Module 4: Bayesian Methods
Lectures 7: Model selection and averaging

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Model selection and averaging

Diabetes example:

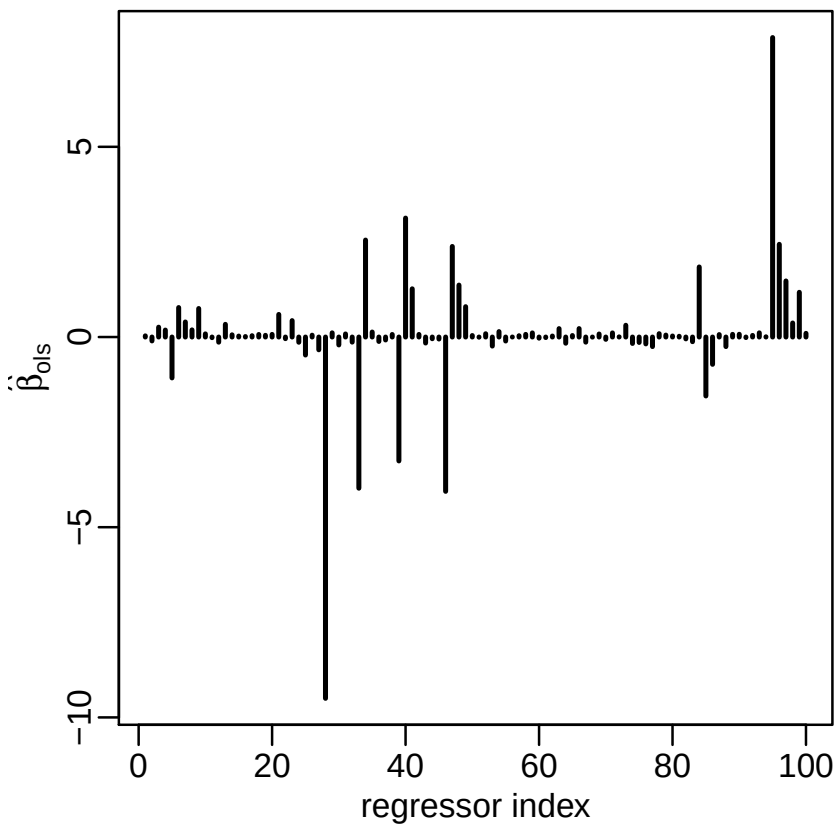
- 342 subjects
- y_i = diabetes progression
- $\mathbf{x}_i$ = explanatory variables.

