1 Least Squares Estimation with Example (Hoff p. 149ff.)

Regression modeling is concerned with describing how the sampling istribution of one random variable Y varies with another variable or set of variables $\mathbf{x} = (x_1, ..., x_p)$. Specifically, a regression model postulates a form for $p(y|\mathbf{x})$, the conditional distribution of Y given $\mathbf{x}$. Estimation of $p(y|\mathbf{x})$ is made using data $y_1, ..., y_n$ that are gathered under a variety of conditions $\mathbf{x}_1, ..., \mathbf{x}_n$.

The normal linear regression model specifies that, in addition to $E[Y|\mathbf{x}]$ being linear, the sampling variability around the mean is i.i.d. normal:

$$\begin{aligned} \epsilon_1, ..., \epsilon_n &\overset{\text{i.i.d}}{\sim} \text{normal}\left(0, \sigma^2\right) \\ Y_i &= \beta^T \mathbf{x}_i + \epsilon_i \end{aligned}$$

This model provides a complete specification of the joint probability density of observed data $y_1, ..., y_n$ conditional upon $\mathbf{x}_1, ..., \mathbf{x}_n$ and values of β and σ^2:

$$\begin{aligned} p\left(y_1, ...y_n|\mathbf{x}_1, ..., \mathbf{x}_n, \beta, \sigma^2\right) &= \prod_{i=1}^{n} p\left(y_i|\mathbf{x}_i, \beta, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta^T\mathbf{x}_i\right)^2\right\} \end{aligned} \tag{1}$$

Another way to write this joint probability density is in terms of the multivariate normal distribution: Let $\mathbf{y}$ be the n-dimensional column vector $(y_1, ..., y_n)^T$ and let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,p}\}$. Then the normal regression model is

$$\{\mathbf{y}|\mathbf{X}, \beta, \sigma^2\} \sim \text{multivariate normal}\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right),$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and

$$\mathbf{X}\beta = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} E\left[Y_1|\beta, \mathbf{x}_1\right] \\ \vdots \\ E\left[Y_n|\beta, \mathbf{x}_n\right] \end{pmatrix}$$

The density (1) depends on β through the residuals $\left(y_i - \beta^T\mathbf{x}_i\right)$. We compute the ordinary least squares estimates

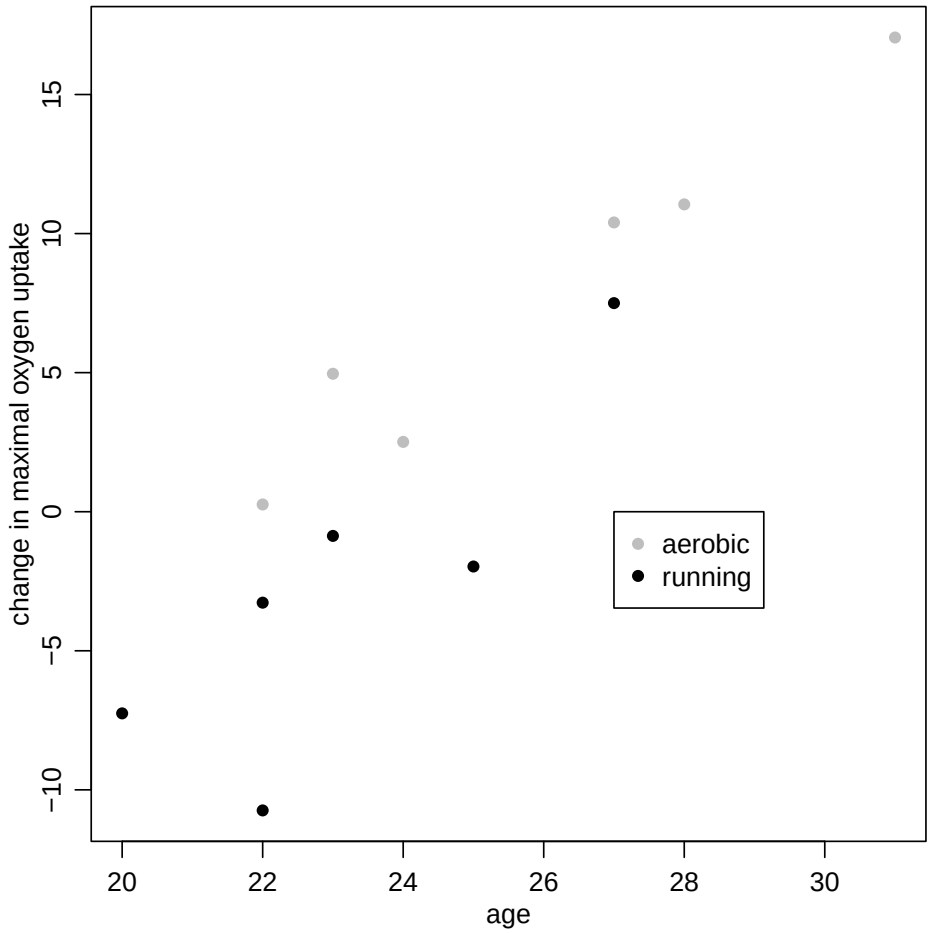
$$\hat{\beta}_{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

