0. Some Review over Linear Regression

- **Essence of LR Model**: ***Linear Conditional Mean*** E[y | x]

$$E[y_i|x_i] = \int y\, p(y|x)\, dx = \beta^T x_i$$

With Normal Assumption of ***Normal Error*** $\epsilon_i \sim N(0, \sigma^2)$

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

- **Likelihood**: MVN form

$$y \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n) \;\; (ith\ row\ of\ X\beta = E[y_i \mid x_i])$$

which then in a full form,

$$p(y \mid \beta, \sigma^2) = \prod N(y_i; \beta, \sigma^2) = \left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right)$$

- **Frequentist inference**: A single $\hat{\beta}$ optimized to the data(OLS = MLE for LR)

$$\hat{\beta} = \arg\max_{\beta} p\left(y \mid \beta, \sigma^2\right) = \left(X^T X\right)^{-1} X^T y$$

Since $\hat{\beta}$ is a linear combination of y, it also follows Normal distribution

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

We have used this for inference such as interval estimation, hypothesis test and ect.

1. Preliminaries

- **Matching trick for Normal family**: $y \sim N(\mu, \Sigma)$

$$\begin{aligned} p(y \mid \mu, \Sigma) &\propto \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right) \\ &\propto \cdots \\ &\propto \exp\left(-\frac{1}{2} y^T \Sigma^{-1} y + y^T \Sigma^{-1} \mu\right) \end{aligned}$$

if $p(y) \propto \exp\left(-\frac{1}{2} y^T A y + y^T b\right)$, then $E(y) = A^{-1} b,\ V(y) = A^{-1}$

- **Inverse Gamma pdf**: $\sigma^2 \sim \Gamma^{-1}\left(\nu_0/2, \sigma_0^2/2\right) = \chi^{-2}\left(\nu_0, \sigma_0^2\right)$

$$p\left(\sigma^2 \mid \nu_0, \sigma_0^2\right) = \frac{\left(\nu_0 \sigma_0^2/2\right)^{\nu_0/2}}{\Gamma\left(\nu_0/2\right)} \left(\frac{1}{\sigma^2}\right)^{\nu_0+1} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right)$$

1. Frequentist approach: $\arg\max_\theta p(X|\theta) = \hat{\theta}$

2. Bayesian approach:

$$\begin{aligned} P(\theta \mid x) &= \frac{P(x \mid \theta)P(\theta)}{\int P(x,\theta)P(\theta)d\theta} \\ &\propto P(x \mid \theta)P(\theta) = P(\theta, x) \;:\; \textit{full probability model} \end{aligned}$$

Here we are going to discuss parameters β, σ^2 for linear regression

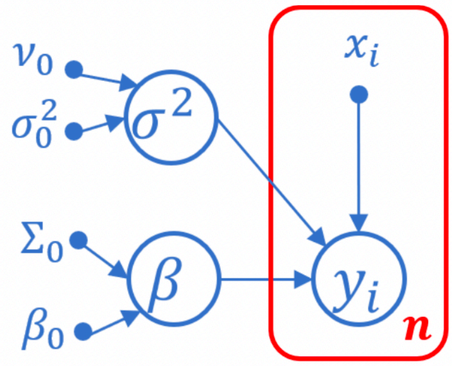
2. Bayesian Treatment of Linear Regression

- Full prob. model: $p(y, \beta, \sigma^2) = p(\beta, \sigma^2)p(y \mid \beta, \sigma^2)$

  Our goal: How are we gonna set the prior for $p(\beta, \sigma^2)$***??***

I will indtroduce two methods based on the FCB book and hun-learning lecture.

