STAT3016/4116/7016
Introduction to Bayesian Data Analysis

RSFAS,College of Business and Economics, ANU

Linear Regression

Introduction

OLS

- (1) Relationship between Bayesian and ordinary least squares regression
- (2) Bayesian approach to model selection

① OLS

$$\sum_{i=1}^{n} (y_i - \beta^T x_i)^2$$

Bayesian approach:

β is not fixed, it is a r.v. which varies.

$\downarrow$

posterior of β

$\downarrow$

inference.

② In which situation we do model selection?

when # of variables is large.

(to avoid multicolinearity).

methods (frequentists: BE, FS, stepwise, AIC, BIC Criterions) ...

More advanced method like Lasso.

drawback: slow when # large.

- can add a penalty term: (Lasso)

$$\sum_{i=1}^{n} (y_i - \beta^T x_i) + \lambda \sum_{j=1}^{p} |\beta_j|$$ freq.

Baysian approach to model selection (2nd part this week)

The linear regression model

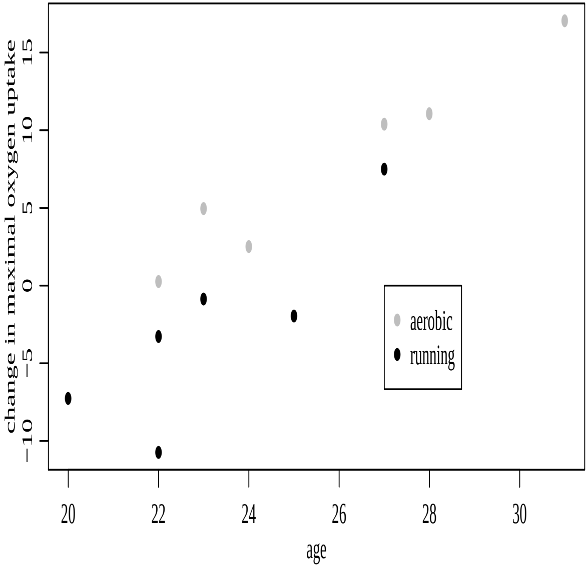
response

- Concerned with how the sampling distribution of one random variable Y varies with a set of variables $\mathbf{x} = \{x_1, ..., x_p\}$ covariates
- Assume a form for $p(y|\mathbf{x})$. Estimate $p(y|\mathbf{x})$ using data $y_1, ..., y_p$ gathered under a variety of conditions of $\mathbf{x}_1$,....,$\mathbf{x}_p$

$$y_i = \beta^T x_i + \varepsilon_i \quad , \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\varepsilon_1, \dots, \varepsilon_n$$

The linear regression model - Oxygen uptake example



The linear regression model - Oxygen uptake example

We would like to estimate the conditional distribution of oxygen uptake for a given exercise program and age.
The linear regression model has the following form:

$$\int yp(y|x)dy = E[y|x] = \beta_1 x_1 + ... + \beta_p x_p = \beta^T \mathbf{x}$$

For the oxygen example, a reasonable model for $p(y|\mathbf{x})$ could be:

$$Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i$$

$x_{i,2} = 0$ if subject i is on the running program, 1 if on aerobic
$x_{i,3} =$ age of subject i
$x_{i,4} = x_{i,2} \times x_{i,3}$
What if we assume $\beta_2 = \beta_4 = 0$? What if we assume $\beta_4 = 0$?

The above model tells us about $E[Y|\mathbf{x}]$. What about the sampling variability??

