Bayesian Linear Models

Sudipto Banerjee\textsuperscript{1} and Andrew O. Finley\textsuperscript{2}

\textsuperscript{1} Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
\textsuperscript{2} Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

March 12, 2009
Ingredients of a linear model include an $n \times 1$ response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ and an $n \times p$ design matrix (e.g. including regressors) $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, assumed to have been observed without error. The linear model:

$$\mathbf{y} = X \beta + \epsilon; \quad \epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

The linear model is the most fundamental of all serious statistical models encompassing:

- **ANOVA**: $\mathbf{y}$ is continuous, $\mathbf{x}_i$’s are categorical
- **REGRESSION**: $\mathbf{y}$ is continuous, $\mathbf{x}_i$’s are continuous
- **ANCOVA**: $\mathbf{y}$ is continuous, some $\mathbf{x}_i$’s are continuous, some categorical.

Unknown parameters include the regression parameters $\beta$ and the variance $\sigma^2$. We assume $X$ is observed without error and all inference is conditional on $X$. 
The classical unbiased estimates of the regression parameter $\beta$ and $\sigma^2$ are

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

$$
\hat{\sigma}^2 = \frac{1}{n - p} (y - X \hat{\beta})^T (y - X \hat{\beta}).
$$

The above estimate of $\beta$ is also a least-squares estimate. The predicted value of $y$ is given by

$$
\hat{y} = X \hat{\beta} = P_X y \text{ where } P_X = X (X^T X)^{-1} X^T.
$$

$P_X$ is called the projector of $X$. It projects any vector to the space spanned by the columns of $X$.

The model residual is estimated as:

$$
\hat{e} = (y - X \hat{\beta})^T (y - X \hat{\beta}) = y^T (I - P_X) y.
$$
For Bayesian analysis, we will need to specify priors for the unknown regression parameters $\beta$ and the variance $\sigma^2$.

Consider independent flat priors on $\beta$ and $\log\sigma^2$:

$$p(\beta) \propto 1; \quad p(\log(\sigma^2)) \propto 1 \quad \text{or equivalently} \quad p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$ 

None of the above two “distributions” are valid probabilities (they do not integrate to any finite number). So why is it that we are even discussing them?

It turns out that even if the priors are improper (that’s what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.
Marginal and conditional distributions

- With a flat prior on $\beta$ we obtain, after some algebra, the conditional posterior distribution:

$$p(\beta | \sigma^2, y) = N(\beta | (X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1}).$$

- The conditional posterior distribution of $\beta$ would have been the desired posterior distribution had $\sigma^2$ been known.

- Since that is not the case, we need to obtain the marginal posterior distribution by integrating out $\sigma^2$ as:

$$p(\beta | y) = \int p(\beta | \sigma^2, y)p(\sigma^2 | y) d\sigma^2$$

- Can we solve this integration using composition sampling? YES: if we can generate samples from $p(\sigma^2 | y)$!
So, we need to find the marginal posterior distribution of $\sigma^2$. With the choice of the flat prior we obtain:

$$p(\sigma^2 | \mathbf{y}) \propto \frac{1}{(\sigma^2)^{(n-p)/2+1}} \exp \left( -\frac{(n-p)s^2}{2\sigma^2} \right)$$

$$= IG \left( \sigma^2 \mid \frac{n-p}{2}, \frac{(n-p)s^2}{2} \right),$$

where $s^2 = \hat{\sigma}^2 = \frac{1}{n-p} \mathbf{y}^T (I - P_X) \mathbf{y}$.

This is known as an *inverted Gamma* distribution (also called a *scaled chi-square* distribution)

$IG(\sigma^2 \mid (n-p)/2, (n-p)s^2/2)$.

In other words: $[(n-p)s^2/\sigma^2 \mid \mathbf{y}] \sim \chi^2_{n-p}$ (with $n-p$ degrees of freedom). A striking similarity with the classical result: The distribution of $\hat{\sigma}^2$ is also characterized as $(n-p)s^2/\sigma^2$ following a chi-square distribution.
Now we are ready to carry out composition sampling from $p(\beta, \sigma^2 \mid y)$ as follows:

- Draw $M$ samples from $p(\sigma^2 \mid y)$:
  \[
  \sigma^2(j) \sim IG \left( \frac{n-p}{2}, \frac{(n-p)s^2}{2}(n-p) \right), \ j = 1, \ldots M
  \]

- For $j = 1, \ldots, M$, draw from $p(\beta \mid \sigma^2(j), y)$:
  \[
  \beta(j) \sim N \left( (X^T X)^{-1} X^T y, \sigma^2(j) (X^T X)^{-1} \right)
  \]

The resulting samples $\{\beta(j), \sigma^2(j)\}_{j=1}^M$ represent $M$ samples from $p(\beta, \sigma^2 \mid y)$.