2-1.Semi-conjugate, independent prior:



Indep. but semi-conj. prior

$$\begin{aligned} p\left(y, \beta, \sigma^2\right) &= \left[p(\beta)p\left(\sigma^2\right)\right] p\left(y \mid \beta, \sigma^2\right) \\ \beta &\sim N\left(\beta_0, \Sigma_0\right) \\ \sigma^2 &\sim \Gamma^{-1}\left(\nu_0/2, \nu_0\sigma_0^2/2\right) \end{aligned}$$

1. **Slopes posterior** $\beta \mid y, \sigma^2 \sim N(\beta_n, \Sigma_n)$ / assumed σ^2 known

$$\begin{aligned} p\left(\beta \mid y, \sigma^2\right) &\propto p\left(y \mid \beta, \sigma^2\right) p(\beta) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left(y^T y - 2\beta^T X^T y + \beta^T X^T X\beta\right) - \frac{1}{2}\left(\beta^T \Sigma_0^{-1}\beta - 2\beta^T\Sigma_0^{-1}\beta_0\right)\right) \\ &\propto \exp(-\frac{1}{2}\beta^T \underbrace{\left(X^T X/\sigma^2 + \Sigma_0^{-1}\right)}_{\Sigma_n^{-1}}\beta + \beta^T \underbrace{\left(X^T y/\sigma^2 + \Sigma_0^{-1}\beta_0\right)}_{\Sigma_n^{-1}\beta_n}) \end{aligned}$$

$$\therefore \Sigma_n^{-1} = \underbrace{X^T X/\sigma^2}_{\text{data precision}} + \underbrace{\Sigma_0^{-1}}_{\text{prior precision}} \quad , \quad \beta_n = \Sigma_n(X^T y \underbrace{/\sigma^2}_{\text{data weight}} + \underbrace{\Sigma_0^{-1}}_{\text{prior weight}} \beta_0)$$

- $|\Sigma_0^{-1}| < \epsilon$ weak prior -> $\beta_n \approx \hat{\beta}_{mle}$
- Want to reduce bias from β_0 significantly?

  $\beta_0 = 0,\ \Sigma_0 = g\sigma^2(X^T X)^{-1}$ -> $\beta \mid y, \sigma^2 \sim N\left(\frac{g}{g+1}\hat{\beta}_{mle}, \frac{g}{g+1}V(\hat{\beta}_{mle})\right)$

  g-prior! Higher g means weaker prior / we usually give as n

2. **Error Variance posterior** $\sigma^2 \mid y, \beta \sim \chi^{-2}(\nu_n, \sigma_n^2)$

   Let $SSR(\beta) = ||y - X\beta||_2^2$

$$
\begin{aligned}
p\left(\sigma^2 \mid y, \beta\right) &\propto p\left(\sigma^2\right) p\left(y \mid \beta, \sigma^2\right) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\nu_0/2+1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{\mathrm{SSR}(\beta)}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sigma^2}\right)^{(\nu_0+n)/2+1} \exp\left(-\frac{\nu_0\sigma_0^2 + \mathrm{SSR}(\beta)}{2\sigma^2}\right) \\
\therefore \nu_n &= \nu_0 + n \quad \text{prior+data} \quad \text{(pooled) sample size} \\
\sigma_n^2 &= \left(\nu_0\sigma_0^2 + SSR(\beta)\right)/\nu_n \qquad \text{pooled variance}
\end{aligned}
$$