The linear regression model - Oxygen uptake example

$$\epsilon_1, ..., \epsilon_n \overset{\text{iid}}{\sim} \text{normal}(0, \sigma^2)$$

$$Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$$

$Y_i \sim N(\beta^T x_i, \sigma^2)$

$$p(y_1, ..., y_n | \mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)$$

$$= (2\pi\sigma^2)^{-n/2} exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \right\}$$

$$\text{or } \mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I})$$

maximize likelihood $\Longleftrightarrow$ minimize square loss

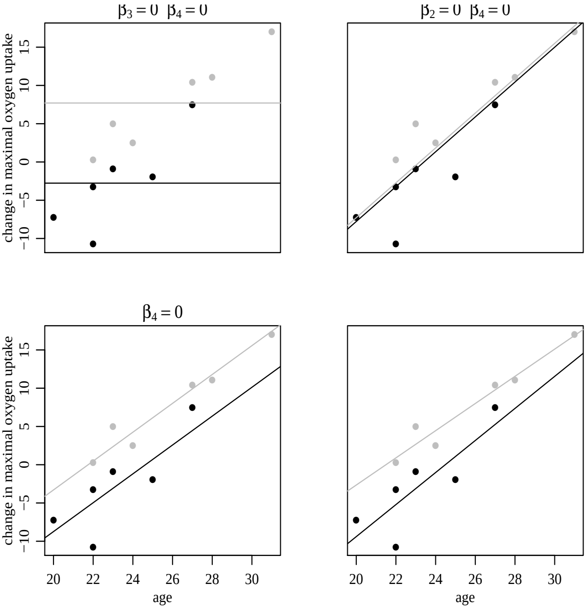
Minimize $SSR(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2$

$\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{y}$; $Var(\hat{\boldsymbol{\beta}}_{ols}) = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\sigma^2$

Bayesian linear regression model - Oxygen uptake example

• Why 2 parallel lines ? Because ...

• We keep age, type, but NO interactions!



Bayesian linear regression model - Prior distributions

(probably most frequently used prior for β)

Semi-conjugate prior distribution for β

Let $\beta \sim MVN(\beta_0, \Sigma_0)$

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)\times p(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}(-2\boldsymbol{\beta}^T\mathbf{X}^\mathbf{T}\mathbf{y}/\sigma^2+\boldsymbol{\beta}^T\mathbf{X}^\mathbf{T}\mathbf{X}\boldsymbol{\beta}/\sigma^2)-\frac{1}{2}(-2\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta_0}+\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta})\right\} \\ &= \exp\left\{\boldsymbol{\beta}^T(\Sigma_0^{-1}\boldsymbol{\beta_0}+\mathbf{X}^\mathbf{T}\mathbf{y}/\sigma^2)-\frac{1}{2}\boldsymbol{\beta}^T(\Sigma_0^{-1}+\mathbf{X}^\mathbf{T}\mathbf{X}/\sigma^2)\boldsymbol{\beta}\right\} \end{aligned}$$

What probability density function is $p(\beta|\mathbf{y},\mathbf{X},\sigma^2)$ proportional to?

$E(\beta|y,X,\sigma^2) = (\Sigma_0^{-1}+X^TX/\sigma^2)^{-1}(\Sigma_0^{-1}\beta_0+X^Ty/\sigma^2)$

$Var(\beta|y,X,\sigma^2) = (\Sigma_0^{-1}+X^TX/\sigma^2)^{-1}$

Var term × another term

Bayesian linear regression model - Prior distributions

Semi-conjugate prior distribution σ^2

$\sigma^2 \sim InvGamma(\nu_0/2, \nu_0\sigma_0^2/2)$

$$
\begin{aligned}
p(\sigma^2|\mathbf{y}, \mathbf{X}\boldsymbol{\beta}) &\propto p(\sigma^2)p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \\
&\propto (\sigma^2)^{-(\nu_0/2+1)} \exp^{-(\nu_0\sigma_0^2/2)/\sigma^2} \times (\sigma^2)^{-n/2} \exp\left\{SSR(\boldsymbol{\beta})/2\sigma^2\right\} \\
&= (\sigma^2)^{-((\nu_0+n)/2+1} \exp\left\{-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + SSR(\boldsymbol{\beta}))\right\}
\end{aligned}
$$

What probability density function is $p(\sigma^2|\mathbf{y}, \mathbf{X}\boldsymbol{\beta})$ proportional to?

