# ECO545A: Bayesian Econometrics

March 19, 2025

## Chapter 7: Simulation by MCMC Methods

The basis of an MCMC algorithm is the construction of a transition kernel $p(x, y)$ that has an invariant density as the target density.

Given such a kernel, the process can be started at $x_0$ to yield a draw $x_1$ from $p(x_0, x_1)$, $x_2$ from $p(x_1, x_2)$, ..., and $x_G$ from $p(x_{G-1}, x_G)$ where G is the desired number of simulations. After a transient period, the distribution of $x_g$ is approximately equal to the target distribution.

There is a general principle for finding such kernels known as the Metropolis-Hastings (MH) Algorithm. A special case of MH algorithm is the Gibbs sampling.

### Gibbs Sampling

Gibbs sampling is a well known MCMC algorithm and is actually a special case of the MH algorithm. Gibbs sampling is applicable when the conditional distributions of all the parameters have a known tractable form and sampling is possible.

Suppose, there is a non standard joint distribution $f(x_1, x_2)$ from which we cannot sample. However, we can sample from each of its conditional distribution i.e., $f(x_1|x_2)$ and $f(x_2|x_1)$. Then the Gibbs algorithm proceeds as follows:

**Algorithm 1 (Gibbs Sampling)**

---

- Choose a starting value $x_2^{(1)}$
- At the 2nd-iteration, draw $x_1^{(2)}$ from $f(x_1|x_2^{(1)})$ and $x_2^{(2)}$ from $f(x_2|x_1^{(2)})$.
- $\vdots$
- At the g-th iteration, draw $x_1^{(g)}$ from $f(x_1|x_2^{(g-1)})$ and $x_2^{(g)}$ from $f(x_2|x_1^{(g)})$

---

The Gibbs process is continued until the desired number of iterations is obtained. The roles of $x_1$ and $x_2$ may be interchanged.

Note that the starting value is arbitrary and not drawn from an invariant distribution. So, some portion of the initial sample is discarded: this is called the transient, burn-in or warm-up

sample. Let B denote the burn-in size. For $g > B$, the distribution of the draws is approximately the target distribution.

Below, we have a simple proof that shows that the invariant distribution of the Gibbs kernel is the target distribution. Let $x = (x_1, x_2)$ be the values of the random variable at the beginning of the one iteration and $y = (y_1, y_2)$ be the values at the end of the iteration. The Gibbs kernel is $p(x, y) = f(y_1|x_2)f(y_2|y_1)$ from which we compute,

$$\begin{aligned}\int p(x,y)f(x)dx &= \int f(y_1|x_2)f(y_2|y_1)f(x_1,x_2)dx_1dx_2 \\ &= f(y_2|y_1)\int f(y_1|x_2)f(x_2)dx_2 \\ &= f(y_2|y_1)f(y_1) = f(y)\end{aligned}$$

which proves that $f(\cdot)$ is an invariant distribution of the Gibbs kernel.

The proof that the invariant distribution of the Gibbs kernel is the target distribution is necessary but not a sufficient condition for the kernel to converge to the target. Such conditions are very technical and difficult to verify for particular cases. Tierney (1994) states that most Gibbs samplers satisfy the conditions of the following theorem:

**Theorem 7.1**

Suppose P is $\pi-$ irreducible and has $\pi$ as its invariant distribution. If $P(x, .)$ is absolutely continuous with respect to $\pi$ for all x, then P is Harris recurrent.

Extending the Gibbs sampling to more than two blocks is easy. Suppose there are $d$ blocks of random variable and their conditional densities are tractable from which we can draw random values. Then the Gibbs process looks like the following:

**Algorithm 2 (Gibbs Sampling with $d$ blocks)**

---

- Choose $x_2^{(1)}, \ldots, x_d^{(1)}$
- In the second iteration draw values as follows:

$$\begin{aligned} x_1^{(2)} &\sim f(x_1|x_2^{(1)}, \ldots, x_d^{(1)}), \\ x_2^{(2)} &\sim f(x_2|x_1^{(2)}, x_3^{(1)} ..., x_d^{(1)}), \\ &\cdots \\ x_d^{(2)} &\sim f(x_d|x_1^{(2)}, x_2^{(2)}, \ldots, x_{d-1}^{(2)}). \end{aligned}$$

- $\vdots$
- At the $g$-th iteration, draw

$$\begin{aligned} x_1^{(g)} &\sim f(x_1|x_2^{(g-1)}, \ldots, x_d^{(g-1)}), \\ x_2^{(g)} &\sim f(x_2|x_1^{(g)}, x_3^{(g-1)}, \ldots, x_d^{(g-1)}), \\ &\cdots \\ x_d^{(g)} &\sim f(x_d|x_1^{(g)}, x_2^{(g)}, \ldots, x_{d-1}^{(g)}). \end{aligned}$$

---

We next look at 2 examples for Gibbs sampling.

