected accelerated intervals are better
- It has been shown that (for fixed p) the asymptotic ($N \to \infty$) distribution of the lasso has point mass at zero (which leads to that bootstrapping not having optimal properties).

The authors of the *penalized package* take the following view Section 6: A note on standard errors and confidence intervals in the Penalized user manual https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

"Unfortunately, in most applications of penalized regression it is impossible to obtain a sufficiently precise estimate of the bias. Any bootstrap-based calculations can only give an assessment of the variance of the estimates. Reliable estimates of the bias are only available if reliable unbiased estimates are available, which is typically not the case in situations in which penalized estimates are used."

"It is certainly a mistake to make confidence statements that are only based on an assessment of the variance of the estimates, such as bootstrap-based confidence intervals do."

Reliable confidence intervals around the penalized estimates can be obtained in the case of low dimensional models using the standard generalized linear model theory as implemented in lm, glm and coxph."

Outline

- Prediction vs statistics inference: what are the aims?
- Sampling distributions
- Bayesian lasso
- Boostrapping
- Debiased lasso

WE are here now!

- Sample splitting
- Inference after selection (forward regression example, polyhedral result, PoSI)
- Reproducibility crisis and selective inference
- Conclusions