Inv Gamma

Bayesian linear regression model - Prior distributions

Unit information prior (a weakly informative prior distribution): contains the same amount of information as that would be contained in only a single observation.

$$\Sigma_0^{-1} = \mathbf{X}^\mathsf{T}\mathbf{X}/(n\sigma^2)$$

$$\boldsymbol{\beta}_\mathbf{0} = \hat{\boldsymbol{\beta}}_{ols}$$

$$\nu_0 = 1$$

$$\sigma_0^2 = \hat{\sigma}^2_{ols}$$

Bayesian linear regression model - Prior distributions

A standard noninformative prior

$$p(\beta, \sigma^2 | X) \propto \sigma^{-2}$$

Posterior

$$\beta | \sigma^2, y, X \sim MVN(\hat{\beta}, V_\beta \sigma^2)$$

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

$$V_\beta = (X^T X)^{-1}$$

and

$$\sigma^2 | y \sim InvGamma(\frac{n-k}{2}, \frac{(n-k)s^2}{2})$$

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T (y - X\hat{\beta})$$

When is the posterior distribution proper?

Bayesian linear regression model - Prior distributions

g-prior

Suppose we require parameter estimation to be invariant to changes in the scale of the covariates.

If not invariant, the prediction will be changed.

Let $\tilde{\mathbf{X}} = \mathbf{XH}$. If we obtain the posterior distribution of $\boldsymbol{\beta}$ from $\mathbf{y}$ and $\mathbf{X}$ and the posterior distribution of $\tilde{\boldsymbol{\beta}}$ from $\mathbf{y}$ and $\tilde{\mathbf{X}}$, then the invariance principle says that the posterior distribution of $\tilde{\boldsymbol{\beta}}$ should be the same as that of $H\boldsymbol{\beta}$.

This is achieved if:

$$\boldsymbol{\beta}_{\mathbf{0}} = \mathbf{0}$$

$$\Sigma_0 = k(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$$

Bayesian linear regression model - Prior distributions

*Note: Monte Carlo simulation. Gibbs not needed. Not full conditional.
(while semiconjugate need Gibbs).

g-prior

Usually we set $k = g\sigma^2$ Under this prior specification, we can show that

under g-prior, posterior $\beta \mid y, X, \sigma^2 \sim MVN(,)$

$$Var[\beta|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X}^\mathbf{T}\mathbf{X}/(g\sigma^2) + \mathbf{X}^\mathbf{T}\mathbf{X}/\sigma^2]^{-1}$$

$$E[\beta|\mathbf{y}, \mathbf{X}, \sigma^2] = \frac{g}{g+1}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}$$

also assumme $\sigma^2 \sim InvGamma(\nu_0/2, \nu_0\sigma_0^2/2)$ and we can show

$$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{InvGamma}((\nu_0 + n)/2, [\nu_0\sigma_0^2 + SSR_g]/2)$$

where $SSR_g = \mathbf{y}^T(\mathbf{I} - \frac{g}{g+1}\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T})\mathbf{y}$

Bayesian linear regression model - Prior distributions

nrows ncols.

```
n<-dim(X)[1] ; p<-dim(X)[2]
Hg<- (g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X)
```

H_g

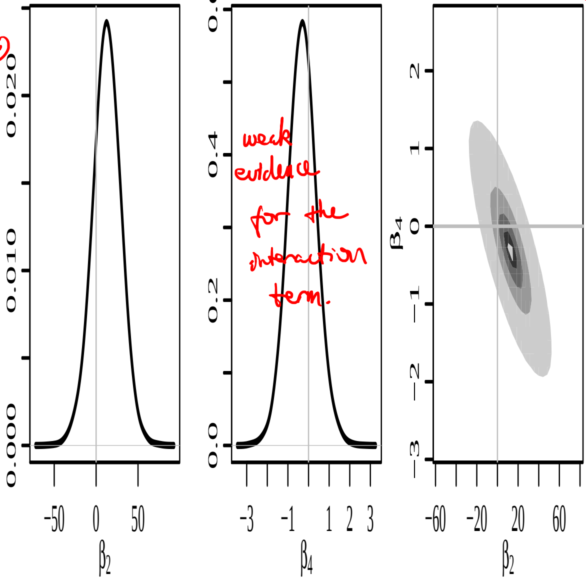
```
SSRg<- t(y)%*%( diag(1,nrow=n)  - Hg ) %*%y

s2<-1/rgamma(S, (nu0+n)/2, (nu0*s20+SSRg)/2 )

Vb<- g*solve(t(X)%*%X)/(g+1)
Eb<- Vb%*%t(X)%*%y

E<-matrix(rnorm(S*p,0,sqrt(s2)),S,p)
beta<-t(  t(E%*%chol(Vb)) +c(Eb))
```