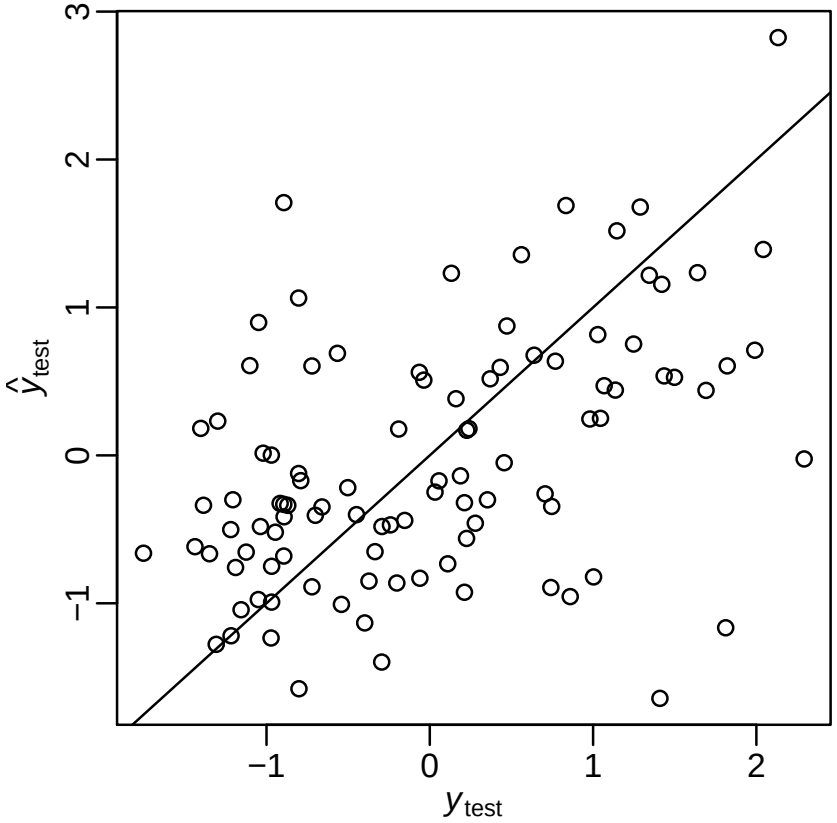
Each $\mathbf{x}_i$ includes

- 13 subject specific measurements $(x_{\text{age}}, x_{\text{sex}}, \ldots)$;
- $78 = \binom{13}{2}$ interaction terms $(x_{\text{age}} \cdot x_{\text{sex}}, \ldots)$;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

OLS regression





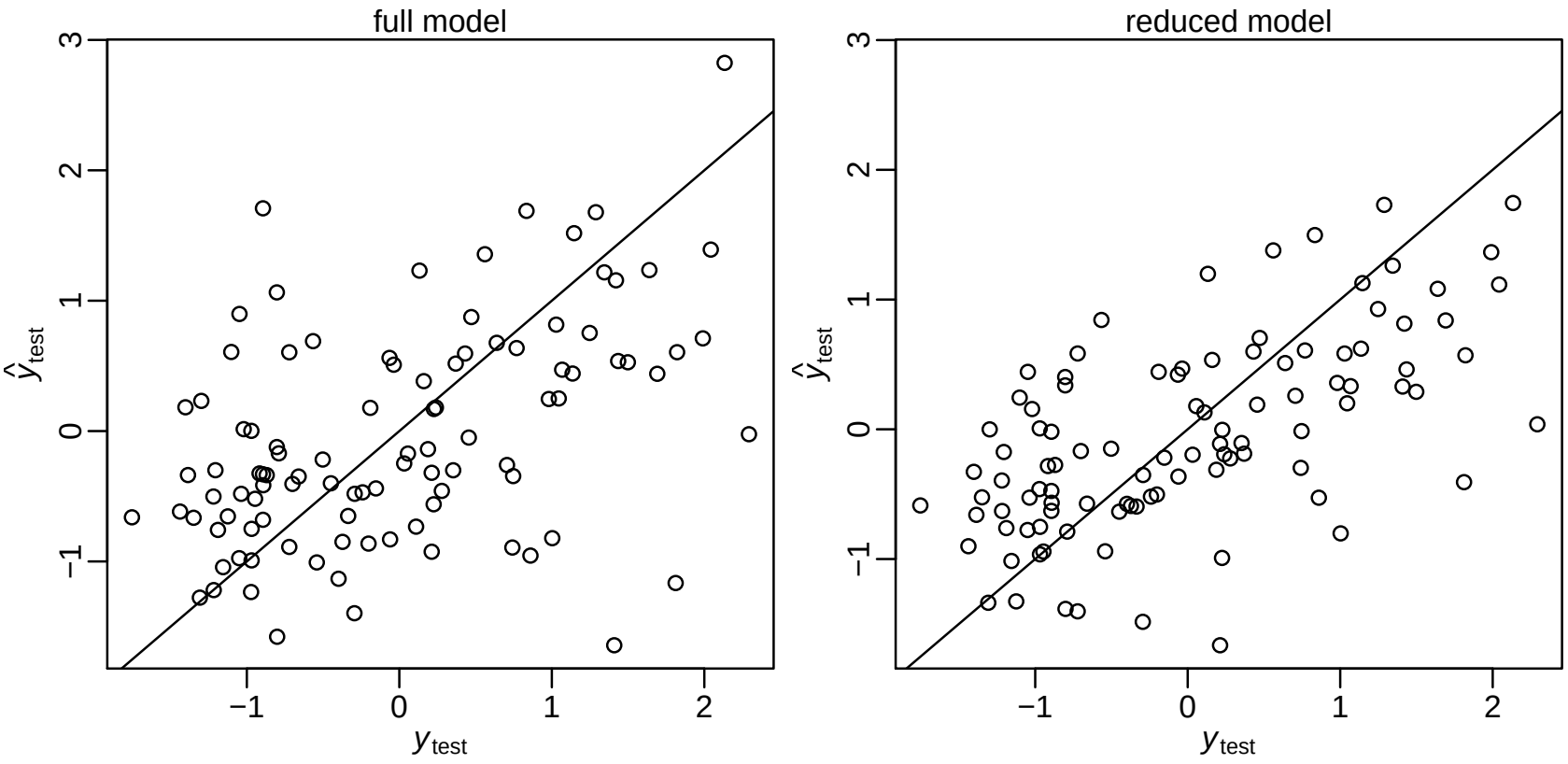
$$\frac{1}{100}\sum(y_{\text{test},i} - \hat{y}_{\text{test},i})^2 = 0.9263$$