$\{\beta(j)\}_{j=1}^M$ are samples from the marginal posterior distribution $p(\beta \mid y)$. This is a multivariate $t$ density:

\[
p(\beta \mid y) = \frac{\Gamma(n/2)}{(\pi(n-p))^{p/2}\Gamma((n-p)/2)s^2(X^T X)^{-1}} \left[ 1 + \frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{(n-p)s^2} \right]^{-n/2}.
\]
The marginal distribution of each individual regression parameter $\beta_j$ is a non-central univariate $t_{n-p}$ distribution. In fact,

$$
\frac{\beta_j - \hat{\beta}_j}{s \sqrt{(X^T X)^{-1}_{jj}}} \sim t_{n-p}.
$$

The 95\% credible intervals for each $\beta_j$ are constructed from the quantiles of the $t$-distribution. The credible intervals exactly coincide with the 95\% classical confidence intervals, but the interpretation is direct: the probability of $\beta_j$ falling in that interval, given the observed data, is 0.95.

Note: an intercept only linear model reduces to the simple univariate $N(\bar{y} | \mu, \sigma^2/n)$ likelihood, for which the marginal posterior of $\mu$ is:

$$
\frac{\mu - \bar{y}}{s / \sqrt{n}} \sim t_{n-1}.
$$
Suppose we have observed the new predictors \( \tilde{X} \), and we wish to predict the outcome \( \tilde{y} \). We specify \( p(\tilde{y}, y | \theta) \) to be a normal distribution:

\[
\begin{pmatrix} y \\ \tilde{y} \end{pmatrix} \sim N \left( \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta, \sigma^2 I \right)
\]

Note \( p(\tilde{y} | y, \beta, \sigma^2) = p(\tilde{y} | \beta, \sigma^2) = N(\tilde{y} | \tilde{X} \beta, \sigma^2 I) \).

The posterior predictive distribution:

\[
p(\tilde{y} | y) = \int p(\tilde{y} | y, \beta, \sigma^2)p(\beta, \sigma^2 | y)d\beta d\sigma^2
\]

\[
= \int p(\tilde{y} | \beta, \sigma^2)p(\beta, \sigma^2 | y)d\beta d\sigma^2.
\]

By now we are comfortable evaluating such integrals:

- First obtain: \( (\beta^{(j)}, \sigma^2^{(j)}) \sim p(\beta, \sigma^2 | y) \), \( j = 1, \ldots, M \)
- Next draw: \( \tilde{y}^{(j)} \sim N(\tilde{X} \beta^{(j)}, \sigma^2^{(j)} I) \).
Suppose that $\theta = (\theta_1, \theta_2)$ and we seek the posterior distribution $p(\theta_1, \theta_2 \mid y)$.

For many interesting hierarchical models, we have access to full conditional distributions $p(\theta_1 \mid \theta_2, y)$ and $p(\theta_1 \mid \theta_2, y)$.

The Gibbs sampler proposes the following sampling scheme. Set starting values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$. For $j = 1, \ldots, M$

- Draw $\theta_1^{(j)} \sim p(\theta_1 \mid \theta_2^{(j-1)}, y)$
- Draw $\theta_2^{(j)} \sim p(\theta_2 \mid \theta_1^{(j)}, y)$

This constructs a Markov Chain and, after an initial “burn-in” period when the chains are trying to find their way, the above algorithm guarantees that

$\{\theta_1^{(j)}, \theta_2^{(j)}\}_{j=M_0+1}^{M}$ will be samples from $p(\theta_1, \theta_2 \mid y)$, where $M_0$ is the burn-in period.
More generally, if $\theta = (\theta_1, \ldots, \theta_p)$ are the parameters in our model, we provide a set of initial values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ and then performs the $j$-th iteration, say for $j = 1, \ldots, M$, by updating successively from the full conditional distributions:

- $\theta_1^{(j)} \sim p(\theta_1 | \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, y)$
- $\theta_2^{(j)} \sim p(\theta_2 | \theta_1^{(j)}, \theta_3^{(j)}, \ldots, \theta_p^{(j-1)}, y)$
  
  \ldots

  (the generic $k^{th}$ element)

- $\theta_k^{(j)} \sim p(\theta_k | \theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j)}, \ldots, \theta_p^{(j-1)}, y)$
  
  \ldots

- $\theta_p^{(j)} \sim p(\theta_p | \theta_1^{(j)}, \ldots, \theta_{p-1}^{(j)}, y)$
Example: Consider the linear model. Suppose we set 
\[ p(\sigma^2) = IG(\sigma^2 | a, b) \text{ and } p(\beta) \propto 1. \]