Bayesian linear regression model - Oxygen uptake example

β_2 ~~has~~ β

(breathing program)

β_4 (interaction) has no significant impact on oxygen uptake.



weak evidence for the interaction term.

β_2 & β_4 are highly correlated.

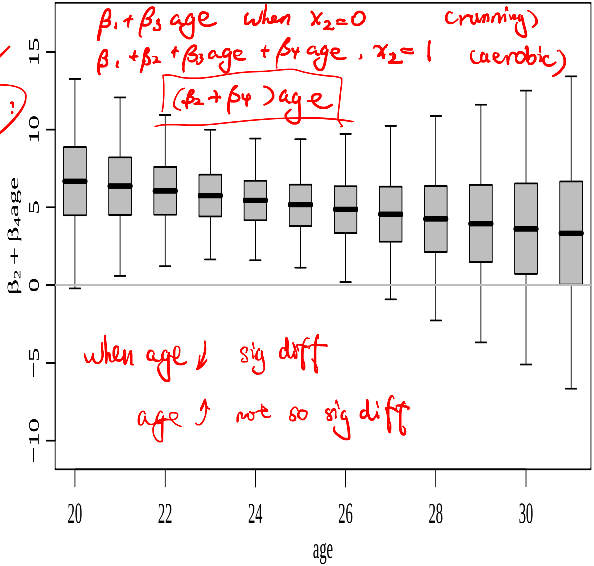
The linear regression model - Oxygen uptake example

What we should really interested in:

$\beta_1 + \beta_3\,\text{age}$ when $x_2 = 0$ (running)

$\beta_1 + \beta_2 + \beta_3\,\text{age} + \beta_4\,\text{age}$, $x_2 = 1$ (aerobic)

diff: $\boxed{(\beta_2 + \beta_4)\,\text{age}}$



when age ↓ sig diff

age ↑ not so sig diff

Bayesian linear regression model - Model checking and sensitivity analysis

frequentist
replaced by ...

- outliers
- normality assumption
- posterior predictive checks

Model selection

- What is the purpose of model selection?
- What are some standard model selection procedures?
- What are some problems with the standard model selection procedures?

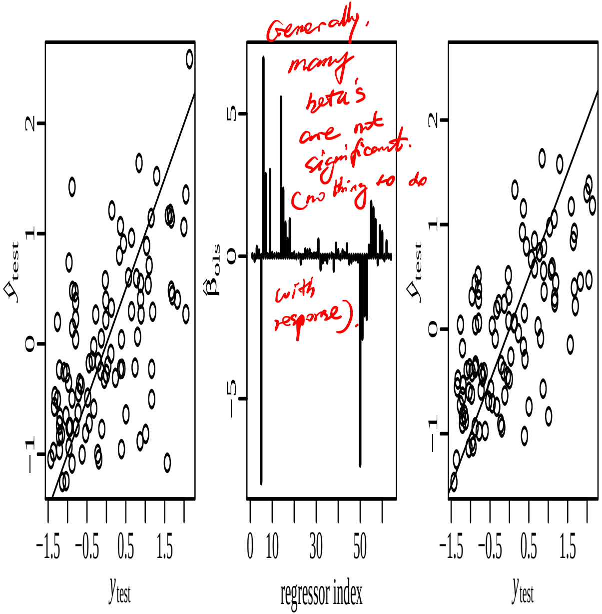
Model selection - diabetes data

Baseline data for ten variables $x_1, ..., x_{10}$ on a group of 442 diabetes patients were gathered, as well as a measure of disease progression y. The aim is to build a predictive model for y based on the baseline measurements. We are interested to assess a model with all main effects, two-way interactions and quadratic terms (p=64 regressors).