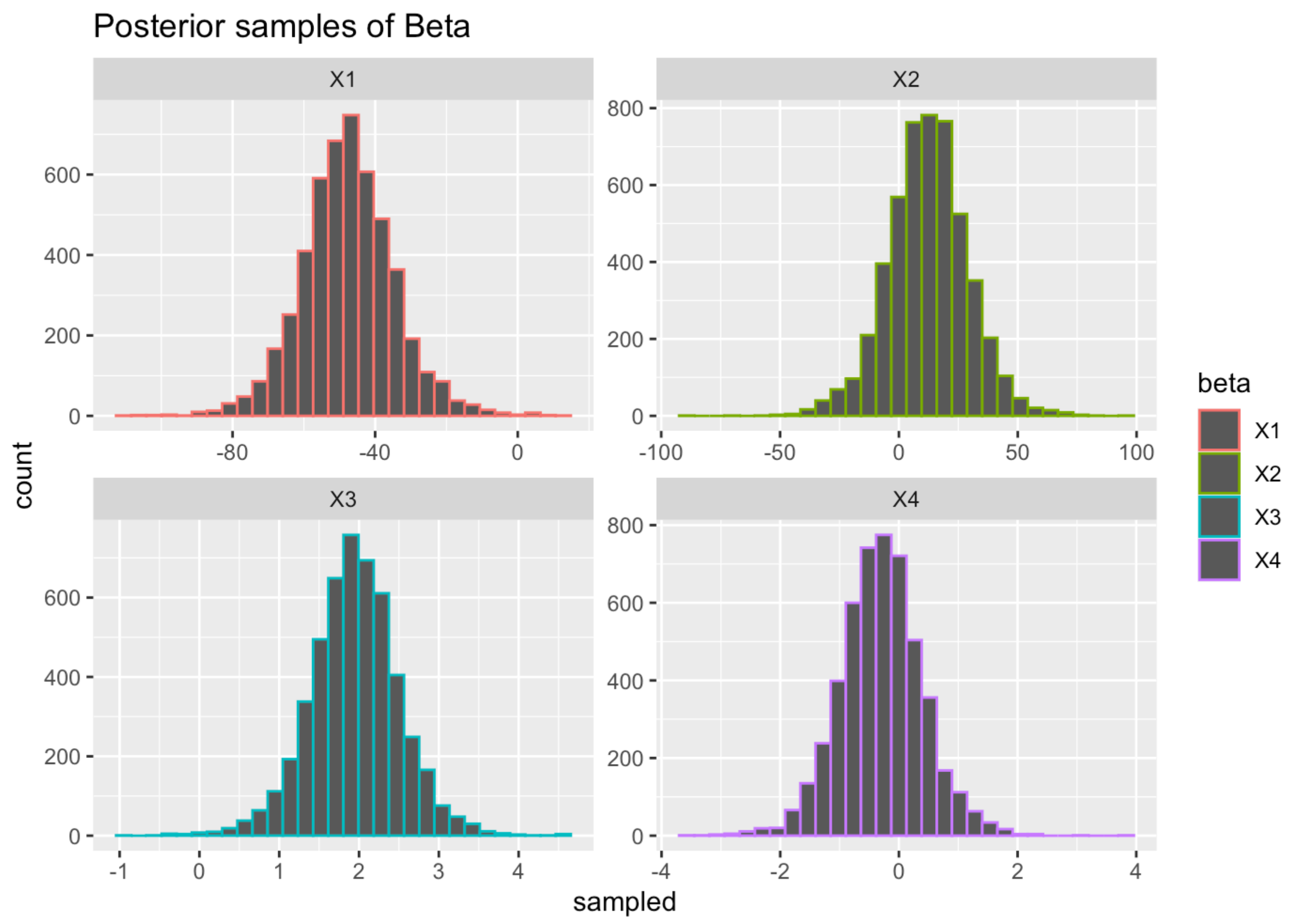
- Full conditional posteriors + Gibbs Sampling -> Joint posterior $\beta, \sigma^2 \mid y$
  1. Set initial estimate σ_0^2 and β_0 - usually straight out from the data (or doesn't matter since we are going to ditch the first half)
  2. Sample $\sigma^2 \sim \chi^{-2}(\nu_n, \sigma_n^2)$
  3. Sample $\beta \sim N\left(\frac{g}{g+1}\hat{\beta}_{mle}, \frac{g}{g+1}V(\hat{\beta}_{mle})\right)$
  4. Repeat 2~3, ditch the first half, use the rest for posterior inference!!

