Notes on Bayesian Linear Regression

Adam N. Smith

July 31, 2021

This note derives the posterior distribution of a Bayesian linear regression model with conjugate priors and may be used as a companion to chapter 2.8 in Rossi et al. (2005). We first define the model and derive the posterior. We conclude with a discussion of efficient posterior sampling based on the Cholesky decomposition.

1 Model

The standard multiple linear regression model relates a response variable y_i to a k-dimensional vector of predictor variables $x_i = (x_{i1}, \ldots, x_{ik})$ for $i = 1, \ldots, n$.

$$y_i = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \tag{1}$$

The parametric assumption on the error terms induces a distribution on y given x_i. Collecting the predictor variables into a matrix X allows us to rewrite the model in matrix notation:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{2}$$

where y is an n-dimensional vector of response variables, X is an $n \times k$ design matrix, β is a k-dimensional vector of regression coefficients, and ε is an n-dimensional vector of errors assumed to have a $N(0, \sigma^2 I_n)$ distribution.

The two key ingredients for a Bayesian analysis of the regression model above are the likelihood (i.e., distribution of the data) and prior (i.e., distribution of parameters).

Likelihood Assuming the error vector is distributed $N(0, \sigma^2 I_n)$ induces a multivariate normal likelihood:

$$p(y|X, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}. \tag{3}$$

Prior We choose conjugate priors for (β, σ^2) to ensure an analytic expression for the posterior distribution. With both β and σ^2 unknown, the conjugate prior is specified as $p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2)$ where

$$\beta|\sigma^2 \sim N(\bar{\beta}, \sigma^2 A^{-1}) \tag{4}$$
$$\sigma^2 \sim \frac{\nu_0 s_0^2}{\chi^2_{\nu_0}}. \tag{5}$$

Hence, $\beta|\sigma^2$ has a normal prior and σ^2 has scaled inverse chi-squared prior.[1]

Posterior The joint posterior then takes the form:

$$\begin{aligned} p(\beta, \sigma^2|y, X) &= p(y|X, \beta, \sigma^2)\ p(\beta|\sigma^2)\ p(\sigma^2) \\ &\propto (\sigma^2)^{-\frac{n}{2}}\ \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\} \\ &\quad \times (\sigma^2)^{-\frac{k}{2}}\ \exp\left\{-\frac{1}{2\sigma^2}(\beta - \bar{\beta})'A(\beta - \bar{\beta})\right\} \\ &\quad \times (\sigma^2)^{-(\frac{\nu_0}{2}+1)}\ \exp\left\{-\frac{\nu_0 s_0^2}{2\sigma^2}\right\}. \end{aligned} \tag{6}$$

2 Deriving the Posterior

Now, the goal is to derive expressions for the marginal posterior distributions of $\beta|\sigma^2$ and σ^2.

Step 1. Quadratic Forms We can simplify (6) by noticing that $p(y|X, \beta, \sigma^2)$ and $p(\beta|\sigma^2)$ both contain quadratic forms in β. The first step is to then expand out the sum of the two quadratic forms.

$$\begin{aligned} (y{-}X\beta)'&(y - X\beta) + (\beta - \bar{\beta})'A(\beta - \bar{\beta}) \\ &= (y' - \beta'X')(y - X\beta) + (\beta' - \bar{\beta}')A(\beta - \bar{\beta}) \\ &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta + \beta'A\beta - \beta'A\bar{\beta} - \bar{\beta}'A\beta + \bar{\beta}'A\bar{\beta} \\ &= \beta'X'X\beta + \beta'A\beta - 2\beta'X'y - 2\beta'A\bar{\beta} + y'y + \bar{\beta}'A\bar{\beta} \end{aligned} \tag{7}$$

[1] Note the equivalence between the scaled inverse chi-squared distribution and the inverse gamma distribution: if $\sigma^2 \sim (\nu_0 s_0^2)/\chi^2_{\nu_0}$ then $\sigma^2 \sim IG(\nu_0/2, \nu_0 s_0^2/2)$.

The last line uses the fact that $y'X\beta$ and $\beta'A\bar{\beta}$ are both scalars, so $y'X\beta = (y'X\beta)'$ and $\beta'A\bar{\beta} = (\beta'A\bar{\beta})'$. Now write (7) as

$$\Big[\beta'(X'X + A)\beta - \beta'(2X'y + 2A\bar{\beta})\Big] + y'y + \bar{\beta}'A\bar{\beta}. \tag{8}$$

We can further simplify the terms in $[\cdot]$ by completing the square in β.