**Example 1**: *Mean and Precision for Normal Model*

$$\begin{aligned} \text{Model}: \quad y_i &\sim N(\mu, h^{-1}), \qquad \text{for } \; i = 1, 2, \ldots, n \\ \text{Priors}: \quad \mu &\sim N(\mu_0, h_0^{-1}), \\ h &\sim Ga\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right). \end{aligned}$$

Using Bayes' theorem, the conditional posterior distribution can be expressed as the product of the likelihood and prior distributions. Collecting terms one at a time, while holding the other fixed, we can derive the expression for the conditional posterior distribution. The conditional posterior distribution for $h|\mu, y$ is,

$$\begin{aligned} \pi(h|\mu, y) &\propto h^{\frac{(\alpha_0+n)}{2}-1} \exp\left[-h\frac{\delta_0 + \sum(y_i - \mu)^2}{2}\right] \\ h|\mu, y &\sim Ga\left[\frac{(\alpha_0 + n)}{2}, \frac{\delta_0 + \sum(y_i - \mu)^2}{2}\right], \end{aligned}$$

and the conditional posterior distribution for $\mu|h, y$ is,

$$\pi(\mu|h, y) \propto \exp\left[-\frac{h_0 + hn}{2}\left(\mu - \frac{h_0\mu_0 + hn\bar{y}}{h_0 + hn}\right)^2\right]$$

$$\mu|h,y \quad \sim \quad N\left[\frac{h_0\mu_0 + hn\bar{y}}{h_0 + hn}, (h_0 + hn)^{-1}\right].$$

**Algorithm 3 (Gibbs Sampling for Normal Model)**

---

- Choose starting value for $\mu = \mu^{(1)}, h = h^{(1)}$.
- At the 2-nd iteration, draw

$$\begin{aligned}\mu^{(2)}|h,y &\sim N\left[\frac{h_0\mu_0 + h^{(1)}n\bar{y}}{h_0 + h^{(1)}n}, \left(h_0 + h^{(1)}n\right)^{-1}\right] \\ h^{(2)}|\mu,y &\sim Ga\left[\frac{\alpha_0 + n}{2}, \frac{\delta_0 + \sum\left(y_i - \mu^{(2)}\right)^2}{2}\right]\end{aligned}$$

- At the g-th iteration, draw

$$\begin{aligned}\mu^{(g)}|h,y &\sim N\left[\frac{h_0\mu_0 + h^{(g-1)}n\bar{y}}{h_0 + h^{(g-1)}n}, \left(h_0 + h^{(g-1)}n\right)^{-1}\right] \\ h^{(g)}|\mu,y &\sim Ga\left[\frac{\alpha_0 + n}{2}, \frac{\delta_0 + \sum\left(y_i - \mu^{(g)}\right)^2}{2}\right]\end{aligned}$$

---

**Example 2**: *Poisson Model with Changing Parameters*
Suppose, the random variable $y$ has the following *pmf*:

$$p(y_i) = \begin{cases} \frac{e^{-\theta_1}\theta_1^{y_i}}{y_i!}, & \text{for } i = 1, 2, \ldots, k \\ \\ \frac{e^{-\theta_2}\theta_2^{y_i}}{y_i!}, & \text{for } i = k+1, \ldots, n \end{cases}$$

where $y_i = 0, 1, ..., n$ and the switch point $k$ is unknown.