$$\frac{1}{100}\sum(y_{\text{test},i} - 0)^2 = \sum y_{\text{test},i}^2 = 1.0095$$

Backwards elimination

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
   2.1 find the regressor j_{min} having the smallest value of $|t_j|$ and remove column j_{min} from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

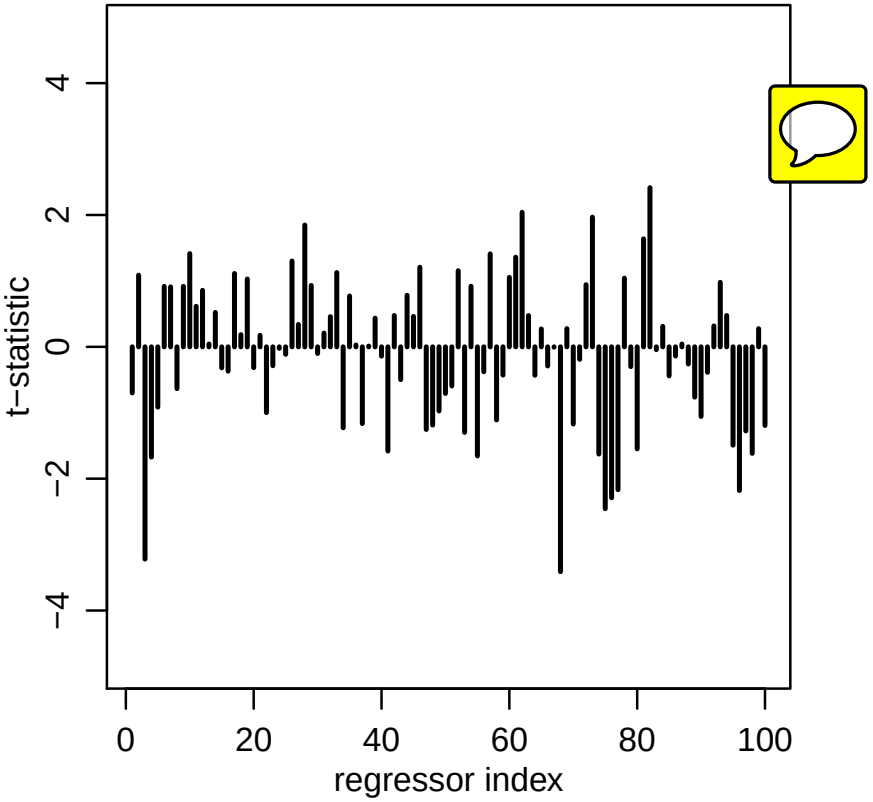
Backwards elimination



$$\frac{1}{100}\sum(y_{\text{test},i} - \hat{y}_{\text{test}^{bel},i})^2 = 0.6392$$