Step 2. Completing the Square The matrix version of completing the square is given by:

$$X'MX + X'n + p = (X - h)'M(X - h) + k \tag{9}$$

where $h = -\frac{1}{2}M^{-1}n$ and $k = p - \frac{1}{4}n'M^{-1}n$. Next, we plug in the matrices from (8) into the general form given above.

$$\begin{aligned} M &= X'X + A \\ n &= -2(X'y + A\bar{\beta}) \\ h &= (X'X + A)^{-1}(X'y + A\bar{\beta}) \\ k &= -(X'y + A\bar{\beta})'(X'X + A)^{-1}(X'y + A\bar{\beta}) \\ p &= 0 \end{aligned}$$

If we let $\tilde{\beta} = h = (X'X + A)^{-1}(X'y + A\bar{\beta})$, then the bracketed terms in (8) become

$$\begin{aligned} &\beta'(X'X + A)\beta - \beta'(2X'y + 2A\bar{\beta}) \qquad (10) \\ &= (\beta - \tilde{\beta})'(X'X + A)(\beta - \tilde{\beta}) - (X'y + A\bar{\beta})'(X'X + A)^{-1}(X'y + A\bar{\beta}) \\ &= (\beta - \tilde{\beta})'(X'X + A)(\beta - \tilde{\beta}) - (X'y + A\bar{\beta})'\tilde{\beta} \end{aligned}$$

Now since $(X'X + A)^{-1}$ is symmetric and $I = \Big[(X'X + A)^{-1}(X'X + A)\Big]$, we can write the rightmost term above as

$$\begin{aligned} (X'y + A\bar{\beta})'\tilde{\beta} &= (X'y + A\bar{\beta})'\Big[(X'X + A)^{-1}(X'X + A)\Big]\tilde{\beta} \qquad (11) \\ &= (X'y + A\bar{\beta})'\big((X'X + A)^{-1}\big)'(X'X + A)\tilde{\beta} \\ &= \Big[(X'X + A)^{-1}(X'y + A\bar{\beta})\Big]'(X'X + A)\tilde{\beta} \\ &= \tilde{\beta}'(X'X + A)\tilde{\beta}. \end{aligned}$$

Therefore, using the results of equations (8), (10), and (11), (7) simplifies to

$$\begin{aligned}(y-X\beta)'(y-X\beta) &+ (\beta-\bar{\beta})'A(\beta-\bar{\beta}) \\ &= (\beta-\tilde{\beta})'(X'X+A)(\beta-\tilde{\beta}) + y'y + \bar{\beta}'A\bar{\beta} - \tilde{\beta}'(X'X+A)\tilde{\beta}.\end{aligned} \tag{12}$$

Step 3. Main Result The joint posterior distribution is then

$$\begin{aligned}p(\beta,\sigma^2|y,X) \propto (\sigma^2)^{-\frac{n}{2}}\ &\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\} \\ &\times (\sigma^2)^{-\frac{k}{2}}\ \exp\left\{-\frac{1}{2\sigma^2}(\beta-\bar{\beta})'A(\beta-\bar{\beta})\right\} \\ &\times (\sigma^2)^{-(\frac{\nu_0}{2}+1)}\ \exp\left\{-\frac{\nu_0 s_0^2}{2\sigma^2}\right\} \\ = (\sigma^2)^{-\frac{k}{2}}\ &\exp\left\{-\frac{1}{2\sigma^2}(\beta-\tilde{\beta})'(X'X+A)(\beta-\tilde{\beta})\right\} \\ &\times (\sigma^2)^{-(\frac{n+\nu_0}{2}+1)}\ \exp\left\{-\frac{(\nu_0 s_0^2 + y'y + \bar{\beta}'A\bar{\beta} - \tilde{\beta}'(X'X+A)\tilde{\beta})}{2\sigma^2}\right\}.\end{aligned} \tag{13}$$

But now we see that the joint posterior distribution factors into two parts: the conditional posterior of $\beta|\sigma^2$ and the marginal posterior of σ^2. Formally, we have

$$\beta|\sigma^2, y, X \sim N\Big(\tilde{\beta}, \sigma^2(X'X+A)^{-1}\Big) \tag{14}$$

$$\sigma^2|y, X \sim \frac{\nu_n s_n^2}{\chi^2_{\nu_n}} \tag{15}$$

where

$$\tilde{\beta} = (X'X+A)^{-1}(X'y + A\bar{\beta}) \tag{16}$$

$$\nu_n = \nu_0 + n \tag{17}$$

$$s_n^2 = \frac{\nu_0 s_0^2 + y'y + \bar{\beta}'A\bar{\beta} - \tilde{\beta}'(X'X+A)\tilde{\beta}}{\nu_0 + n}. \tag{18}$$

3 Sampling from the Posterior

Sampling from the posterior above is exact by virtue of conjugacy. That is, we can generate iid draws from the posteriors of $\beta|\sigma^2$ and σ^2 using standard software. However, drawing from $\beta|\sigma^2$ requires computing $(X'X+A)^{-1}$ (i.e., the inverse of the posterior precision) which is the inverse of a $k \times k$ matrix. This type of matrix inverse regularly appear in the computation of posterior moments, especially in Bayesian regression models. When k is large, this matrix inverse becomes more computationally demanding and can be a bottleneck in a posterior sampling routine.