We assume conditionally conjugate priors on the parameters: $\theta_1 \sim Ga(\alpha_{10}, \beta_{10})$, $\theta_2 \sim Ga(\alpha_{20}, \beta_{20})$, and $\pi(k = j) = 1/n$ for $j = 1, 2, \ldots, n$. Note the last one is a discrete uniform distribution to $k$ over the values $1, 2, \ldots, n$ which includes the possibility that no change occurs i.e. $k = n$. The *pdf*'s of the prior distributions has the following form:

$$\begin{aligned}\pi(\theta_1) &= \frac{\beta_{10}^{\alpha_{10}}}{\Gamma(\alpha_{10})}\,\theta_1^{\alpha_{10}-1}e^{-\beta_{10}\theta_1} \\ \pi(\theta_2) &= \frac{\beta_{20}^{\alpha_{20}}}{\Gamma(\alpha_{20})}\,\theta_2^{\alpha_{20}-1}e^{-\beta_{20}\theta_2} \\ \pi(k = j) &= 1/n.\end{aligned}$$

The Poisson model with changing parameters yields the following likelihood function:

$$f(y|\theta_1, \theta_2, k) = \frac{1}{y!}\prod_{i=1}^{k} e^{-\theta_1}\theta_1^{y_i} \prod_{i=k+1}^{n} e^{-\theta_2}\theta_2^{y_i}.$$

Using the Bayes' theorem, the joint posterior distribution is obtained as the product of the likelihood and the prior distributions. The resulting expression is,

$$\pi(\theta_1, \theta_2, k|y) \propto \theta_1^{\alpha_{10}-1} e^{-\beta_{10}\theta_1}\, \theta_2^{\alpha_{20}-1} e^{-\beta_{20}\theta_2} \prod_{i=1}^{k} e^{-\theta_1}\theta_1^{y_i} \prod_{i=k+1}^{n} e^{-\theta_2}\theta_2^{y_i}.$$

The conditional posterior distributions are derived by collecting terms for one parameters, while holding the remaining parameters fixed. Following this procedure, we find the following conditional posterior distributions:

$$\begin{aligned}
\pi(\theta_1|\theta_2, k, y) &\propto \theta_1^{\alpha_{10}-1} e^{-\beta_{10}\theta_1} \prod_{i=1}^{k} e^{-\theta_1}\theta_1^{y_i} \sim Ga(\alpha_{10} + \sum_{i=1}^{k} y_i, \beta_{10} + k) \\
\pi(\theta_2|\theta_1, k, y) &\propto \theta_2^{\alpha_{20}-1} e^{-\beta_{20}\theta_2} \prod_{i=k+1}^{n} e^{-\theta_2}\theta_2^{y_i} \sim Ga(\alpha_{20} + \sum_{i=k+1}^{n} y_i, \beta_{20} + n - k) \\
\pi(k|\theta_2, \theta_1, y) &\propto e^{-k\theta_1}\theta_1^{\sum_{i=1}^{k} y_i} e^{-(n-k)\theta_2}\theta_2^{\sum_{k+1}^{k} y_i} \\
&\propto e^{-k(\theta_1-\theta_2)}\theta_1^{\sum_{i=1}^{k} y_i}\theta_2^{\sum_{1}^{n} y_i - \sum_{1}^{k} y_i} \\
&\propto e^{-k(\theta_1-\theta_2)}\left(\frac{\theta_1}{\theta_2}\right)^{\sum_{1}^{k} y_i}
\end{aligned}$$

For the conditional posterior of $k$, we know that $\sum_k \pi(k|\theta_2, \theta_1, y) = 1$. Therefore, we have,

$$\pi(k|\theta_2, \theta_1, y) = \frac{e^{-k(\theta_1-\theta_2)}\left(\frac{\theta_1}{\theta_2}\right)^{\sum_{1}^{k} y_i}}{\sum_{k=1}^{n} e^{-k(\theta_1-\theta_2)}\left(\frac{\theta_1}{\theta_2}\right)^{\sum_{1}^{k} y_i}}.$$

The Gibbs sampler usually works well in practice but there are some situations in which it does not. For example, if there are high correlations between one or more of the random variables in different blocks, the algorithm may not mix well.

## Marginal Likelihood using Gibbs Output

Marginal likelihood is required for calculating the Bayes factor. There exists several method for calculating the marginal likelihood, but the most robust method are those proposed by Chib (1995) and Chib and Jeliazkov (2001).