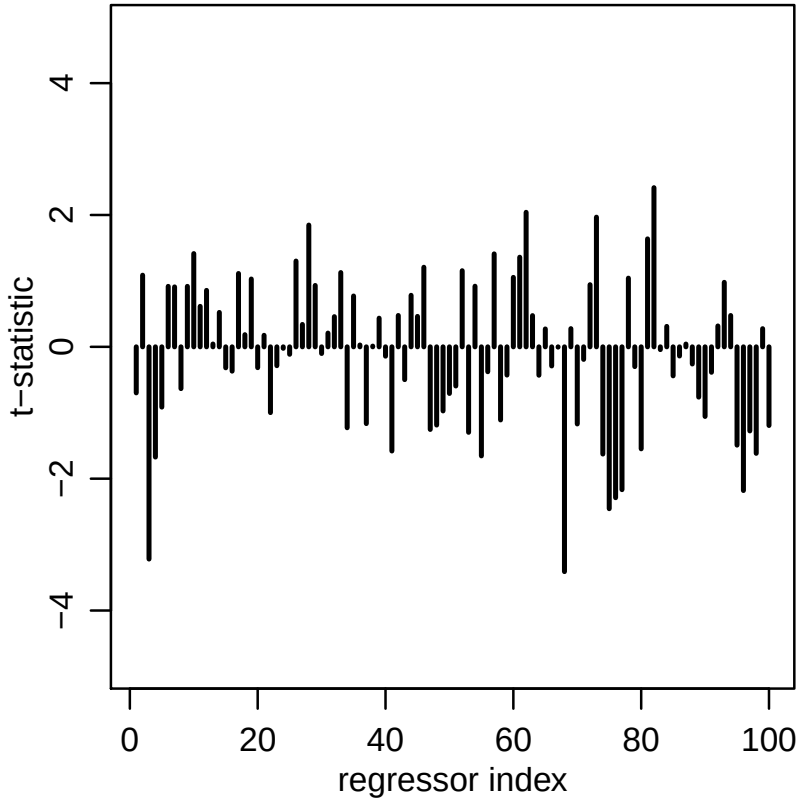
Spurious associations

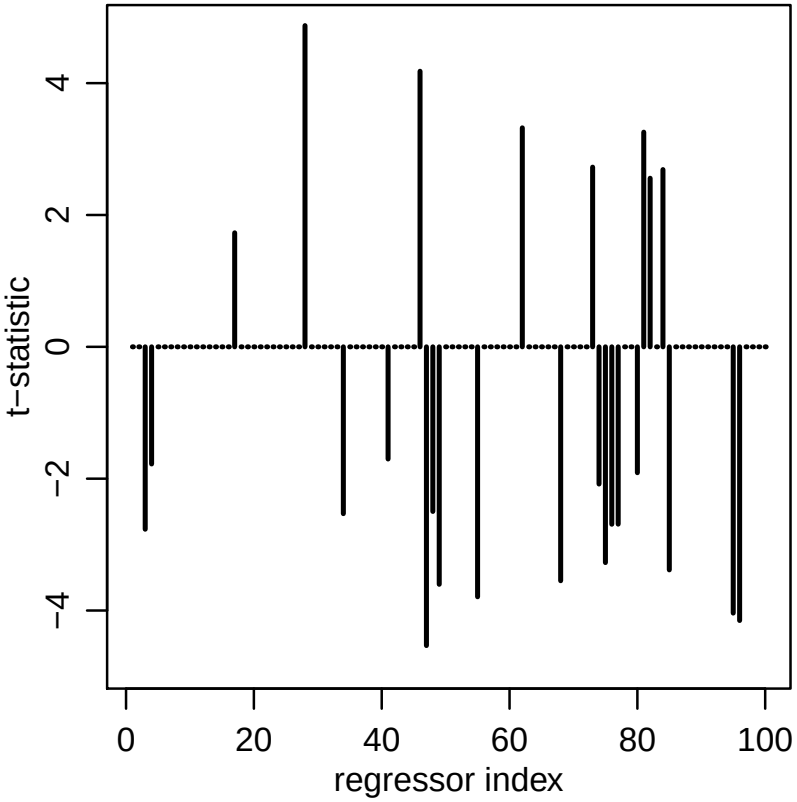
Now try modeling $y_{\pi i} = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$



Spurious associations

Now try modeling $y_{\pi i} = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$





Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Bayesian model selection

Prior belief: $\beta_j \approx 0$ for many j's.

Formulation: Write $\beta_j = z_j \times b_j$, where $z_j \in \{0, 1\}$ and $b_j \in \mathbb{R}$.

$$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i.$$

For example, in the FTO experiment,

$$\begin{aligned}
\mathrm{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,0,1,0)] &= b_1 x_1 + b_3 x_3 \\
&= b_1 + b_3 \times \mathrm{age} \\
\mathrm{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,1,0,0)] &= b_1 x_1 + b_2 x_2 \\
&= b_1 + b_2 \times \mathrm{group} \\
\mathrm{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,1,1,0)] &= b_1 x_1 + b_2 x_2 + b_3 x_3 \\
&= b_1 + b_2 \times \mathrm{group} + b_3 \times \mathrm{age}.
\end{aligned}$$

Each value of $\mathbf{z} = (z_1, \ldots, z_p)$ corresponds to a *different model*.