Rossi et al. (2005) describe the Bayesian regression model with an eye towards efficient computation. The goal of this section is to provide the necessary background information to understand their approach. We start by defining the Cholesky decomposition which is a common method for matrix factorization.

Definition 1. *The Cholesky decomposition of a symmetric positive-definite matrix Σ is defined as $\Sigma = U'U$ where U is the upper triangular "Cholesky root" matrix.*

Consider the following simple example (based in R).

```
> Sigma
     [,1] [,2]
[1,]    2    1
[2,]    1    3
> U = chol(Sigma)
> U
         [,1]      [,2]
[1,] 1.414214 0.7071068
[2,] 0.000000 1.5811388
> t(U)%*%U
     [,1] [,2]
[1,]    2    1
[2,]    1    3
```

Now consider the problem of inverting Σ. The most straightforward approach is to use the `solve()` function in R.

```
> solve(Sigma)
     [,1] [,2]
[1,]  0.6 -0.2
[2,] -0.2  0.4
```

However, a more efficient approach is to use the Cholesky decomposition of Σ.

Definition 2. *If Σ is a symmetric positive-definite matrix with Cholesky decomposition $\Sigma = U'U$, then $\Sigma^{-1} = (U^{-1})(U^{-1})'$.*

This result shows that the inverse of Σ can be computed only using the inverse of the Cholesky root U. That is, we have replaced the problem of inverting Σ with the problem of inverting U. The fact that U is upper triangular leads to faster and more numerically stable inversion methods relative to a dense matrix like Σ. The following R code uses the previous result to compute Σ^{-1}.

```
> U = chol(Sigma)
> U
         [,1]      [,2]
[1,] 1.414214 0.7071068
[2,] 0.000000 1.5811388
> IR = backsolve(U,diag(ncol(U)))
> IR
          [,1]       [,2]
[1,] 0.7071068 -0.3162278
[2,] 0.0000000  0.6324555
> IR%*%t(IR)
     [,1] [,2]
[1,]  0.6 -0.2
[2,] -0.2  0.4
```

Here `IR` refers to the "inverse (Cholesky) root" of Σ. Also notice that `backsolve()` is used in place of `solve()` for computing `IR`. While `solve(U)` is equivalent to `backsolve(U,diag(nrow(U)))`, `backsolve()` is preferred because it recognizes the special structure of U and solves the *triangular* systems of equations.

We can now return to the problem of sampling from the posterior defined in (14). Using the results of the previous section, we first write

$$\Sigma = (X'X + A) = U'U \tag{19}$$

where U is the upper triangular Cholesky root of $(X'X + A)$. It follows that

$$\begin{aligned} \Sigma^{-1} &= (X'X + A)^{-1} \\ &= (U^{-1})(U^{-1})' \\ &= (IR)(IR)' \end{aligned} \tag{20}$$

and so

$$\begin{aligned}\tilde{\beta} &= (X'X + A)^{-1}(X'y + A\bar{\beta}) \\ &= (IR)(IR)'(X'y + A\bar{\beta}).\end{aligned} \tag{21}$$

The following R code generates one draw from the posterior of $\beta|\sigma^2$.

```
k = length(betabar)
U = chol(crossprod(X)+A)
IR = backsolve(U,diag(k))
btilde = crossprod(t(IR))%*%(crossprod(X,y)+A%*%betabar)
beta = btilde + sqrt(sigmasq)*IR%*%rnorm(k)
```

Rossi et al. (2005) take this a step further. Let $A = U'U$ and define

$$z = \begin{pmatrix} y \\ U\bar{\beta} \end{pmatrix} \quad W = \begin{pmatrix} X \\ U \end{pmatrix} \tag{22}$$

so that $W'W = (X'X + A)$ and $W'z = (X'y + A\bar{\beta})$. The `breg()` function in `bayesm` (Rossi, 2019) uses this modified model structure.

```
k = length(betabar)
RA = chol(A)
W = rbind(X,RA)
z = c(y,RA%*%betabar)
IR = backsolve(chol(crossprod(W)),diag(k))
btilde = crossprod(t(IR))%*%crossprod(W,z)
beta = btilde + sqrt(sigmasq)*IR%*%rnorm(k)
```

References

Rossi, P. E. (2019). *bayesm: Bayesian Inference for Marketing/Micro-Econometrics*, R package version 3.1-4 edition.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons.