Bayesian Linear Models

Sudipto Banerjee

September 03–05, 2017

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles

Linear Regression

- Linear regression is, perhaps, *the* most widely used statistical modeling 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 $x_i^\top = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$.

- The first element of $x_i^\top$ 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:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix}.$$

- 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.

Conjugate Bayesian Linear Regression

- A conjugate Bayesian linear model is given by:

$$y_i \mid \mu_i, \sigma^2, X \overset{ind}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n \; ;$$
$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = x_i^\top \beta \; ; \quad \beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top \; ;$$
$$\beta \mid \sigma^2 \sim N(\mu_\beta, \sigma^2 V_\beta) \; ; \quad \sigma^2 \sim IG(a, b) \; .$$

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

- We assume X is observed without error and all inference is conditional on X.

- The above model is often written it terms of the posterior density $p(\theta \mid y) \propto p(\theta, y)$:

$$IG(\sigma^2 \mid a, b) \times N(\beta \mid \mu_\beta, \sigma^2 V_\beta) \times \prod_{i=1}^{n} N(y_i \mid x_i^\top \beta, \sigma^2) \; .$$

Conjugate Bayesian (General) Linear Regression

- A more general conjugate Bayesian linear model is given by:

$$
\begin{aligned}
&y \mid \beta, \sigma^2, X \sim N(X\beta, \sigma^2 V_y) \\
&\beta \mid \sigma^2 \sim N(\mu_\beta, \sigma^2 V_\beta) \; ; \\
&\sigma^2 \sim IG(a, b) \; .
\end{aligned}
$$

- V_y, V_β and μ_β are assumed fixed.
- Unknown parameters include the regression parameters and the variance, i.e. $\theta = \{\beta, \sigma^2\}$.
- We assume X is observed without error and all inference is conditional on X.
- The posterior density $p(\theta \mid y) \propto p(\theta, y)$:

$$
IG(\sigma^2 \mid a, b) \times N(\beta \mid \mu_\beta, \sigma^2 V_\beta) \times N(y \mid X\beta, \sigma^2 V_y)
$$

- The model on the previous slide is a special case with $V_y = I_n$ ($n \times n$ identity matrix).

Conjugate Bayesian (General) Linear Regression

- The joint posterior density can be written as

$$p(\beta, \sigma^2 \mid y) \propto \underbrace{IG(\sigma^2 \mid a^*, b^*)}_{p(\sigma^2 \mid y)} \times \underbrace{N\left(\beta \mid Mm, \sigma^2 M\right)}_{p(\beta \mid \sigma^2, y)} ,$$

where

$$a^* = a + \frac{n}{2} ; \quad b^* = b + \frac{1}{2}\left(\mu_\beta^\top V_\beta^{-1}\mu_\beta + y^\top y - m^\top Mm\right) ;$$
$$m = V_\beta^{-1}\mu_\beta + X^\top V_y^{-1} y ; \quad M^{-1} = V_\beta^{-1} + X^\top V_y^{-1} X .$$

- Exact posterior sampling from $p(\beta, \sigma^2 \mid y)$ will automatically yield samples from $p(\beta \mid y)$ and $p(\sigma^2 \mid y)$.
- For each $i = 1, 2, \ldots, N$ do the following:
  1. Draw $\sigma^2_{(i)} \sim IG(a^*, b^*)$
  2. Draw $\beta_{(i)} \sim N\left(Mm, \sigma^2_{(i)} M\right)$
- The above is sometimes referred to as *composition sampling*.

Exact sampling from joint posterior distributions

- Suppose we wish to draw samples from a joint posterior:

$$p(\theta_1, \theta_2 \mid y) = p(\theta_1 \mid y) \times p(\theta_2 \mid \theta_1, y) \ .$$

- In conjugate models, it is often easy to draw samples from $p(\theta_1 \mid y)$ and from $p(\theta_2 \mid \theta_1, y)$.
- We can draw M samples from $p(\theta_1, \theta_2 \mid y)$ as follows.
- For each $i = 1, 2, \ldots, N$ do the following:
  1. Draw $\theta_{1(i)} \sim p(\theta_1 \mid y)$
  2. Draw $\theta_{2(i)} \sim p(\theta_2 \mid \theta_1, y)$
- Remarkably, the $\theta_{2(i)}$'s drawn above have marginal distribution $p(\theta_2 \mid y)$ because:

$$\begin{aligned} P(\theta_2 \leq u \mid y) &= \mathsf{E}_{\theta_2 \mid y}\left[1(\theta_2 \leq u)\right] = \mathsf{E}_{\theta_1 \mid y}\left\{\mathsf{E}_{\theta_2 \mid \theta_1, y}\left[1(\theta_2 \leq u)\right]\right\} \\ &\approx \frac{1}{N}\sum_{i=1}^{N} \mathsf{E}_{\theta_2 \mid \theta_{1(i)}, y}\left[1(\theta_2 \leq u)\right] \approx \frac{1}{N}\sum_{i=1}^{N} 1(\theta_{2(i)} \leq u) \ . \end{aligned}$$

- "Automatic Marginalization:" We draw samples $p(\theta_1, \theta_2 \mid y)$ and automatically get samples from $p(\theta_1 \mid y)$ and $p(\theta_2 \mid y)$.