Bayesian model comparison

Posterior probability

$$p(\mathbf{z}|\mathbf{y},\mathbf{X}) = \frac{p(\mathbf{z})p(\mathbf{y}|\mathbf{X},\mathbf{z})}{p(\mathbf{y}|\mathbf{X})}$$

Model comparison

$$\frac{p(\mathbf{z}_a|\mathbf{y},\mathbf{X})}{p(\mathbf{z}_b|\mathbf{y},\mathbf{X})} = \frac{p(\mathbf{z}_a)}{p(\mathbf{z}_b)} \times \frac{p(\mathbf{y}|\mathbf{X},\mathbf{z}_a)}{p(\mathbf{y}|\mathbf{X},\mathbf{z}_b)}$$

$$\text{posterior odds} = \text{prior odds} \times \text{“Bayes factor”}$$

Parsimony

The formula for $p(\mathbf{y}|\mathbf{X}, \mathbf{z})$ is messy, but

$$\frac{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_a)}{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_b)} = (1+n)^{(p_{z_b} - p_{z_a})/2} \left(\frac{s^2_{z_a}}{s^2_{z_b}} \right)^{1/2} \times \left(\frac{s^2_{z_b} + \mathrm{SSR}^{z_b}_g}{s^2_{z_a} + \mathrm{SSR}^{z_a}_g} \right)^{(n+1)/2}.$$

A model $\mathbf{z}_a$ is penalized if

- it is too complex (p_A is large)
- it doesn't fit well (SSR^a_g is large)

FTO example

$$
\begin{aligned}
\mathrm{E}[Y_i|\boldsymbol{\beta}, \mathbf{x}_i] &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} \\
&= \beta_1 + \beta_2 \times \mathrm{grp}_i + \beta_3 \times \mathrm{age}_i + \beta_4 \times \mathrm{grp}_i \times \mathrm{age}_i \,.
\end{aligned}
$$

effect of group $\Leftrightarrow$ one of more of β_2, β_4 not zero

z	model	$\log p(\mathbf{y}\|\mathbf{X}, \mathbf{z})$	$p(\mathbf{z}\|\mathbf{y}, \mathbf{X})$
(1,0,0,0)	β_1	-71.82	0
(1,1,0,0)	$\beta_1 + \beta_2 \times \mathrm{grp}_i$	-70.04	0
(1,0,1,0)	$\beta_1 + \beta_3 \times \mathrm{age}_i$	-67.04	0
(1,1,1,0)	$\beta_1 + \beta_2 \times \mathrm{grp}_i + \beta_3 \times \mathrm{age}_i$	-67.04	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times \mathrm{grp}_i + \beta_3 \times \mathrm{age}_i + \beta_4 \times \mathrm{grp}_i \times \mathrm{age}_i$	-61.72	0.37

$$
\begin{aligned}
\Pr(\beta_2 \text{ or } \beta_4 \neq 0) &= 0.60 \\
\Pr(\beta_2 \text{ or } \beta_4 \neq 0|\mathbf{y}, \mathbf{X}) &\approx 1
\end{aligned}
$$

High dimensional regression

Diabetes example: $p = 100 \Rightarrow 2^{100} \approx 10^{30}$ models to consider.

We can't compute $p(\mathbf{z}|\mathbf{y}, \mathbf{X})$ for each $\mathbf{z}$. Instead, we hope to

- search for models $\mathbf{z}$ with high posterior probability;
- approximate $\beta_j = z_j \times b_j$ for each j;
- build a predictive model for $\mathbf{y}$.

This can be achieved via a Monte Carlo method known as *Gibbs sampling*.

The Gibbs sampler

Goal: A Monte Carlo approximation to $p(x, y, z)$

Given $\{x^{(s)}, y^{(s)}, z^{(s)}\}$,

1. simulate $x^{(s+1)} \sim p(x|y^{(s)}, z^{(s)})$,
2. simulate $y^{(s+1)} \sim p(y|x^{(s+1)}, z^{(s)})$,
3. simulate $z^{(s+1)} \sim p(z|x^{s+1)}, y^{(s+1)})$.