The full conditional distributions are:

\[ p(\beta | y, \sigma^2) = N(\beta | (X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1}) \]
\[ p(\sigma^2 | y, \beta) = IG \left( \sigma^2 | a + n/2, b + \frac{1}{2} (y - X \beta)^T (y - X \beta) \right). \]

Thus, the Gibbs sampler will initialize \((\beta^{(0)}, \sigma^{2(0)})\) and draw, for \(j = 1, \ldots, M\):

- Draw \(\beta^{(j)} \sim N((X^T X)^{-1} X^T y, \sigma^{2(j-1)} (X^T X)^{-1})\)
- Draw \(\sigma^{2(j)} \sim IG \left( a + n/2, b + \frac{1}{2} (y - X \beta^{(j)})^T (y - X \beta^{(j)}) \right)\)
In principle, the Gibbs sampler will work for extremely complex hierarchical models. The only issue is sampling from the full conditionals. They may not be amenable to easy sampling – when these are not in closed form. A more general and extremely powerful - and often easier to code - algorithm is the Metropolis-Hastings (MH) algorithm.

This algorithm also constructs a Markov Chain, but does not necessarily care about full conditionals.
The Metropolis-Hastings algorithm: Start with an initial value for $\theta = \theta^{(0)}$. Select a candidate or proposal distribution from which to propose a value of $\theta$ at the $j$-th iteration: $\theta^{(j)} \sim q(\theta^{(j-1)}, \nu)$. For example, $q(\theta^{(j-1)}, \nu) = N(\theta^{(j-1)}, \nu)$ with $\nu$ fixed.

Compute

$$r = \frac{p(\theta^* \mid y)q(\theta^{(j-1)} \mid \theta^*, \nu)}{p(\theta^{(j-1)} \mid y)q(\theta^* \mid \theta^{(j-1)}, \nu)}$$

If $r \geq 1$ then set $\theta^{(j)} = \theta^*$. If $r \leq 1$ then draw $U \sim (0, 1)$. If $U \leq r$ then $\theta^{(j)} = \theta^*$. Otherwise, $\theta^{(j)} = \theta^{(j-1)}$.

Repeat for $j = 1, \ldots, M$. This yields $\theta^{(1)}, \ldots, \theta^{(M)}$, which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio $r$ of the sampler through the iterations. Rough recommendations: for vector updates $r \approx 20\%$, for scalar updates $r \approx 40\%$. This can be controlled by “tuning” $\nu$.

Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.
Example: For the linear model, our parameters are \((\beta, \sigma^2)\). We write \(\theta = (\beta, \log(\sigma^2))\) and, at the \(j\)-th iteration, propose \(\theta^* \sim N(\theta^{(j-1)}, \Sigma)\). The log transformation on \(\sigma^2\) ensures that all components of \(\theta\) have support on the entire real line and can have meaningful proposed values from the multivariate normal. But we need to transform our prior to \(p(\beta, \log(\sigma^2))\).

Let \(z = \log(\sigma^2)\) and assume \(p(\beta, z) = p(\beta)p(z)\). Let us derive \(p(z)\). **REMEMBER:** we need to adjust for the jacobian. Then \(p(z) = p(\sigma^2) |d\sigma^2/dz| = p(e^z)e^z\). The jacobian here is \(e^z = \sigma^2\).

Let \(p(\beta) = 1\) and an \(p(\sigma^2) = IG(\sigma^2 \mid a, b)\). Then log-posterior is:

\[-(a + n/2 + 1)z + z - \frac{1}{e^z} \{b + \frac{1}{2}(Y - X\beta)^T(Y - X\beta)\}\.\]

A symmetric proposal distribution, say \(q(\theta^* \mid \theta^{(j-1)}, \Sigma) = N(\theta^{(j-1)}, \Sigma)\), cancels out in \(r\). In practice it is better to compute \(\log(r)\): \(\log(r) = \log(p(\theta^* \mid y) - \log(p(\theta^{(j-1)} \mid y))\). For the proposal, \(N(\theta^{(j-1)}, \Sigma)\), \(\Sigma\) is a \(d \times d\) variance-covariance matrix, and \(d = \text{dim}(\theta) = p + 1\).

If \(\log r \geq 0\) then set \(\theta^{(j)} = \theta^*\). If \(\log r \leq 0\) then draw \(U \sim (0, 1)\). If \(U \leq r\) (or \(\log U \leq \log r\) then \(\theta^{(j)} = \theta^*\). Otherwise, \(\theta^{(j)} = \theta^{(j-1)}\).

Repeat the above procedure for \(j = 1, \ldots, M\) to obtain samples \(\theta^{(1)}, \ldots, \theta^{(M)}\).