and

$$\hat{\sigma}^2_{ols} = \frac{SSR\left(\hat{\beta}_{ols}\right)}{(n-p)} = \frac{\sum\left(y_i - \hat{\beta}^T_{ols}x_i\right)^2}{(n-p)}.$$

Example: Oxygen uptake (from Kuehl (2000), Hoff p. 149ff)

Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program. The maximum oxygen uptake of each subject was measured (in liters per minute) while running on an inclined treadmill, both before and after the 12-week program. Of interest is how a subject's change in maximal oxygen uptake may depend on which program they were assigned to. However, other factors, such as age, are expected to affect the change in maximal uptake as well. The results are shown here:



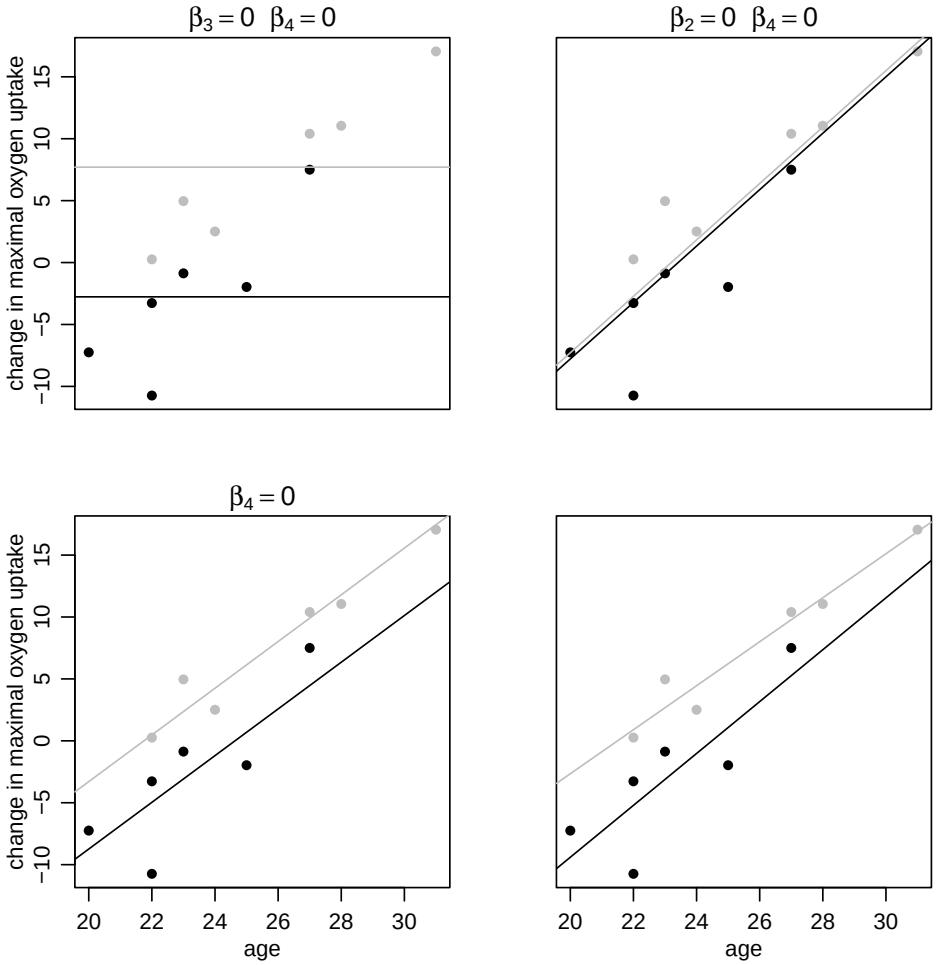
Hoff's regression model:

$$
\begin{aligned}
Y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where} \\
x_{i,1} &= 1 \text{ for each subject } i \\
x_{i,2} &= 0 \text{ if subject } j \text{ is on the running program, 1 if on aerobic} \\
x_{i,3} &= \text{ age of subject } i \\
x_{i,4} &= x_{i,2} \times x_{i,3}
\end{aligned}
\tag{2}
$$

Under this model the conditional expectations of Y for the two different levels of $x_{i,1}$ are

$$E[Y|\mathbf{x}] = \beta_1 + \beta_3 \times \text{ age if } x_1 = 0, \text{ and}$$
$$E[Y|\mathbf{x}] = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{ age if } x_1 = 1$$

In other words, the model assumes that the relationship is linear in age for both exercise groups, with the difference in intercepts given by β_2 and the difference in slopes given by β_4. If we assumed that $\beta_2 = \beta_4 = 0$, then we would have identical lines for both groups. If we assumed $\beta_4 = 0$ then we would have a different line for each group but they would be parallel. Allowing all coefficients to be non-zero gives us two unrelated lines. Some different possibilities are depicted graphically below:



Let's find the least squares regression estimates for the model in 2, and use the results to evaluate the differences between the two exercise groups. The ages of the 12 subjects, along with their observed changes in maximal oxygen uptake, are

$$\mathbf{x}_3 = (23, 22, 22, 25, 27, 20, 31, 23, 27, 28, 22, 24)$$
$$\mathbf{y} = (-0.87, -10.74, -3.27, -1.97, 7.50, -7.25, 17.05, 4.96, 10.40, 11.05, 0.26, 2.51),$$

with the first six elements of each vector corresponding to the subjects in the running group and the latter six corresponding ot sujects in the aerobics group. After constructing

the 12 matrix $\mathbf{X}$ out of the vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ defined as in (2), the matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ can be computed, from which we get $\beta_{ols} = (-51.29, 13.11, 2.09, -0.32)^T$:

```
> n <- length(y)
> x1 <- rep(1,n)
> x4 <- x2*x3
> X = cbind(x1,x2,x3,x4)
> p = ncol(X)
> beta.ols<- solve(t(X)%*%X)%*%t(X)%*%y
> sig2.ols = (t(y-X%*%beta.ols)%*%(y-X%*%beta.ols)/(n-p))[1]
> sig2.ols

[1] 8.542477

> #sampling variance-covariance matrix of beta.ols:
> SIG2.ols = solve(t(X)%*%X)*sig2.ols
> #standard errors for the components of beta.ols:
> SE.ols = sqrt(diag(SIG2.ols))
> betadata = cbind(beta.ols,SE.ols)
> colnames(betadata) = c("beta.ols","SE.ols")
> betadata

      beta.ols     SE.ols
x1 -51.2939459 12.2522126
x2  13.1070904 15.7619762
x3   2.0947027  0.5263585
x4  -0.3182438  0.6498086
```

This means that the estimated linear relationship between uptake and age has an intercept and slope of -51.29 and 2.09 for the running group, and -51.29 + 13.11 = -38.18 and 2.09 - 0.32 = 1.77 for the aerobics group. These two lines are plotted in the fourth panel of Figure XX. We obtain unbiased estimate $\sigma^2 = SSR(\hat{\beta}_{ols})/(n-p) = 8.54$, and use this to compute the standard error of the components of $\hat{\beta}_{ols}$, which are 12.25, 15.76, 0.53, and 0.65, respectively. comparing the values of $\hat{\beta}_{ols}$ to their standard errors suggests that the evidence for differences between the two exercise regimens is not very strong.

