Bayesian Linear Regression

Sudipto Banerjee

[1] Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

September 15, 2010

- 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, $\mathbf{x}$?
- The quantity y is called the *response* or *outcome variable*. Some people simply refer to it as the *dependent variable*.
- The variable(s) $\mathbf{x}$ are called *explanatory variables*, *covariates* or simply *independent variables*.
- In general, we are interested in the conditional distribution of y, given $\mathbf{x}$, parametrized as $p(y \mid \boldsymbol{\theta}, \mathbf{x})$.

- 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:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.$$

- 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.

- The Bayesian or hierarchical linear model is given by:

$$
\begin{aligned}
&y_i \mid \mu_i, \sigma^2, \mathbf{X} \overset{ind}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n;\\
&\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \boldsymbol{\beta}; \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p);\\
&\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X} \sim p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X})\ .
\end{aligned}
$$

- Unknown parameters include the regression parameters and the variance, i.e. $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$.
- $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}) \equiv p(\boldsymbol{\theta} \mid \mathbf{X})$ 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 $p(\boldsymbol{\beta}, \sigma^2)$ completes the hierarchical model.
- All inference proceeds from $p(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y})$
- With no prior information, we specify

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

- 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 $\boldsymbol{\beta}$ and σ^2 as:
$$p(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y}) = p(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}) \times p(\sigma^2 \,|\, \mathbf{y}) \; .$$
- The *conditional posterior* distribution of $\boldsymbol{\beta}$, given σ^2:
$$\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2 \mathbf{V}_{\beta}),$$
where, using some algebra, one finds
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad \mathbf{V}_{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \; .$$

- The *marginal posterior* distribution of σ^2: Let $k = (p+1)$ be the number of columns of $\mathbf{X}$.

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

  where

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

  is the classical unbiased estimate of σ^2 in the linear regression model.

- The *marginal posterior* distribution $p(\boldsymbol{\beta} \,|\, \mathbf{y})$, averaging over σ^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 $p(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y})$ by executing the following steps:
- Step 1: Compute $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}_\beta$.
- Step 2: Compute s^2.
- Step 3: Draw M samples from $p(\sigma^2 \,|\, \mathbf{y})$:

$$\sigma^{2(j)} \sim IG\left(\frac{n-k}{2}, \frac{(n-k)s^2}{2}\right),\ j = 1, \ldots M$$

- Step 4: For $j = 1, \ldots, M$, draw $\boldsymbol{\beta}^{(j)}$ from $p(\boldsymbol{\beta} \,|\, \sigma^{2(j)}, \mathbf{y})$:

$$\boldsymbol{\beta}^{(j)} \sim N\left(\hat{\boldsymbol{\beta}},\ \sigma^{2(j)}\mathbf{V}_\beta\right)$$

- The marginal distribution of each individual regression parameter β_j is a non-central univariate t_{n-p} distribution. In fact,

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

  The 95% credible interval for each β_j is constructed from the quantiles of the t-distribution. This exactly coincides with the 95% classical confidence intervals, but the intepretation is direct: the probability of β_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 μ is:

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}.$$

- Suppose we have observed the new predictors $\tilde{\mathbf{X}}$, and we wish to predict the outcome $\tilde{\mathbf{y}}$.
- If β and σ^2 were known exactly, the random vector $\tilde{\mathbf{y}}$ would follow $N(\tilde{\mathbf{X}}\beta, \sigma^2\mathbf{I})$.
- 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, $p(\tilde{\mathbf{y}} \,|\, \mathbf{y})$
  1. Draw $\{\boldsymbol{\beta}^{(j)}, \sigma^{2(j)}\} \sim p(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y}),\ j = 1, 2, \ldots, M$
  2. Draw $\tilde{\mathbf{y}}^{(j)} \sim N(\tilde{\mathbf{X}}\boldsymbol{\beta}^{(j)}, \sigma^{2(j)} I),\ j = 1, 2, \ldots, M$.

- Predictive Mean and Variance (conditional upon σ^2):

$$\mathsf{E}(\tilde{\mathbf{y}} \,|\, \sigma^2, \mathbf{y}) = \tilde{\mathbf{X}}\hat{\beta}$$
$$\mathrm{var}(\tilde{\mathbf{y}} \,|\, \sigma^2, \mathbf{y}) = (\mathbf{I} + \tilde{\mathbf{X}}\mathbf{V}_\beta\tilde{\mathbf{X}}')\sigma^2.$$

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

Incorporating prior information

$$\begin{aligned}
&y_i \,|\, \mu_i, \sigma^2 \stackrel{ind}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n;\\
&\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \boldsymbol{\beta}; \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p);\\
&\boldsymbol{\beta} \,|\, \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2 \mathbf{R}_\beta); \quad \sigma^2 \sim IG(a_\sigma, b_\sigma)\ ,
\end{aligned}$$

where $\mathbf{R}_\beta$ is a *fixed* correlation matrix. Alternatively,

$$\begin{aligned}
&y_i \,|\, \mu_i, \sigma^2 \stackrel{ind}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n;\\
&\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_p = \mathbf{x}_i' \boldsymbol{\beta}; \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p);\\
&\boldsymbol{\beta} \,|\, \Sigma_\beta \sim N(\boldsymbol{\beta}_0, \Sigma_\beta); \quad \Sigma_\beta \sim IW(\nu, \mathbf{S}); \quad \sigma^2 \sim IG(a_\sigma, b_\sigma)\ ,
\end{aligned}$$

where Σ_β is a *random* covariance matrix.

- The Gibbs sampler: If $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$ are the parameters in our model, we provide a set of initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_p^{(0)})$ and then performs the j-th iteration, say for $j = 1, \ldots, M$, by updating successively from the *full conditional* distributions:

$$\boldsymbol{\theta}_1^{(j)} \sim p(\boldsymbol{\theta}_1^{(j)} \,|\, \boldsymbol{\theta}_2^{(j-1)}, \ldots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

$$\boldsymbol{\theta}_2^{(j)} \sim p(\boldsymbol{\theta}_2 \,|\, \boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

$$\vdots$$

(the generic k^{th} element)

$$\boldsymbol{\theta}_k^{(j)} \sim p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_1^{(j)}, \ldots, \boldsymbol{\theta}_{k-1}^{(j)}, \boldsymbol{\theta}_{k+1}^{(j)}, \ldots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

$$\vdots$$

$$\boldsymbol{\theta}_p^{(j)} \sim p(\boldsymbol{\theta}_p \,|\, \boldsymbol{\theta}_1^{(j)}, \ldots, \boldsymbol{\theta}_{p-1}^{(j)}, \mathbf{y})$$

- 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.