Bayesian Linear Models

Andrew Finley[1] & Jeffrey Doser[2]

May 15, 2023

[1]Department of Forestry, Michigan State University.

[2]Department of Integrative Biology, Michigan State University.

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 $\mathbf{x}_i^\top = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$.
- The first element of $\mathbf{x}_i^\top$ 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:

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

- 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:

  $$y_i \mid \boldsymbol{\beta}, \sigma^2, \mathbf{x}_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n \; ;$$
  $$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i^\top \boldsymbol{\beta} \; ; \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \; ;$$
  $$\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\mu}_\beta, \sigma^2 \mathbf{V}_\beta) \; ; \quad \sigma^2 \sim IG(a, b) \; .$$

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

- We assume $\mathbf{X}$ is observed without error and all inference is conditional on $\mathbf{X}$.

- The above model is often written in terms of the posterior density $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\boldsymbol{\theta}, \mathbf{y})$:

  $$IG(\sigma^2 \mid a, b) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \sigma^2 \mathbf{V}_\beta) \times \prod_{i=1}^{n} N(y_i \mid \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2).$$

Conjugate Bayesian (General) Linear Regression

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

$$\mathbf{y} \,|\, \beta, \sigma^2, \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{V}_y)$$
$$\beta \,|\, \sigma^2 \sim N(\boldsymbol{\mu}_\beta, \sigma^2 \mathbf{V}_\beta) \;;$$
$$\sigma^2 \sim IG(a, b) \;.$$

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

$$IG(\sigma^2 \,|\, a, b) \times N(\beta \,|\, \boldsymbol{\mu}_\beta, \sigma^2 \mathbf{V}_\beta) \times N(\mathbf{y} \,|\, \mathbf{X}\beta, \sigma^2 \mathbf{V}_y)$$

- The model on the previous slide is a special case with $\mathbf{V}_y = \mathbf{I}_n$ ($n \times n$ identity matrix).

Conjugate Bayesian (General) Linear Regression

- The joint posterior density can be written as

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

where

$$a^* = a + \frac{n}{2} \; ; \quad b^* = b + \frac{1}{2}\left(\boldsymbol{\mu}_\beta^\top \mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{y}^\top\mathbf{V}_y^{-1}\mathbf{y} - \mathbf{m}^\top\mathbf{Mm}\right) \; ;$$

$$\mathbf{m} = \mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}^\top\mathbf{V}_y^{-1}\mathbf{y} \; ; \quad \mathbf{M}^{-1} = \mathbf{V}_\beta^{-1} + \mathbf{X}^\top\mathbf{V}_y^{-1}\mathbf{X} \; .$$

- Exact posterior sampling from $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$ will automatically yield samples from $p(\boldsymbol{\beta} \mid \mathbf{y})$ and $p(\sigma^2 \mid \mathbf{y})$.
- For each $j = 1, 2, \ldots, N$ do the following:
  1. Draw $\sigma^2_{(j)} \sim IG(a^*, b^*)$
  2. Draw $\beta_{(j)} \sim N\left(\mathbf{Mm}, \sigma^2_{(j)}\mathbf{M}\right)$
- 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:

$$p(\theta_1, \theta_2 \,|\, \mathbf{y}) = p(\theta_1 \,|\, \mathbf{y}) \times p(\theta_2 \,|\, \theta_1, \mathbf{y}) \; .$$

- In conjugate models, it is often easy to draw samples from $p(\theta_1 \,|\, \mathbf{y})$ and from $p(\theta_2 \,|\, \theta_1, \mathbf{y})$.
- We can draw N samples from $p(\theta_1, \theta_2 \,|\, \mathbf{y})$ as follows.
- For each $j = 1, 2, \ldots, N$ do the following:
  1. Draw $\theta_{1(j)} \sim p(\theta_1 \,|\, \mathbf{y})$
  2. Draw $\theta_{2(j)} \sim p(\theta_2 \,|\, \theta_{1(j)}, \mathbf{y})$
- Remarkably, the $\theta_{2(j)}$'s drawn above have marginal distribution $p(\theta_2 \,|\, \mathbf{y})$ (see, Gelfand and Smith 1990).
- "Automatic Marginalization" we draw samples $p(\theta_1, \theta_2 \,|\, \mathbf{y})$ and automatically get samples from $p(\theta_1 \,|\, \mathbf{y})$ and $p(\theta_2 \,|\, \mathbf{y})$.