We first describe how to calculate the marginal likelihood based on Gibbs output as described by Chib (1995). The Chib (1995) method begins with the identity,

$$\pi(\theta^*|y) = \frac{f(y|\theta^*)\ \pi(\theta^*)}{f(y)},$$

where $\theta^*$ is a particular value of $\theta$ and $f(y)$ is the marginal likelihood. For numerical accuracy, $\theta^*$ is typically chosen to be a high density point such as the mean or mode of sample values. This identity can be re-written as follows,

$$f(y) = \frac{f(y|\theta^*)\pi(\theta^*)}{\pi(\theta^*|y)}.$$

The Chib method computes the quantities on the right hand side from the Gibbs output. The likelihood $f(y|\theta^*)$ and prior $\pi(\theta^*)$ are readily available. The main problem is to compute $\pi(\theta^*|y)$ for which the normalizing constant is unknown.

Consider the simple case where Gibbs Algorithm is run in two blocks, denoted by $\theta_1$ and $\theta_2$.

$$\pi(\theta_1^*, \theta_2^*|y) = \pi(\theta_1^*|\theta_2^*, y)\pi(\theta_2^*|y).$$

The first term is readily available when running the Gibbs sampler. To compute the second term, Chib employs the identity,

$$\begin{aligned} \pi(\theta_2^*|y) &= \int \pi(\theta_1, \theta_2^*|y)\ d\theta_1 \\ &= \int \pi(\theta_2^*|y, \theta_1)\, \pi(\theta_1|y)\ d\theta_1 \end{aligned}$$

which can be approximated by,

$$\hat{\pi}(\theta_2^*|y) = \frac{1}{G}\sum \pi(\theta_2^*|y, \theta_1^{(g)})$$

where the values $\theta_1^{(g)}$ are taken from the Gibbs output.

When there are three or more blocks, the computation requires additional simulations. For example, consider the three block case,

$$f(y) = \frac{f(y|\theta_1^*, \theta_2^*, \theta_3^*)\pi(\theta_1^*, \theta_2^*, \theta_3^*)}{\pi(\theta_1^*, \theta_2^*, \theta_3^*|y)}.$$

The denominator can be written as,

$$\pi(\theta_1^*, \theta_2^*, \theta_3^*|y) = \pi(\theta_1^*|y)\, \pi(\theta_2^*|\theta_1^*, y)\, \pi(\theta_3^*|\theta_1^*, \theta_2^*, y).$$

The first term on the right hand side $\pi(\theta_1^*|y)$ can be approximated as,

$$\hat{\pi}(\theta_1^*|y) = \frac{1}{G} \sum \pi(\theta_1^*|\theta_2^{(g)}, \theta_3^{(g)}, y).$$

For the second term $\pi(\theta_2^*|\theta_1^*, y)$, we can re-write it as,

$$\pi(\theta_2^*|\theta_1^*, y) = \int \pi(\theta_2^*|\theta_1^*, \theta_3, y)\, \pi(\theta_3|\theta_1^*, y)\, d\theta_3$$

which can be can be approximated based on Gibbs output as follows,

$$\hat{\pi}(\theta_2^*|\theta_1^*, y) = \frac{1}{G} \sum \pi(\theta_2^*|\theta_1^*, \theta_3^{(g)}, y),$$

where the values $\theta_3^{(g)}$ are generated from a "reduced run" in which $\theta_2^{(g)}$ and $\theta_3^{(g)}$ are sampled from $\pi(\theta_2|\theta_1^*, \theta_3, y)$ and $\pi(\theta_3|\theta_1^*, \theta_2, y)$ respectively and $\theta_1$ is fixed at $\theta_1^*$. Computations for the reduced runs can use the same code as the original run but $\theta_1$ is held constant at $\theta_1^*$. Finally, the value $\pi(\theta_3^*|\theta_1^*, \theta_2^*, y)$ is available directly from the conditional distribution.

### Metropolis Hastings Algorithm

The MH Algorithm is more general than Gibbs sampler because it does not require the knowledge of conditional distributions for sampling.

To generate a sample from $f(X)$ where $X$ may be a scalar or vector random variable, the first step is to find a kernel $p(X, Y)$ that has $f(\cdot)$ as its invariant distribution. To this end, we introduce the idea of reversible kernel, defined as kernel $q(\cdot, \cdot)$ for which $f(x)q(x, y) = f(y)q(y, x)$. If $q(\cdot, \cdot)$ is reversible,

$$\begin{aligned} P(y \in A) &= \int_A \int_{R^d} f(x)q(x, y)\, dx\, dy \\ &= \int_A \int_{R^d} f(y)q(y, x)\, dx\, dy = \int_A f(y)dy. \end{aligned}$$

This shows that $f(\cdot)$ is the invariant distribution for the kernel $q(\cdot, \cdot)$ because the probability that $y$ is contained in $A$ is computed from $f(\cdot)$.