Bayesian predictions from linear regression

- Let $\tilde{y}$ denote an $m \times 1$ vector of outcomes we seek to predict based upon predictors $\tilde{X}$.
- We seek the posterior predictive density:

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta, y) p(\theta \mid y) \mathrm{d}\theta \ .$$

- Posterior predictive inference: sample from $p(\tilde{y} \mid y)$.
- For each $i = 1, 2, \ldots, N$ do the following:
    1. Draw $\theta_{(i)} \sim p(\theta \mid y)$
    2. Draw $\tilde{y}_{(i)} \sim p(\tilde{y} \mid \theta_{(i)}, y)$

Bayesian predictions from linear regression (contd.)

- For legitimate probabilistic predictions (forecasting), the conditional distribution $p(\tilde{y} \,|\, \theta, y)$ must be well-defined.
- For example, consider the case with $V_y = I_n$. Specify the linear model:

$$\begin{bmatrix} y \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta + \begin{bmatrix} \epsilon \\ \tilde{\epsilon} \end{bmatrix} \; ; \quad \begin{bmatrix} \epsilon \\ \tilde{\epsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} I_n & O \\ O & I_m \end{bmatrix} \right) \; .$$

- Easy to derive the conditional density:

$$p(\tilde{y} \,|\, \theta, y) = p(\tilde{y} \,|\, \theta) = N(\tilde{y} \,|\, \tilde{X}\beta, \sigma^2 I_m)$$

- Posterior predictive density:

$$p(\tilde{y} \,|\, y) = \int N(\tilde{y} \,|\, \tilde{X}\beta, \sigma^2 I_m) p(\beta, \sigma^2 \,|\, y) \mathrm{d}\beta \mathrm{d}\sigma^2 \; .$$

- For each $i = 1, 2, \ldots, N$ do the following:
    1. Draw $\{\beta_{(i)}, \sigma^2_{(i)}\} \sim p(\beta, \sigma^2 \,|\, y)$
    2. Draw $\tilde{y}_{(i)} \sim N(\tilde{X}\beta_{(i)}, \sigma^2_{(i)} I_m)$

Bayesian predictions from general linear regression

- For example, consider the case with general V_y. Specify:

$$\begin{bmatrix} y \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta + \begin{bmatrix} \epsilon \\ \tilde{\epsilon} \end{bmatrix} \; ; \quad \begin{bmatrix} \epsilon \\ \tilde{\epsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} V_y & V_{y\tilde{y}} \\ V_{y\tilde{y}}^{\top} & V_{\tilde{y}} \end{bmatrix} \right) .$$

- Derive the conditional density $p(\tilde{y} \mid \theta, y) = N\left(\tilde{y} \mid \mu_{\tilde{y}|y}, \sigma^2 V_{\tilde{y}|y}\right)$:

$$\mu_{\tilde{y}|y} = \tilde{X}\beta + V_{y\tilde{y}}^{\top} V_y^{-1}(y - X\beta) \; ; \quad V_{\tilde{y}|y} = V_{\tilde{y}} - V_{y\tilde{y}}^{\top} V_y^{-1} V_{y\tilde{y}} .$$

- Posterior predictive density:

$$p(\tilde{y} \mid y) = \int N\left(\tilde{y} \mid \mu_{\tilde{y}|y}, \sigma^2 V_{\tilde{y}|y}\right) p(\beta, \sigma^2 \mid y) \mathrm{d}\beta \mathrm{d}\sigma^2 .$$

- For each $i = 1, 2, \ldots, N$ do the following:
  1. Draw $\{\beta_{(i)}, \sigma^2_{(i)}\} \sim p(\beta, \sigma^2 \mid y)$
  2. Compute $\mu_{\tilde{y}|y}$ using $\beta_{(i)}$ and draw $\tilde{y}_{(i)} \sim N(\mu_{\tilde{y}|y}, \sigma^2_{(i)} V_{\tilde{y}})$

Application to Bayesian Geostatistics

- Consider the spatial regression model

$$y(s_i) = x^\top(s_i)\beta + w(s_i) + \epsilon(s_i)\ ,$$

  where $w(s_i)$'s are spatial random effects and $\epsilon(s_i)$'s are unstructured errors ("white noise").
- $w = (w(s_1), w(s_2), \ldots, w(s_n))^\top \sim N(0, \sigma^2 R(\phi))$
- $\epsilon = (\epsilon(s_1), \epsilon(s_2), \ldots, \epsilon(s_n))^\top \sim N(0, \tau^2 I_n)$
- Integrating out random effects leads to a Bayesian model:

$$IG(\sigma^2 \,|\, a, b) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times N(y \,|\, X\beta, \sigma^2 V_y)$$

  where $V_y = R(\phi) + \alpha I_n$ and $\alpha = \tau^2/\sigma^2$.
- Fixing ϕ and α (e.g., from variogram or other EDA) yields a conjugate Bayesian model.
- Exact posterior sampling is easily achieved as before.