The variables are first centered to zero and scaled to have variance one.

We use cross-validation to evaluate the models: Training sample - 342 subjects; Testing sample - 100 subjects. Calculate average squared prediction error: $\frac{1}{100} \sum (y_{test,i} - \hat{y}_{test,i})^2$

Model selection - diabetes data

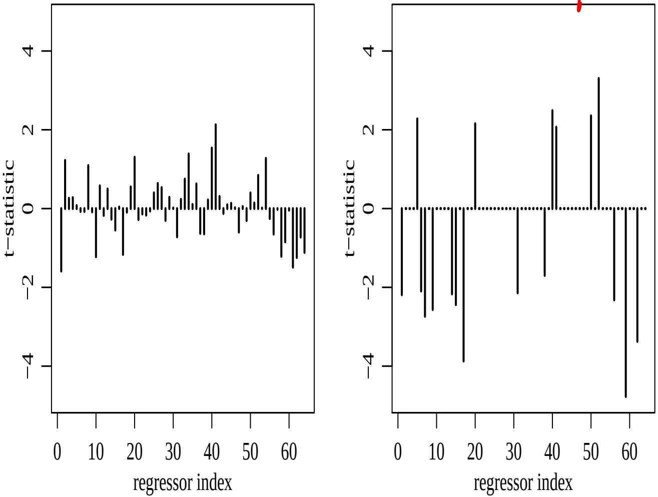


(note: the third plot is after backwards elimination)

Model selection - diabetes data

Suppose we create a new data vector $\tilde{\mathbf{y}}$ by randomly permuting the values of $\mathbf{y}$ and then regress this on $\mathbf{X}$. (The true association between $\tilde{\mathbf{y}}$ and the columns of $\mathbf{X}$ is zero).

BE, does n't take account of all the models.



Bayesian model comparison

sparsity.

Prior belief reflects the possibility that some of the regression coefficients are potentially equal to zero.

Let $\beta_j = z_j \times b_j$ where $z_j \in \{0, 1\}$ and b_j is some real number.

each combination of z corresponds to a model

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

Each value of $\mathbf{z} = (z_1, ..., z_p)$ corresponds to a different model. Let's obtain a posterior distribution for $\mathbf{z}$. Now we need a joint prior distribution on $\{z, \beta, \sigma^2\}$, so we can compute:

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

or

compare ratio

$$\mathrm{odds}(z_a, z_b|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{z_a}|\mathbf{y}, \mathbf{X})}{p(\mathbf{z_b}|\mathbf{y}, \mathbf{X})} = \frac{p(z_a)}{p(z_b)} \times \frac{p(\mathbf{y}|\mathbf{z_a}, \mathbf{X})}{p(\mathbf{y}|\mathbf{z_b}, \mathbf{X})}$$

Computing the marginal probability

$= \iint p(y, \beta, \sigma^2 | X, z) d\beta, d\sigma^2$

Assuming a g-prior distribution we can show that
$p(\mathbf{y}|\mathbf{X}, \mathbf{z}) = \pi^{-n/2} \frac{\Gamma([\nu_0+n]/2)}{\Gamma(\nu_0/2)} (1+g)^{-p_z/2} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + SSR_g^z)^{(\nu_0+n)/2}}$

Assume a unit information prior for $p(\sigma^2)$ and set g=n, $\nu_0 = 1$ and $\sigma_0^2 = \sigma^2_{ols,z}$. We have:

$$\frac{p(\mathbf{y}|\mathbf{X}, z_a)}{p(\mathbf{y}|\mathbf{X}, z_b)} = (1+n)^{(p_{z_b} - p_{z_a})/2} \left(\frac{s^2_{z_b} + SSR_g^{z_b}}{s^2_{z_a} + SSR_g^{z_a}} \right)$$

When is model z_b penalised relative to model z_a? When is model z_a penalised relative to model z_b??

- $SSR_g = y^T(\mathrm{I} - \frac{g}{g+1} X(X^T X)^{-1} X^T) y$.

Oxygen uptake example

z	model	log $p(\mathbf{y}\|\mathbf{Z}, \mathbf{X})$	$p(z\|\mathbf{y}, \mathbf{X})$
(1,0,0,0)	β_1	-44.33	0.00
(1,1,0,0)	$\beta_1 + \beta_2 \times \text{group}$	-42.53	0.00
(1,1,1,0)	$\beta_1 + \beta_3 \times \text{age}$	-37.66	0.18
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{group} + \beta_3 \times \text{age}$	-36.42	0.63
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{group} + \beta_3 \times \text{age} + \beta_4 \times \text{group} \times \text{age}$	-37.60	0.19

- assume all models are equally likely a priori
- unit information prior for σ^2 and g-prior for $\boldsymbol{\beta}$.
- Which model is most probable? Comment on the evidence for the effect of age and group respectively.

Gibbs sampling and model averaging

If there are p regression coefficients, how many different models are there to consider??

The number of models to search for can be really large, how can we carry out model selection in an efficient manner?

Gibbs sampling and model averaging

underlying idea.
(only do models with high prob.)

Implement a Gibbs sampling scheme:

$z^{(s)} \rightarrow \sigma^{2(s)} \rightarrow \beta^{(s)}$

$\downarrow$

$z^{(s+1)} \rightarrow \sigma^{2(s+1)} \rightarrow \beta^{(s+1)}$

The Gibbs sampler searches through the model space for values of z with higher posterior probability. For current value $z^{(s)} = (z_1, ..., z_p)$, generate a new value for z_j (j=1,...,p) by sampling from $p(z_j|\mathbf{y}, \mathbf{X}, z_{-j})$ (where z_{-j} refers to the values of z except the one corresponding to regressor j). Define conditional odds

$$o_j = \frac{Pr(z_j = 1|\mathbf{y}, \mathbf{X}, \mathbf{z}_{-j})}{Pr(z_j = 0|\mathbf{y}, \mathbf{X}, z_{-j})} = \frac{Pr(z_j = 1)}{Pr(z_j = 0)} \times \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_{-j}, z_j = 1)}{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_{-j}, z_j = 0)}$$

$$\frac{o_j}{1+o_j}$$

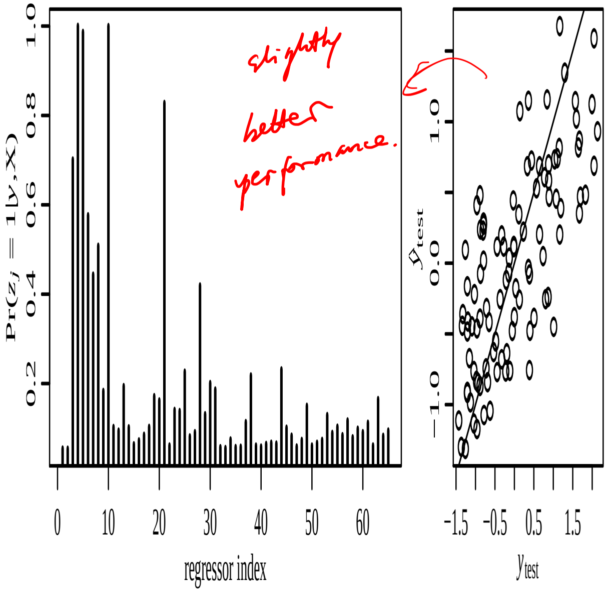
Gibbs sampling and model averaging

```
> lpy.X #calculates log of p(y|X)
function(y,X,
   g=length(y),nu0=1,s20=try(summary(lm(y~-1+X))$sigma^2,
        silent=TRUE))
{
  n<-dim(X)[1] ; p<-dim(X)[2]
  if(p==0) { s20<-mean(y^2) }
  H0<-0 ; if(p>0) { H0<- (g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X)
  SS0<- t(y)%*%( diag(1,nrow=n)  - H0 ) %*%y

  -.5*n*log(2*pi) +lgamma(.5*(nu0+n)) - lgamma(.5*nu0)
    - .5*p*log(1+g) + .5*nu0*log(.5*nu0*s20)
     -.5*(nu0+n)*log(.5*(nu0*s20+SS0))
}
```

Gibbs sampling and model averaging

```
z<-rep(1,dim(X)[2] ) #start off with all z_j ==1
lpy.c<-lpy.X(y,X[,z==1,drop=FALSE]) #starting value of log p(y|X
for(s in 1:S)
{
      for(j in sample(1:p)) #random permutation of j=1,...,p
    {
      zp<-z ; zp[j]<-1-zp[j]  (switch value of z_j)
      #recompute log p(y|X)
      lpy.p<-lpy.X(y,X[,zp==1,drop=FALSE])
       #conditional odds that z_j==1
      r<- (lpy.p - lpy.c)*(-1)^(zp[j]==0)
       #generate a value of z given conditional probability
      z[j]<-rbinom(1,1,1/(1+exp(-r)))
      #retain value of log p(y|X), if new draw of z_j
           is same as before
      if(z[j]==zp[j]) {lpy.c<-lpy.p}
     }
  Z[s,]<-z
```

Gibbs sampling and model averaging - diabetes example



slightly better performance.

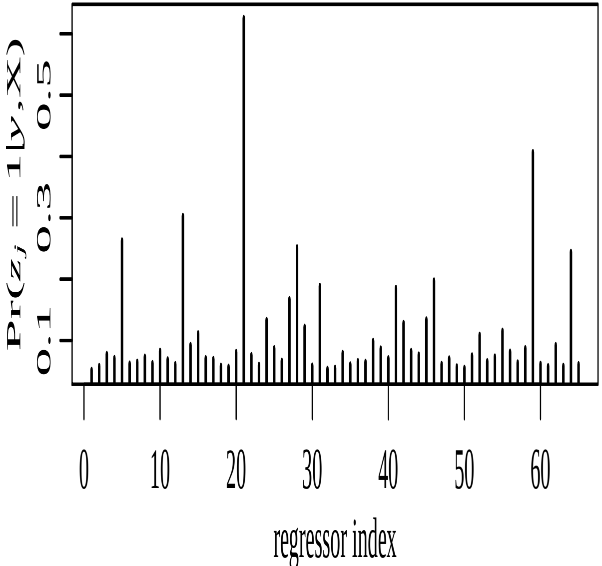
post. prob.

$$\hat{\beta}_{bma} = \sum_{s=1}^{S} \frac{\beta^s}{S} \times w^{(s)}$$

& basically just average the parameters.

It's your choice to chose a "S" based on how many models are "~~..~~ of high prob".

Gibbs sampling and model averaging - diabetes example (random permuation of y)



Exercise - Model selection

The dataset `achievement` contains information on 109 Austrian schoolchildren. The following variables were measured: `gender` (0 for male and 1 for female), `age` (in months), `IQ` , `Read1` , a test on assessing reading speed, and `Read2` , a test for assessing reading comprehension. One is interested in using a normal linear regression model to understand the variation in each of the reading tests based on the predictors gender, age, and IQ.

(a) Suppose one is interested in finding the best model to predict the `Read1` reading score. How many possible models are there? Use a Zellner g prior, and a Bayesian modelling stategy, which is the best model?

(b) Use a classical model-checking strategy to find the best regression model, and compare the best model with the best model chosen in part (a).