This generates $\{x^{(s+1)}, y^{(s+1)}, z^{(s+1)}\}$.

The Gibbs sampler

Repeated many times, this generates $\{x^{(1)}, y^{(1)}, z^{(1)}\}, \ldots, \{x^{(S)}, y^{(S)}, z^{(S)}\}$

The distribution of this sequence *approximates* $p(x, y, z)$:

$$\begin{aligned}
\frac{1}{S}\sum x^{(s)} &\approx \mathrm{E}[x] = \int x \; p(x,y,z) \; dx \; dy \; dz \\
\frac{\#(x^{(s)} \in A)}{S} &\approx \Pr(x \in A) = \int\int\int_A p(x,y,z) \; dx \; dy \; dz \\
\frac{\#(\{x^{(s)}, y^{(s)}, z^{(s)}\} \in B)}{S} &\approx \int\int\int_B p(x,y,z) \; dx \; dy \; dz
\end{aligned}$$

By necessity, the sequence will frequently visit regions where $p(x, y, z)$ is large.

Gibbs sampling for model selection

Goal Approximate $p(z_1, \ldots, z_p | \mathbf{y}, \mathbf{X})$.

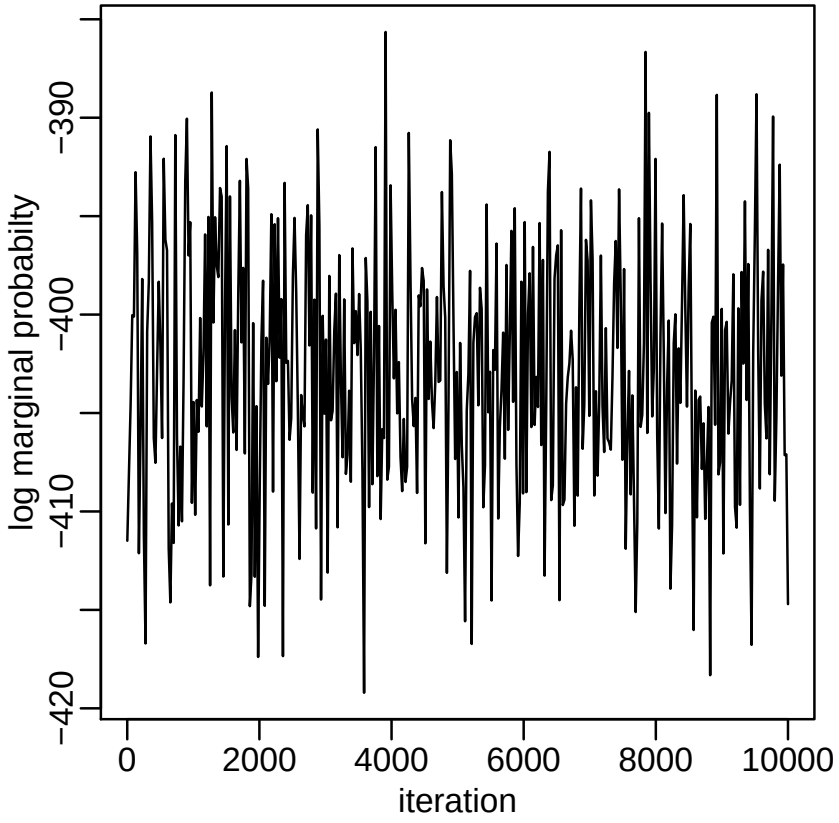
Gibbs sampler: Given $\mathbf{z}^{(s)} = (z_1^{(s)}, \ldots, z_p^{(s)})$,

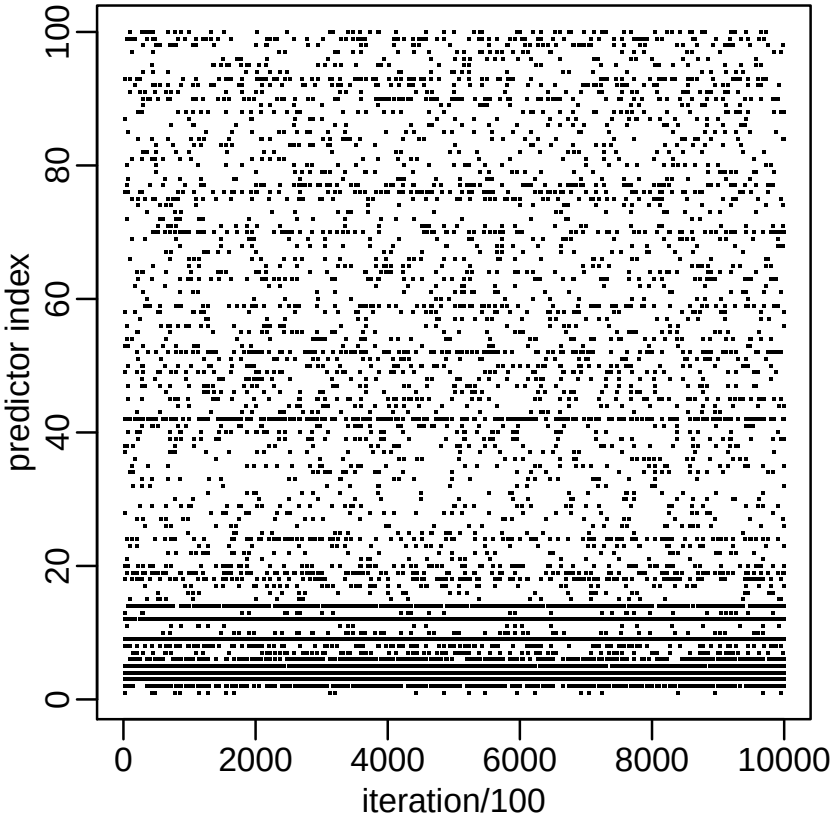
$$\begin{aligned}
z_1^{(s+1)} &\sim p(z_1 | z_2^{(s)}, \ldots, z_p^{(s)}, \mathbf{y}, \mathbf{X}) \\
z_2^{(s+1)} &\sim p(z_2 | z_1^{(s+1)}, z_3^{(s)}, \ldots, z_p^{(s)}, \mathbf{y}, \mathbf{X}) \\
&\vdots \\
z_p^{(s+1)} &\sim p(z_p | z_1^{(s+1)}, \ldots, z_{p-1}^{(s+1)}, \mathbf{y}, \mathbf{X})
\end{aligned}$$

This generates $\mathbf{z}^{(s+1)}$ from $\mathbf{z}^{(s)}$.

Repeating this generates $z^{(1)}, \ldots, z^{(S)}$ with which to approximate $p(\mathbf{z} | \mathbf{y}, \mathbf{X})$.

Diabetes example





Marginal inference

What is the estimate of $\boldsymbol{\beta}$?

Recall

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) = (b_1 z_1, \ldots, b_p, z_p)$$

Our Monte Carlo samples are

$$\begin{array}{rcccccc} \boldsymbol{\beta}^{(1)} & = & (0 & -.299 & 0 & .427 & \cdots & .845) \\ \boldsymbol{\beta}^{(2)} & = & (0 & -.235 & .834 & .374 & \cdots & 0) \\ \vdots & & & & \vdots & & & \\ \boldsymbol{\beta}^{(S)} & = & (0 & -.315 & 0 & .536 & \cdots & 0) \end{array}$$

A posterior mean for $\boldsymbol{\beta}$ is obtained in the usual way:

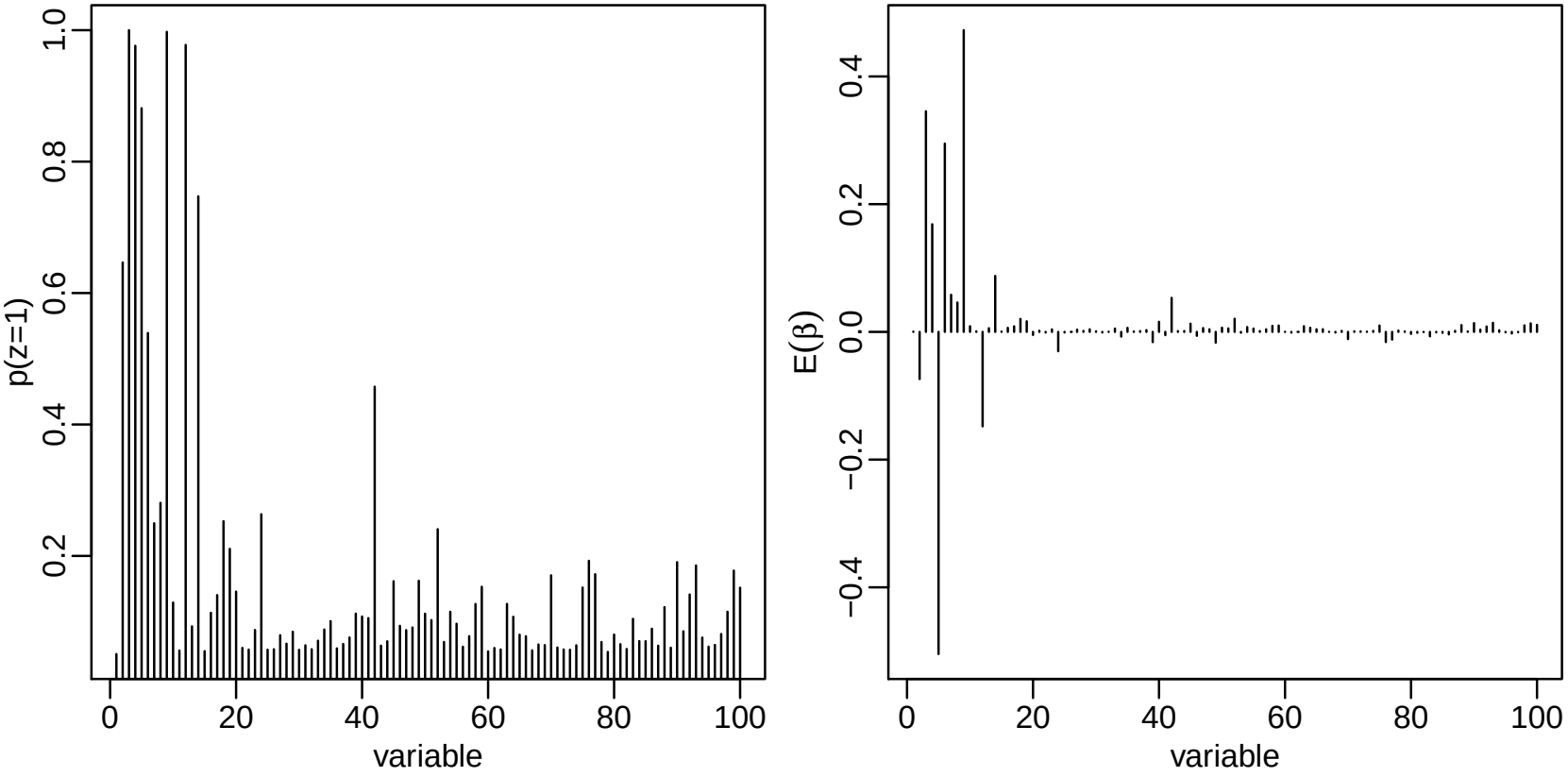
$$\hat{\boldsymbol{\beta}}^{\text{bayes}} = \frac{1}{S} \sum \boldsymbol{\beta}^{(s)} \approx \mathrm{E}[\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}]$$

Out of sample predictions can be made with $\hat{\boldsymbol{\beta}}_{\text{bayes}}$:

$$\hat{y}^{bayes}_{\text{test,i}} = \hat{\boldsymbol{\beta}}^T_{\text{bayes}} \mathbf{x}_{\text{test,i}}$$

Out of sample prediction error: $\frac{1}{S} \sum (y_{\text{test,i}} - \hat{y}^{bayes}_{\text{test,i}})^2 = 0.4853$

Marginal inference



Important variables

```
colnames(X)[ order(z.pmean,decreasing=TRUE)[1:10] ]

##  [1] "bmi"     "ltg"     "g2"      "map"     "tc"      "sex.age" "sex"
##  [8] "ldl"     "ltg.age" "tch"

colnames(X)[ order(b.pmean,decreasing=TRUE)[1:10] ]

##  [1] "ltg"     "bmi"     "ldl"     "map"     "sex.age" "hdl"     "ltg.age"
##  [8] "tch"     "glu.bmi" "map.sex"
```

Checking the null