2 Bayesian Estimation for a Regression Model (Hoff p. 154ff)

2.1 A semiconjugate prior distribution

A semiconjugate prior distribution for β and σ^2 is used when there is information available about the parameters. The sampling density of the data (Equation 1) is

$$p(\mathbf{y}|\mathbf{X},\beta,\sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}\text{SSR}(\beta)\} = \exp\{-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta]\}.$$

The role that β plays in the exponent looks very similar to that played by $\mathbf{y}$, and the distribution of $\mathbf{y}$ is multivariate normal. This suggests that a multivariate normal prior distribution for β is conjugate: if $\beta \sim$ multivariate normal(β_0, Σ_0), then

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

This is proportional to a multivariate normal density, with

$$Var[\beta|\mathbf{y},\mathbf{X},\sigma^2] = (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1} \tag{3}$$

$$\text{E}[\beta|\mathbf{y},\mathbf{X},\sigma^2] = (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\beta_0 + \mathbf{X}^T\mathbf{y}/\sigma^2). \tag{4}$$

As usual, we can gain some understanding of these formulae by considering some limiting cases. If the elements of the prior precision matrix Σ_0^{-1} are small in magnitude, then the conditional expectation $\text{E}[\beta|\mathbf{y},\mathbf{X},\sigma^2]$ is approximately equal to $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, the least squares estimate. On the other hand, if the measurement precision is very small (σ^2 is very large), then the expectation is approximately β_0, the prior expectation.

As in most normal sampling problems, the semiconjugate prior distribution for σ^2 is an inverse-gamma distriution. Letting $\gamma = 1/\sigma^2$ be the measurement precision, if $\gamma \sim$ gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$, then

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

which we recognize as a gamma density, so that

$$\{\sigma^2|\mathbf{y},\mathbf{X},\beta\} \sim \text{inverse-gamma}([\nu_0+n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\beta)]/2).$$

Constructing a Gibbs sampler to approximate the joint posterior distribution $p(\beta,\sigma^2|\mathbf{y},\mathbf{X})$ is then straightforward: given current values $\{\beta^{(s)}, \sigma^{2(s)}\}$, new values can be generated by

1. updating β:
   (a) compute $\mathbf{V} = \text{Var}[\beta|\mathbf{y}, \mathbf{X}, \sigma^{2(s)}]$ and $\mathbf{m} = \text{E}[\beta|\mathbf{y}, \mathbf{X}, \sigma^{2(s)}]$
   (b) sample $\beta^{(s+1)} \sim$ multivariate normal$(\mathbf{m}, \mathbf{V})$
2. updating σ^2:
   (a) compute $\text{SSR}(\beta^{(s+1)})$
   (b) sample $\sigma^{2(s+1)} \sim$ inverse-gamma$([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\beta^{(s+1)})]/2)$.

```
> #### Bayesian estimation via MCMC
> x2<-c(0,0,0,0,0,0,1,1,1,1,1,1)
> x3<-c(23,22,22,25,27,20,31,23,27,28,22,24)
> y<-c(-0.87,-10.74,-3.27,-1.97,7.50,-7.25,17.05,4.96,10.40,11.05,0.26,2.51)
> x1 <- rep(1,n)
> x4 <- x2*x3
> X = cbind(x1,x2,x3,x4)
> n<-length(y)
> #X<-cbind(rep(1,n),x1,x2,x1*x2)
> p<-dim(X)[2]
> fit.ls<-lm(y~-1+ X) #linear regression fit omitting intercept term
> beta.0<-rep(0,p) ; Sigma.0<-diag(c(150,30,6,5)^2,p) # SOME INITIAL
> nu.0<-1 ; sigma2.0<- 15^2                          # VALUES
> beta.0<-fit.ls$coef                                # beta.0 is the coefficients o
> nu.0<-1  ; sigma2.0<-sum(fit.ls$res^2)/(n-p)       # nu.0 = # prior obs, sigma2.0
> Sigma.0<- solve(t(X)%*%X)*sigma2.0*n               # Sigma.0 is sampling variance
> S<-5000
> rmvnorm<-function(n,mu,Sigma)
+ { # samples from the multivariate normal distribution
+   E<-matrix(rnorm(n*length(mu)),n,length(mu))
+   t(  t(E%*%chol(Sigma)) +c(mu))
+ }
> ## some convenient quantites
> n<-length(y)
> p<-length(beta.0)
> iSigma.0<-solve(Sigma.0)                           #iSigma.0 = inverse of Sigma.0
> XtX<-t(X)%*%X
> ## store mcmc samples in these objects
> beta.post<-matrix(nrow=S,ncol=p)                   #storage for S instances for e
> sigma2.post<-rep(NA,S)                             #storage for S instances of va
> ## starting value
> set.seed(1)
> sigma2<- var( residuals(lm(y~0+X)) )               #starting with the variance of
> ## MCMC algorithm
> for( scan in 1:S) {
+
+ #update beta                                       #Formulas and steps Hoff p. 15
```

```
+ V.beta<- solve(  iSigma.0 + XtX/sigma2 )            #Conditional variance of the r
+ E.beta<- V.beta%*%( iSigma.0%*%beta.0 + t(X)%*%y/sigma2 ) #Conditional mean of the
+ beta<-t(rmvnorm(1, E.beta,V.beta) )                #Gibbs sampler step:  update b
+
+ #update sigma2
+ nu.n<- nu.0+n                                      #numerator of 1st term of inve
+ ss.n<-nu.0*sigma2.0 + sum(  (y-X%*%beta)^2 )       #numerator of 2nd term of inve
+ sigma2<-1/rgamma(1,nu.n/2, ss.n/2)                 #Gibbs ampler step:  update si
+
+ #save results of this scan
+ beta.post[scan,]<-beta                             #Store updated beta in current
+ sigma2.post[scan]<-sigma2                          #Store updated sigma^2 in curr
+                          }
> round( apply(beta.post,2,mean), 3)                 #compute mean of Gibbs sampled

[1] -50.943  12.656   2.079  -0.300

>
```

