Bayesian Linear Regression

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

\textsuperscript{1} Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

\textsuperscript{2} Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

October 30, 2013
Linear regression models: a Bayesian perspective

- Linear regression is, perhaps, *the* most widely used statistical modelling tool.

- It addresses the following question: How does a quantity of primary interest, $y$, vary as (depend upon) another quantity, or set of quantities, $x$?

- The quantity $y$ is called the *response* or *outcome variable*. Some people simply refer to it as the *dependent variable*.

- The variable(s) $x$ are called *explanatory variables*, *covariates* or simply *independent variables*.

- In general, we are interested in the conditional distribution of $y$, given $x$, parametrized as $p(y \mid \theta, x)$. 
Typically, we have a set of *units* or *experimental subjects* 
\[ i = 1, 2, \ldots, n. \]

For each of these units we have measured an outcome \( y_i \) and a set of explanatory variables \( \mathbf{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ip}) \).

The first element of \( \mathbf{x}_i' \) is often taken as 1 to signify the presence of an “intercept”.

We collect the outcome and explanatory variables into an \( n \times 1 \) vector and an \( n \times (p + 1) \) matrix:

\[
\begin{align*}
\mathbf{y} &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \\
\mathbf{x} &= \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.
\end{align*}
\]
The linear model is the most fundamental of all serious statistical models underpinning:

- **ANOVA**: $y_i$ is continuous, $x_{ij}$’s are *all* categorical
- **REGRESSION**: $y_i$ is continuous, $x_{ij}$’s are continuous
- **ANCOVA**: $y_i$ is continuous, $x_{ij}$’s are continuous for some $j$ and categorical for others.
The Bayesian or hierarchical linear model is given by:

\[ y_i \mid \mu_i, \sigma^2, X \overset{\text{ind}}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n; \]
\[ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = x_i' \beta; \quad \beta = (\beta_0, \beta_1, \ldots, \beta_p); \]
\[ \beta, \sigma^2 \mid X \sim p(\beta, \sigma^2 \mid X). \]

Unknown parameters include the regression parameters and the variance, i.e. \( \theta = \{\beta, \sigma^2\} \).

\[ p(\beta, \sigma^2 \mid X) \equiv p(\theta \mid X) \] is the joint prior on the parameters.

We assume \( X \) is observed without error and all inference is conditional on \( X \).

We suppress dependence on \( X \) in subsequent notation.
Specifying \( p(\beta, \sigma^2) \) completes the hierarchical model.

All inference proceeds from \( p(\beta, \sigma^2 | y) \)

With no prior information, we specify

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

The above is NOT a probability density (they do not integrate to any finite number). So why is it that we are even discussing them?

Even if the priors are improper, as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.
Computing the posterior distribution

- Strategy: Factor the joint posterior distribution for $\beta$ and $\sigma^2$ as:
  \[ p(\beta, \sigma^2 \mid y) = p(\beta \mid \sigma^2, y) \times p(\sigma^2 \mid y). \]

- The *conditional posterior* distribution of $\beta$, given $\sigma^2$:
  \[ \beta \mid \sigma^2, y \sim N(\hat{\beta}, \sigma^2 V_\beta), \]
  where, using some algebra, one finds
  \[ \hat{\beta} = (X'X)^{-1}X'y \quad \text{and} \quad V_\beta = (X'X)^{-1}. \]
The *marginal posterior* distribution of $\sigma^2$: Let $k = (p + 1)$ be the number of columns of $X$.

$$
\sigma^2 \mid y \sim IG \left( \frac{n - k}{2}, \frac{(n - k)s^2}{2} \right),
$$

where

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

is the classical unbiased estimate of $\sigma^2$ in the linear regression model.

The *marginal posterior* distribution $p(\beta \mid y)$, averaging over $\sigma^2$, is *multivariate t* with $n - k$ degrees of freedom. But we rarely use this fact in practice.

Instead, we *sample* from the posterior distribution.
Algorithm for sampling from the posterior distribution

- We draw samples from \( p(\beta, \sigma^2 \mid y) \) by executing the following steps:

  - Step 1: Compute \( \hat{\beta} \) and \( V_\beta \).
  
  - Step 2: Compute \( s^2 \).