3. Code

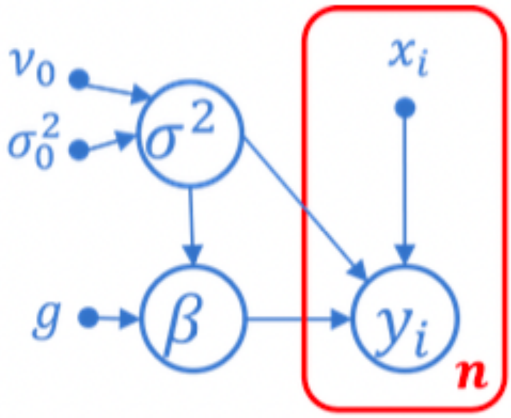
```
library(ggplot2)
library(reshape2)
library(dplyr)
library(ggpubr)
library(mvtnorm)
library(latex2exp)
library(tidyr)

DF = dget('http://www2.stat.duke.edu/~pdh10/FCBS/Inline/yX.o2uptake')
y = DF[,1]; X = DF[,-1]; inv = solve
### set prior and get necessary statistics
g = length(y) # g-prior for beta
nu0 = 1; s20 = summary(lm(y~-1+X))$sigma^2 # prior for sig^2
n = length(y); p = ncol(X)
### MCMC setup
S = 10000; set.seed(0827)
BETA = matrix(NA, nrow=S, ncol=p)
sigma2 = matrix(NA, nrow=S, ncol=1)
BETA[1,] = inv(t(X) %*% X) %*% t(X) %*% y # initial estimate
sigma2[1,] = s20 # initial estimate
### gibbs sampling
nun = nu0 + n
betan = (g/(g+1)) * inv(t(X) %*% X) %*% t(X) %*% y
```

```
for(s in 2:S){
  s2n = nu0*s20 + t(y- X %*% BETA[s-1,]) %*% (y- X %*% BETA[s-1,])
  sigma2[s,] = 1/rgamma(1, shape = nun/2, rate = s2n/2)
  Sigman = (g/(g+1)) * sigma2[s,] * inv(t(X) %*% X)
  BETA[s,] = MASS::mvrnorm(n=1,betan, Sigman) }
### display results
p = data.frame(BETA[(S/2+1):S,]) %>% gather(beta, sampled)%>%
  ggplot(aes(x=sampled, color=beta))+
  geom_histogram(bins=30)+ facet_wrap(~beta, scales='free')+
  labs(title="Posterior samples of Beta")
p
```



2-2.Full-conjugate, dependent prior:



Dep. but full conj. prior

$$
\begin{aligned}
p\left(y, \beta, \sigma^2\right) &= \left[p\left(\sigma^2\right) p\left(\beta \mid \sigma^2\right)\right] p\left(y \mid \beta, \sigma^2\right) \\
\beta \mid \sigma^2 &\sim N\left(0, g\sigma^2\left(X^T X\right)^{-1}\right) \qquad \text{g-prior!} \\
\sigma^2 &\sim \Gamma^{-1}\left(\nu_0/2, \nu_0\sigma_0^2/2\right)
\end{aligned}
$$

- **Posterior for full-conjugate prior**

$$
p\left(\beta, \sigma^2 \mid y\right) = \underbrace{p\left(\sigma^2 \mid y\right)}_{??} \underbrace{p\left(\beta \mid \sigma^2, y\right)}_{\text{posterior of g-prior}}
$$

-> $p(\sigma^2 \mid y) \propto p(\sigma^2)p(y \mid \sigma^2)$

Let m, V be post. mean and var. of $\beta \mid \sigma^2, y$ $\left(m = \frac{g}{g+1}\left(X^T X\right)^{-1} X^T y, V = \frac{g}{g+1}\sigma^2\left(X^T X\right)^{-1}\right)$

-
-
-

$$
\begin{aligned}
p\left(y \mid \sigma^2\right) &\propto \int p\left(y \mid \beta, \sigma^2\right) p\left(\beta \mid \sigma^2\right) d\beta \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \frac{1}{\left|g\sigma^2\left(X^T X\right)^{-1}\right|^{1/2}} \int \exp\left[-\frac{1}{2\sigma^2}\left(\|y - X\beta\|^2\right) - \frac{1}{2g\sigma^2}\beta^T X^T X\beta\right] d\beta \\
&= \left(\frac{1}{\sigma^2}\right)^{n/2} \frac{1}{\left|g\sigma^2\left(X^T X\right)^{-1}\right|^{1/2}} \exp\left[-\frac{1}{2\sigma^2} y^T y\right] \int \exp[\frac{1}{\sigma^2}\beta^T X^T y - \frac{1 + 1/g}{2\sigma^2} \underbrace{\beta^T X^T X\beta}_{\text{subtract } m \text{ from } \beta \text{, use V}}] d\beta \\
&= \left(\frac{1}{\sigma^2}\right)^{n/2} \frac{1}{\left|g\sigma^2\left(X^T X\right)^{-1}\right|^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left(y^T y - \sigma^2 m^T V^{-1} m\right)\right] \int \exp\left[-\frac{1}{2}(\beta - m)^T V^{-1}(\beta - m)\right] d\beta \\
&\propto \left(\frac{1}{2\pi}\right)^{n/2} (1 + g)^{-p/2} \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} SSR(g)\right]
\end{aligned}
$$