Bayesian predictions from linear regression

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

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

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

Bayesian predictions from linear regression (cont'd)

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

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

- Easy to derive the conditional density:

$$p(\tilde{\mathbf{y}} \,|\, \boldsymbol{\theta}, \mathbf{y}) = p(\tilde{\mathbf{y}} \,|\, \boldsymbol{\theta}) = N(\tilde{\mathbf{y}} \,|\, \tilde{\mathbf{X}}\beta, \sigma^2 \mathbf{I}_m)$$

- Posterior predictive density:

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

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

Bayesian predictions from general linear regression

- For example, consider the case with general $\mathbf{V}_y$. Specify:

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

- Derive the conditional density $p(\tilde{\mathbf{y}} \,|\, \boldsymbol{\theta}, \mathbf{y}) = N\left(\tilde{\mathbf{y}} \,|\, \boldsymbol{\mu}_{\tilde{y}|y}, \sigma^2 \mathbf{V}_{\tilde{y}|y}\right)$:

$$\boldsymbol{\mu}_{\tilde{y}|y} = \tilde{\mathbf{X}}\beta + \mathbf{V}_{y\tilde{y}}^{\top}\mathbf{V}_y^{-1}(\mathbf{y} - \mathbf{X}\beta) \; ; \quad \mathbf{V}_{\tilde{y}|y} = \mathbf{V}_{\tilde{y}} - \mathbf{V}_{y\tilde{y}}^{\top}\mathbf{V}_y^{-1}\mathbf{V}_{y\tilde{y}}.$$

- Posterior predictive density:

$$p(\tilde{\mathbf{y}} \,|\, \mathbf{y}) = \int N\left(\tilde{\mathbf{y}} \,|\, \boldsymbol{\mu}_{\tilde{y}|y}, \sigma^2 \mathbf{V}_{\tilde{y}|y}\right) p(\beta, \sigma^2 \,|\, \mathbf{y}) d\beta d\sigma^2.$$

- For each $j = 1, 2, \ldots, N$ do the following:
  1. Draw $\{\beta_{(j)}, \sigma^2_{(j)}\} \sim p(\beta, \sigma^2 \,|\, \mathbf{y})$
  2. Compute $\boldsymbol{\mu}_{\tilde{y}|y}$ using $\beta_{(j)}$ and draw $\tilde{\mathbf{y}}_{(j)} \sim N(\boldsymbol{\mu}_{\tilde{y}|y}, \sigma^2_{(j)} \mathbf{V}_{\tilde{y}})$

Application to Bayesian Geostatistics

- Consider the spatial regression model
$$y(s_i) = \mathbf{x}^\top(\mathbf{s}_i)\beta + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$
where $w(\mathbf{s}_i)$'s are spatial random effects and $\epsilon(\mathbf{s}_i)$'s are unstructured errors ("white noise").
- $\mathbf{w} = (w(\mathbf{s}_1), w(\mathbf{s}_2), \ldots, w(\mathbf{s}_n))^\top \sim N(\mathbf{0}, \sigma^2\mathbf{R}(\phi))$
- $\boldsymbol{\epsilon} = (\epsilon(\mathbf{s}_1), \epsilon(\mathbf{s}_2), \ldots, \epsilon(\mathbf{s}_n))^\top \sim N(\mathbf{0}, \tau^2\mathbf{I}_n)$
- Integrating out random effects leads to a Bayesian model:
$$IG(\sigma^2 \mid a, b) \times N(\beta \mid \boldsymbol{\mu}_\beta, \sigma^2\mathbf{V}_\beta) \times N(\mathbf{y} \mid \mathbf{X}\beta, \sigma^2\mathbf{V}_y)$$
where $\mathbf{V}_y = \mathbf{R}(\phi) + \alpha\mathbf{I}_n$ and $\alpha = \tau^2/\sigma^2$.
- Fixing ϕ and α (e.g., from variogram or other EDA) yields a conjugate Bayesian model (see `bayesGeostatExact()` in `spBayes` package).
- Exact posterior sampling is easily achieved as before!

Inference on spatial random effects

- Rewrite the model in terms of $\mathbf{w}$ as:

$$IG(\sigma^2 \,|\, a, b) \times N(\beta \,|\, \boldsymbol{\mu}_\beta, \sigma^2 \mathbf{V}_\beta) \times N(\mathbf{w} \,|\, \mathbf{0}, \sigma^2 \mathbf{R}(\phi)) \\ \times N(\mathbf{y} \,|\, \mathbf{X}\beta + \mathbf{w}, \tau^2 \mathbf{I}_n).$$

- Posterior distribution of spatial random effects $\mathbf{w}$:

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

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

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

Inference on the process

- Posterior distribution of $w(\mathbf{s}_0)$ at new location $\mathbf{s}_0$:

$$p(w(\mathbf{s}_0)\,|\,\mathbf{y}) = \int N(w(\mathbf{s}_0)\,|\,\mu_{w(\mathbf{s}_0)|w}, \sigma^2_{w(\mathbf{s}_0)|w}) \times p(\sigma^2, \mathbf{w}\,|\,\mathbf{y}) d\sigma^2 d\mathbf{w} \; ,$$

where

$$\begin{aligned}\mu_{w(\mathbf{s}_0)|w} &= \mathbf{r}^\top(\mathbf{s}_0;\phi)\mathbf{R}^{-1}(\phi)\mathbf{w} \; ; \\ \sigma^2_{w(\mathbf{s}_0)|w} &= \sigma^2\{1 - \mathbf{r}^\top(\mathbf{s}_0;\phi)\mathbf{R}^{-1}(\phi)\mathbf{r}(\mathbf{s}_0,\phi)\}\end{aligned}$$

- For each $j = 1, 2, \ldots, N$ do the following:
  1. Compute $\mu_{w(\mathbf{s}_0)|w}$ and $\sigma^2_{w(\mathbf{s}_0)|w}$ from $\mathbf{w}_{(j)}$ and $\sigma^2_{(j)}$.
  2. Draw $w_{(j)}(\mathbf{s}_0) \sim N(\mu_{w(\mathbf{s}_0)|w}, \sigma^2_{w(\mathbf{s}_0)|w})$.

Bayesian "kriging" or prediction

- Posterior predictive distribution at new location $\mathbf{s}_0$ is $p(y(\mathbf{s}_0) \mid \mathbf{y})$:

$$\int N(y(\mathbf{s}_0) \mid \mathbf{x}^\top(s_0)\beta + w(\mathbf{s}_0), \alpha\sigma^2) \times p(\beta, \sigma^2, \mathbf{w} \mid \mathbf{y}) \mathrm{d}\beta \mathrm{d}\sigma^2 \mathrm{d}\mathbf{w} \,,$$

- For each $j = 1, 2, \ldots, N$ do the following:
  1. Draw $y_{(j)}(\mathbf{s}_0) \sim N(\mathbf{x}^\top(\mathbf{s}_0)\beta_{(j)} + w_{(j)}(s_0), \alpha\sigma^2_{(j)})$.

Non-conjugate models: The Gibbs Sampler

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

$$
\begin{aligned}
&\theta_1^{(j)} \sim p(\theta_1^{(j)} \mid \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, \mathbf{y}) \\
&\theta_2^{(j)} \sim p(\theta_2 \mid \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \mathbf{y}) \\
&\vdots \\
&\text{(the generic } k^{th} \text{ element)} \\
&\theta_k^{(j)} \sim p(\theta_k | \theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}, \mathbf{y}) \\
&\vdots \\
&\theta_p^{(j)} \sim p(\theta_p \mid \theta_1^{(j)}, \ldots, \theta_{p-1}^{(j)}, \mathbf{y})
\end{aligned}
$$

- 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.

When we don't want to fix ϕ and $\alpha = \tau^2/\sigma^2$

Latent Bayesian Model

$$N(\mathbf{y} \mid \mathbf{X}\beta + \mathbf{w}, \tau^2\mathbf{I}) \times N(\mathbf{w} \mid \mathbf{0}, \sigma^2\mathbf{R}(\phi)) \times N(\beta \mid \boldsymbol{\mu}_\beta, \mathbf{V}_\beta)$$
$$\times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times Unif(\phi \mid a_\phi, b_\phi)$$

Sampler:

- Full conditionals for β, τ^2, σ^2 and $w(\mathbf{s}_i)$'s
- Metropolis step for updating ϕ
- Pros: Full conditional distributions for all parameters except ϕ, easy to code up
- Cons: High-dimensional parameter space can mean slow convergence

When we don't want to fix ϕ and $\alpha = \tau^2/\sigma^2$ (cont'd)

Collapsed Bayesian Model

$$N(\mathbf{y} \,|\, \mathbf{X}\beta, \sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \mathbf{V}_\beta) \\ \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times Unif(\phi \,|\, a_\phi, b_\phi)$$

Sampler:

- Full conditional for $\boldsymbol{\beta}$
- Metropolis step for updating τ^2, σ^2, ϕ
- Pros: Low-dimensional parameter space
- "Recover" $w(\mathbf{s}_i)$'s in a posterior predictive fashion

We can also integrate out β! See Finley et al. (2015) for details https://www.jstatsoft.org/article/view/v063i13 and implementation in the spBayes package.

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 $j = 1, \ldots N$. This yields $\theta^{(1)}, \ldots, \theta^{(N)}$, 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 $j = 1, \ldots N$ to obtain samples $\theta^{(1)}, \ldots, \theta^{(N)}$.