(Note: In Markov Chain theory, we start with a transition kernel and determine conditions under which that it converges to the invariant distribution. In MCMC theory, we start with an invariant distribution and try to find a transition kernel that converges to the invariant distribution. We focus on the latter as far as this course is concerned.)

The fact that a reversible kernel has this property can help in finding a kernel that has the desired target distribution. Starting with a non reversal proposal density, the trick is to make

an irreversible kernel reversible. If a kernel is not reversible, for some pair $(x, y)$, we have,

$$f(x)q(x, y) > f(y)q(y, x)$$

This means that the kernel goes from $x$ to $y$ with greater probability than it goes from $y$ to $x$.

The MH Algorithm multiplies both sides by $\alpha(\cdot, \cdot)$ that turns the irreversible kernel $q(\cdot, \cdot)$ into the reversible kernel,

$$p(x, y) = \alpha(x, y)q(x, y).$$

Consequently, we have the following,

$$f(x)\alpha(x, y)q(x, y) = f(y)\alpha(y, x)q(y, x).$$

The expression $\alpha(x, y)q(x, y)$ is interpreted as follows: Starting from x, generate a value y from the kernel $q(x, y)$ and make a move to y with probability $\alpha(x, y)$.

How to define $\alpha(x, y)$ is the next question. Suppose that,

$$f(x)q(x, y) > f(y)q(y, x)$$

Roughly speaking, this means that the kernel goes from $x$ to $y$ with greater probability than it goes from $y$ to $x$. Accordingly, if the process is at $y$ and the kernel proposes a move to $x$, that move should be made with high probability. So, we set $\alpha(y, x) = 1$. As a result, then $\alpha(x, y)$ is determined and we have,

$$f(x)\alpha(x, y)q(x, y) = f(y)\alpha(y, x)q(y, x),$$

which implies that $\alpha(x, y)$ can be written as,

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1\right\}, & \text{if } f(x)q(x, y) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The condition $f(x)q(x, y) \neq 0$ is not a problem because we always begin $(x)$ in the support of distribution and choose a value in the support $(y)$.

To summarize in algorithmic form:

**Algorithm 4 (Metropolis-Hastings Algorithm)**

---

- Given $x$, generate $y$ from the proposal density $q(x, y)$.
- Generate $U \sim Unif(0, 1)$. Compute $\alpha(x, y)$ as above. If $u \leq \alpha(x, y)$ return $y$. Otherwise, return $x$ and go to step 1.

---

Note: The unknown constant in the target distribution is not needed to compute $\alpha(\cdot,\cdot)$, because it cancels out via the fraction $f(y)/f(x)$.

The next implementation issue is how to choose the proposal density $q(x,y)$.

The kernel should generate proposals that have a reasonably good probability of acceptance, otherwise the same value will be repeated leading to poor mixing. On the other hand, if the kernel generate proposals too close to the current, the acceptance will be large but the sample will not be able to explore the support leading to poor mixing.

Two straightforward kernels are random-walk and independent kernels.

(1) *Random Walk*: The proposal is generated as $y = x + u$ where $x$ is the current value and $u \sim D$, where $D$ is some distribution. If $D$ is symmetric around zero, then $h(u) = h(-u)$ and the kernel has the property $q(x,y) = q(y,x)$. The MH acceptance probability reduces to,

$$\alpha(x,y) = \begin{cases} \min\left\{\frac{f(y)}{f(x)}, 1\right\}, & \text{if } f(x) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

A move to a higher density point $f(y) > f(x)$ is always accepted but move to a lower density point is accepted with probability $\frac{f(y)}{f(x)}$ where $f(y) < f(x)$. With a random-walk MH algorithm, the optimal acceptance rate is 30 percent.

(2) *Independent Chain*: The proposal density is independent of the current state of the chain, $q(x,y) = q(y)$. For this type of kernel,

$$\alpha(x,y) = \begin{cases} \min\left\{\frac{f(y)/q(y)}{f(x)/q(x)}, 1\right\}, & \text{if } f(x)q(y) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note: The optimal acceptance rate for a random walk MH algorithm is 33 percent. For independent MH algorithm, we aim for high acceptance rate. Anything above 90 percent acceptance is good.