-

$$
\begin{aligned}
p\left(\sigma^2 \mid y\right) & \propto p\left(\sigma^2\right) p\left(y \mid \sigma^2\right) \\
& \propto\left(\frac{1}{\sigma^2}\right)^{\left(\nu_0+n\right) / 2+1} \exp \left(-\frac{\nu_0 \sigma_0^2+S S R(g)}{2 \sigma^2}\right)
\end{aligned}
$$

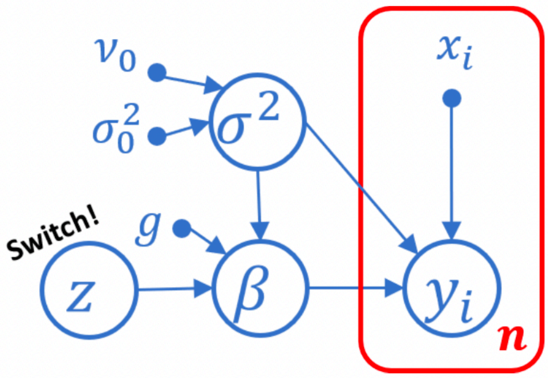
- With full conjugate posterior, the it's just sampling
  1. Sample $\sigma^2 \sim \chi^{-2}\left(\nu_0+n, \frac{\nu_0 \sigma_0^2+SSR(g)}{\nu_0+n}\right)$
  2. Sample $\beta \mid \sigma^2 \sim N\left(\frac{g}{g+1} \hat{\beta}_{mle}, \frac{g}{g+1} V\left(\hat{\beta}_{mle}\right)\right)$
  3. Use sampled $\left(\sigma^2, \beta\right)$ for inference

Bayesian Model Selection

Frequentist - Combinatorial Optimization using single metric e.g. AIC, BIC

- Idea : Introduce $z_j \in\{0,1\}$ which decides whether $\beta_j=z_i \cdot b_j \neq 0$ (i.e. included) or not.

$$
\begin{aligned}
&\text{Single data model } y_i=z_1 b_1 x_{i, 1}+z_2 b_2 x_{i, 2}+\ldots+z_p b_p x_{i, p}+\epsilon_i \\
&\text{Full prob. model} \quad p\left(y, \beta, \sigma^2, z\right)=p(z) p\left(\sigma^2\right) p\left(\beta \mid \sigma^2, z\right) p\left(y \mid \beta, \sigma^2\right)
\end{aligned}
$$



- Bayesian MS aim to get a distribution of the **"switch"** variable z given y.
- It does not give a single optimal model, but may "probable" models!

$$
p(z \mid y)=\frac{p(z) p(y \mid z)}{\sum_z p(z) p(y \mid z)} \quad \text{intractable denominator}
$$

- Note that $p(z \mid y) \propto p(z) p(y \mid z) \propto p(y \mid z)$ w/ uniform $p(z)$

  Let β_z, X_z, p_z be variables w/ $z_j=1$ and $p\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$

$$
\begin{aligned}
p(y \mid z) &= \iint p\left(y, \beta, \sigma^2 \mid z\right) d\beta d\sigma^2 \\
&= \int p\left(\sigma^2\right) \underbrace{\int p\left(y \mid \beta_z, \sigma^2\right) p\left(\beta_z \mid \sigma^2\right) d\beta d\sigma^2}_{p(y \mid \sigma^2, z)} \\
&\propto (1+g)^{-p_z/2} \int \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left[-\frac{\mathrm{SSR}(g,z)}{2\sigma^2}\right] d\sigma^2 \\
&\propto (1+g)^{-p_z/2}\, \mathrm{SSR}(g,z)^{-n/2} \\
&\quad \left(\because p\left(y \mid \sigma^2\right) \propto \left(\frac{1}{2\pi}\right)^{n/2} (1+g)^{-p/2} \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\mathrm{SSR}(g)\right]\right)
\end{aligned}
$$

- Bayes factor as below.