Inference on spatial random effects

- Rewrite the model in terms of w as:

$$IG(\sigma^2 \,|\, a, b) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times N(w \,|\, 0, \sigma^2 R(\phi)) \\ \times N(y \,|\, X\beta + w, \tau^2 I_n) \; .$$

- Posterior distribution of spatial random effects w:

$$p(w \,|\, y) = \int N(w \,|\, Mm, \sigma^2 M) \times p(\beta, \sigma^2 \,|\, y) \mathrm{d}\beta \mathrm{d}\sigma^2 \; ,$$

where $m = (1/\alpha)(y - X\beta)$ and $M^{-1} = R^{-1}(\phi) + (1/\alpha) I_n$.

- For each $i = 1, 2, \ldots, N$ do the following:
  1. Draw $\{\beta_{(i)}, \sigma^2_{(i)}\} \sim p(\beta, \sigma^2 \,|\, y)$
  2. Compute m from $\beta_{(i)}$ and draw $w_{(i)} \sim N(Mm, \sigma^2_{(i)} M)$

Inference on the process

- Posterior distribution of $w(s_0)$ at new location s_0:

$$p(w(s_0)\,|\,y) = \int N(w(s_0)\,|\,\mu_{w(s_0)|w}, \sigma^2_{w(s_0)|w}) \times p(\sigma^2, w\,|\,y)\mathrm{d}\sigma^2\mathrm{d}w \ ,$$

  where

$$\begin{aligned} \mu_{w(s_0)|w} &= r^\top(s_0;\phi)R^{-1}(\phi)w \ ; \\ \sigma^2_{w(s_0)|w} &= \sigma^2\{1 - r^\top(s_0;\phi)R^{-1}(\phi)r(s_0,\phi)\} \end{aligned}$$

- For each $i = 1, 2, \ldots, N$ do the following:
  1. Compute $\mu_{w(s_0)|w}$ and $\sigma^2_{w(s_0)|w}$ from $w_{(i)}$ and $\sigma^2_{(i)}$.
  2. Draw $w_{(i)}(s_0) \sim N(\mu_{w(s_0)|w}, \sigma^2_{w(s_0)|w})$.

Bayesian "kriging" or prediction

- Posterior predictive distribution at new location s_0 is $p(y(s_0) \mid y)$:

$$\int N(y(s_0) \mid x^\top(s_0)\beta + w(s_0), \alpha\sigma^2) \times p(\beta, \sigma^2, w \mid y) \mathrm{d}\beta \mathrm{d}\sigma^2 \mathrm{d}w \ ,$$

- For each $i = 1, 2, \ldots, N$ do the following:
    1. Draw $y_{(i)}(s_0) \sim N(x^\top(s_0)\beta_{(i)} + w_{(i)}(s_0), \alpha\sigma^2_{(i)})$.

Non-conjugate models: The Gibbs Sampler

- Let $\theta = (\theta_1, \dots, \theta_p)$ be the parameters in our model.
- $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_p^{(0)})$
- For $j = 1, \dots, M$, update successively using the *full conditional* distributions:

$$
\begin{array}{l}
\theta_1^{(j)} \sim p(\theta_1^{(j)} \,|\, \theta_2^{(j-1)}, \dots, \theta_p^{(j-1)}, y) \\
\theta_2^{(j)} \sim p(\theta_2 \,|\, \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, y) \\
\vdots \\
\text{(the generic } k^{th} \text{ element)} \\
\theta_k^{(j)} \sim p(\theta_k | \theta_1^{(j)}, \dots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)}, \dots, \theta_p^{(j-1)}, y) \\
\vdots \\
\theta_p^{(j)} \sim p(\theta_p \,|\, \theta_1^{(j)}, \dots, \theta_{p-1}^{(j)}, y)
\end{array}
$$

- 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.

The Metropolis-Hastings Algorithm

- The Metropolis-Hastings algorithm: Start with a initial value for $\theta = \theta^{(0)}$. Select a *candidate* or *proposal* distribution from which to propose a value of θ 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 ν 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" ν.
- 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 (β, σ^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 σ^2 ensures that all components of θ 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 $q(\theta^* | \theta^{(j-1)}, \Sigma) = N(\theta^{(j-1)}, \Sigma)$, 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, $N(\theta^{(j-1)}, \Sigma)$, Σ is a $d \times d$ variance-covariance matrix, and $d = \dim(\theta) = p + 1$.
- 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 $\theta^{(1)}, \ldots, \theta^{(M)}$.