  - Step 3: Draw \( M \) samples from \( p(\sigma^2 \mid y) \):
    \[
    \sigma^2(j) \sim IG \left( \frac{n - k}{2}, \frac{(n - k)s^2}{2} \right), \quad j = 1, \ldots, M
    \]

  - Step 4: For \( j = 1, \ldots, M \), draw \( \beta^{(j)} \) from \( p(\beta \mid \sigma^2(j), y) \):
    \[
    \beta^{(j)} \sim N \left( \hat{\beta}, \sigma^2(j) V_\beta \right)
    \]
The marginal distribution of each individual regression parameter $\beta_j$ is a non-central univariate $t_{n-p}$ distribution. In fact,

$$\frac{\beta_j - \hat{\beta}_j}{s \sqrt{V_{\beta;jj}}} \sim t_{n-p}.$$ 

The 95% credible interval for each $\beta_j$ is constructed from the quantiles of the $t$-distribution. This exactly coincides 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 $\mathcal{N}(\bar{y} \mid \mu, \sigma^2/n)$ 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 $\tilde{y}$.

If $\beta$ and $\sigma^2$ were known exactly, the random vector $\tilde{y}$ would follow $N(\tilde{X}\beta, \sigma^2 I)$.

But we do not know model parameters, which contribute to the uncertainty in predictions.

Predictions are carried out by sampling from the posterior predictive distribution, $p(\tilde{y} \mid y)$

1. Draw $\{\beta^{(j)}, \sigma^2^{(j)}\} \sim p(\beta, \sigma^2 \mid y)$, $j = 1, 2, \ldots, M$

2. Draw $\tilde{y}^{(j)} \sim N(\tilde{X}\beta^{(j)}, \sigma^2^{(j)} I)$, $j = 1, 2, \ldots, M$. 
Predictive Mean and Variance (conditional upon $\sigma^2$):

$$E(\tilde{y} \mid \sigma^2, y) = \tilde{X}\hat{\beta}$$

$$\text{var}(\tilde{y} \mid \sigma^2, y) = (I + \tilde{X}V\tilde{X}')\sigma^2.$$ 

The posterior predictive distribution, $p(\tilde{y} \mid y)$, is a multivariate $t$ distribution, $t_{n-p}(\tilde{X}\hat{\beta}, s^2(I + \tilde{X}V\tilde{X}')).$
Incorporating prior information

\[ y_i \mid \mu_i, \sigma^2 \sim \text{ind} \ N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n; \]
\[ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \beta; \quad \beta = (\beta_0, \beta_1, \ldots, \beta_p); \]
\[ \beta \mid \sigma^2 \sim N(\beta_0, \sigma^2 R_\beta); \quad \sigma^2 \sim IG(a_\sigma, b_\sigma), \]

where \( R_\beta \) is a \textit{fixed} correlation matrix. Alternatively,

\[ y_i \mid \mu_i, \sigma^2 \sim \text{ind} \ N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n; \]
\[ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \beta; \quad \beta = (\beta_0, \beta_1, \ldots, \beta_p); \]
\[ \beta \mid \Sigma_\beta \sim N(\beta_0, \Sigma_\beta); \quad \Sigma_\beta \sim IW(\nu, \mathbf{S}); \quad \sigma^2 \sim IG(a_\sigma, b_\sigma), \]

where \( \Sigma_\beta \) is a \textit{random} covariance matrix.
The Gibbs sampler: 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:

\[
\begin{align*}
\theta_1^{(j)} &\sim p(\theta_1^{(j)} | \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, y) \\
\theta_2^{(j)} &\sim p(\theta_2 | \theta_1^{(j)}, \theta_3^{(j)}, \ldots, \theta_p^{(j-1)}, y) \\
\vdots & \hspace{100pt} \text{(the generic $k^{th}$ element)} \\
\theta_k^{(j)} &\sim p(\theta_k | \theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j)}, \ldots, \theta_p^{(j-1)}, y) \\
\vdots & \hspace{100pt} \text{...} \\
\theta_p^{(j)} &\sim p(\theta_p | \theta_1^{(j)}, \ldots, \theta_{p-1}^{(j)}, y)
\end{align*}
\]
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.

Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.