1. favors model that **explains data well** (low SSR)
2. **penalizes complex model** (low p)

$$
\frac{p\left(y \mid z_1\right)}{p\left(y \mid z_2\right)} = (1+g)^{(p_2-p_1)/2} \left(\frac{\mathrm{SSR}\left(g, z_2\right)}{\mathrm{SSR}\left(g, z_1\right)}\right)^{n/2}
$$

Gibbs Sampler for Bayesian Model Selection

- $p(z \mid y)$ requires full conditional distribution

$$
p\left(z_j = 1 \mid y, z_{-j}\right) = \frac{p\left(z_j = 1 \mid y, z_{-j}\right)}{p\left(z_j = 1 \mid y, z_{-j}\right) + \left(z_j = 0 \mid y, z_{-j}\right)} = \frac{1}{1+o_j}
$$

where $O_j = \frac{p(z_j=0 \mid y, z_{-j})}{p(z_j=1 \mid y, z_{-j})}$ is the conditional odds that $z_j = 1$, which is a ratio of $p(y \mid z)$

$$
o_j = \frac{p\left(z_j = 0 \mid y, z_{-j}\right)}{p\left(z_j = 1 \mid y, z_{-j}\right)} = \frac{p\left(z_j = 0\right)}{p\left(z_j = 1\right)} \times \frac{p\left(y \mid z_{-j}, z_j = 0\right)}{p\left(y \mid z_{-j}, z_j = 1\right)} = \frac{p\left(y \mid z_{-j}, z_j = 0\right)}{p\left(y \mid z_{-j}, z_j = 1\right)}
$$

- Gibbs sampler for Bayesian MS

1. given initial $z^{(s)}, \sigma^{(s)}, \beta^{(s)}$
2. update z: for each j, replace $z_j = 1 - z_j$ w/ probability $1/(1+o_j)$
3. Given z, draw $\sigma^2 \sim p(\sigma^2 \mid z, y)$, $\beta \sim p(\beta \mid \sigma^2, y)$

- Code

```
load("/Users/kwanseok/Downloads/diabetes.RData")
DF = diabetes
y = as.matrix(DF$y); X = as.matrix((DF$X)[,1:10]); inv = solve

# function of prop ro log p(y \mid z)
lpy = function(y, X, g = length(y), nu0 = 1, s20 =
```

```
                  try(summary(lm(y~-1+ X))$sigma^2, silent=T)){
  n = nrow(X); p = ncol(X);
  if(p==0) Hg = 0; s20 = mean(y^2) # null model
  if(p > 0) Hg = (g/(g+1)) * X %*% inv(t(X) %*% X) %*% t(X)
  SSRg = t(y) %*% (diag(1, n) - Hg) %*% y
  return( -p/2 * log(1+g) - n/2 * log(SSRg) ) } # log p(y \mid z)

# MCMC setup
z = rep(1, ncol(X)) # MCMC setup
lpy.c = lpy(y, X[, z==1, drop=F])
S = 1000; Z = matrix(NA, S, ncol(X))

for(s in 1:S){# Gibbs sampler
  for(j in sample(1:ncol(X))){ # iterate over variables randomly
    zp = z; zp[j] = 1-zp[j]
    lpy.p = lpy(y, X[, zp==1, drop=F])
    o = (lpy.p - lpy.c) * (-1)^(zp[j] == 1)
    z[j] = rbinom(1, 1, 1/(1+exp(o))) # full cond. dist of zj
    if(z[j] == zp[j]) lpy.c = lpy.p }
  Z[s,] = z; if(s %% 100 == 0) cat(s, "\t") }

colz = colMeans(Z[(S/2+1):S,])
```