**Example**: Generate data from a $Beta(3,4)$ distribution with $U(0,1)$ as the proposal density with independent chain.

**Algorithm 5 (MH for Beta(3, 4) with $U(0,1)$ Proposal)**

---

- Set $x^{(1)}$ equal to a number between 0 and 1.
- At the g-th iteration, generate $u_1, u_2$ from U(0,1).
- Compute the MH ratio,

$$\alpha(x^{(g-1)}, u_2) = \frac{u_2^2(1-u_2)^3}{(x^{(g-1)})^2(1-x^{(g-1)})^3}.$$

  If $u_1 \leq \alpha(x^{(g-1)}, u_2)$, set $x^{(g)} = u_2$. Otherwise, $x^{(g)} = x^{(g-1)}$.
- Go to step (2) until the desired number of iterations is obtained.

---

## Marginal Likelihood using MH Output

To calculate the marginal likelihood based on MH output, we follow the approach proposed by Chib and Jeliazkov (2001).

Similar to Chib (1995), we start with the identity:

$$f(y) = \frac{f(y|\theta^*)\pi(\theta^*)}{\pi(\theta^*|y)},$$

where the challenge is to compute the denominator on the right hand side $\pi(\theta^*|y)$. The balancing equation yields,

$$\alpha(\theta, \theta^*|y)q(\theta, \theta^*|y)\pi(\theta|y) = \alpha(\theta^*, \theta|y)q(\theta^*, \theta|y)\pi(\theta^*|y).$$

The last expression can be written to find,

$$\pi(\theta^*|y) = \frac{\int \alpha(\theta, \theta^*|y)q(\theta, \theta^*|y)\pi(\theta|y)d\theta}{\int \alpha(\theta^*, \theta|y)q(\theta^*, \theta|y)d\theta}.$$

The numerator of the above expression can be directly estimated from the MH output as,

$$\frac{1}{G}\sum_g \alpha(\theta^{(g)}, \theta^*|y)\ q(\theta^{(g)}, \theta^*|y)$$

The denominator is estimated by drawing $\theta^{(j)}$ from $q(\theta^*, \theta|y)$, for $j = 1, 2, \ldots, J$, i.e. generate values from the proposal density all of which are conditioned on $\theta^*$ and then compute $\frac{1}{J}\sum_j \alpha(\theta^*, \theta^{(j)}|y)$.

### Numerical Standard Error and Convergence

For an independent sample $z_1, z_2, \ldots, z_G$, we have:

$$\text{Sample Mean :} \qquad \bar{z} = \frac{1}{G}\sum_g z^{(g)}$$

$$\text{Variance :} \qquad \widehat{var}(z) = \frac{1}{G-1}\sum_g (z^{(g)} - \bar{z})^2$$

$$\text{Numerical standard error :} \qquad nse = \sqrt{\frac{var(z)}{G}}.$$

When the samples are not independent, as in MCMC output, the above expression for variance is not valid because we need to account for the covariance. Here we study 'batch means' method. The logic is that autocorrelations of high enough order tend to zero. Calculate the empirical correlation and choose lag length say $j_0$ at which autocorrelation is small, say less than 0.05. Divide the sample of G observations into $N = G/j_0$ groups or batches, adjusting G or $j_0$ to get an integer value of N. For each batch, $b = 1, 2, \ldots, N$, calculate the batch mean $\bar{z}_b$ and variance,

$$\widehat{var}(\bar{z}_b) = \frac{1}{N-1}\sum_{b=1}^{N}(\bar{z}_b - \bar{z})^2$$

where $\bar{z}$ is the overall mean. The numerical standard error is computed as,

$$nse = \sqrt{\frac{\widehat{var}(\bar{z}_b)}{N}}.$$

Positive autocorrelation leads to larger variance than that from independent samples. So several measures have been proposed as the cost of working with correlated samples. One is relative numerical efficiency (RNE) and the other is inefficiency factor (IF). The RNE is gives by the following expression,

$$RNE = \sqrt{\frac{\sum_g \left(z^{(g)} - \bar{z}\right)^2}{G(G-1)} \Big/ \frac{\widehat{var}_c(\bar{z}_b)}{N}},$$

where $\widehat{var}_c(\bar{z})$ is the estimate of the variance corrected for autocorrelation by the batch means method. Small values of RNE indicate substantial autocorrelation in the chain.