2.2 Default and weakly informative prior distributions

In situations where prior information is unavailable or difficult to quantify, an alternative "default" class of prior distributions is given. Specification of the prior parameters (β_0, Σ_0) and (ν_0, σ_0^2) that represent actual prior information for a Bayesian analysis can be difficult. For a prior distribution that is not going to represent real prior information about the parameters, we choose one that is as minimally informative as possible. The resulting posterior distribution, then, will represent the posterior information of someone who began with little knowledge of the population being studied. Here we will employ Zellner's "g-prior" (Zellner, 1986). We choose $\beta_0 = \mathbf{0}$ and $\Sigma_0 = k(\mathbf{X}^T\mathbf{X})^{-1}, k = g\sigma^2, g > 0$, which satisfies a desired condition that the regression parameter estimation be invariant to changes in the scale of the regressors. With this, equations 3 and 4 reduce to

$$\mathrm{Var}[\beta|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X^T X}/(g\sigma^2) + \mathbf{X^T X}/\sigma^2]^{-1} = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1} \tag{5}$$

$$\mathrm{E}[\beta|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X^T X}/(g\sigma^2) + \mathbf{X^T X}/\sigma^2]^{-1}\mathbf{X^T y}/\sigma^2 = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1}\mathbf{X^T y}. \tag{6}$$

Letting

$$\mathbf{V} = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1} \text{ and } \mathbf{m} = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1}\mathbf{X^T y}$$

we arrive at posteriors

$$\{\sigma^2|\mathbf{y}, \mathbf{X}\} \sim \text{inverse-gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \mathrm{SSR}_g]/2) \tag{7}$$

$$\{\beta|\mathbf{y}, \mathbf{X}, \sigma^2\} \sim \text{multivariate normal}\left(\frac{g}{g+1}\hat{\beta}_{ols}, \frac{g}{g+1}\sigma^2[\mathbf{X^T X}]^{-1}\right). \tag{8}$$

Here $\text{SSR}_g = \mathbf{y^T y} - \mathbf{m^T V^{-1} m} = \mathbf{y^T}(\mathbf{I} - \frac{g}{g+1}\mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T})\mathbf{y}$.

Simple Monte Carlo approximation can be used to sample from the joint posterior density $p(\sigma^2, \beta|\mathbf{y}, \mathbf{X})$ as follows:

1. sample $\sigma^2 \sim$ inverse-gamma$([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2)$
2. sample $\beta \sim$ multivariate normal $\left(\frac{g}{g+1}\hat{\beta}_{ols}, \frac{g}{g+1}\sigma^2[\mathbf{X^T X}]^{-1}\right)$.