The reciprocal of RNE is the *inefficiency factor* or *autocorrelation time*. Large values of inefficiency factor indicate the cost of working with correlated samples.

Inefficiency factor also bears a direct relationship with *Effective Sample Size* (ESS), where the latter can be obtained as the total number of (post burn-in) MCMC draws divided by the inefficiency factor.

Another method of calculating the inefficiency factor is by using the formula,

$$1 + 2\sum_{t=1}^{T}\rho_k(t)\left(\frac{T-t}{T}\right),$$

where $\rho_k(t)$ denotes the autocorrelation for the $k$th parameter at lag $t$, and $T$ is the value at which the autocorrelations taper off (typically, 0.05 or 0.10).

# Chapter 8: Linear Regression and Extension

## Normal Linear Regression

The linear regression model is given by the following equation,

$$y = X\beta + \epsilon,$$

where $y = (y_1, ..., y_n)'$ is a column vector of dimension $n \times 1$, $X$ is a matrix of dimension $n \times k$ defined as,

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

$\beta = (\beta_1, \cdots, \beta_k)'$ is a $k \times 1$ column vector of unknown parameters, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$ is an $n \times 1$ vector of errors.

If we assume $\epsilon \sim N(0, \sigma^2 I_n)$, then it yields $y \sim N(X\beta, \sigma^2 I_n)$ which can be used to write the likelihood. We assume the following prior distribution on the parameters: $\beta \sim N_k(\beta_0, B_0)$ and $\sigma^2 \sim IG\big(\alpha_0/2, \delta_0/2\big)$.

The posterior distribution can be obtained as product of the likelihood and prior distributions. This yields the following expression:

$$\begin{aligned} \pi(\beta, \sigma^2|y) \propto (2\pi\sigma^2)^{-n/2} \exp\Big[ -\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\Big] \\ \times \exp\Big[ -\frac{1}{2}(\beta - \beta_0)' B_0^{-1}(\beta - \beta_0)\Big] \times \left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1} \exp\Big[ -\frac{\delta_0}{2\sigma^2}\Big]. \end{aligned} \tag{1}$$

The conditional posteriors are derived by collecting expression for one parameter at a time from the joint posterior (1), while holding the remaining parameters fixed. The kernel for $\beta$ can be written as,

$$\begin{aligned} \pi(\beta|\sigma^2, y) &\propto \exp\Big[ -\frac{1}{2}\Big\{(y - X\beta)'\sigma^{-2}(y - X\beta) + (\beta - \beta_0)' B_0^{-1}(\beta - \beta_0)\Big\}\Big] \\ &\propto \exp\Big[ -\frac{1}{2}\Big\{ -\sigma^{-2} y'X\beta - \sigma^{-2}\beta' X'y + \sigma^{-2}\beta' X'X\beta + \beta' B_0^{-1}\beta \\ &\qquad -\beta' B_0^{-1}\beta_0 - \beta_0' B_0^{-1}\beta\Big\}\Big], \end{aligned}$$

where we have omitted terms that do not involve $\beta$. Collecting terms involving $\beta$, we have,

$$\begin{aligned} \pi(\beta|\sigma^2, y) &\propto \exp\Big[ -\frac{1}{2}\Big\{\beta'(\sigma^{-2}X'X + B_0^{-1})\beta - \beta'(\sigma^{-2}X'y + B_0^{-1}\beta_0) - (\sigma^{-2}y'X + \beta' B_0^{-1})\beta\Big\}\Big] \\ &\propto \exp\Big[ -\frac{1}{2}\Big\{\beta' B_1^{-1}\beta - \beta' B_1^{-1}\bar{\beta} - \bar{\beta}' B_1^{-1}\beta\Big\}\Big] \\ &\propto \exp\Big[ -\frac{1}{2}\Big\{(\beta - \bar{\beta})' B_1^{-1}(\beta - \bar{\beta}) - \bar{\beta}' B_1^{-1}\bar{\beta}\Big\}\Big] \end{aligned}$$

$$\propto \quad \exp\Big[-\frac{1}{2}\Big\{(\beta-\bar{\beta})'B_1^{-1}(\beta-\bar{\beta})\Big\}\Big]$$

where the second line introduces the terms,

$$B_1 = \Big[\sigma^{-2}X'X + B_0^{-1}\Big]^{-1}, \quad \text{and} \quad \bar{\beta} = B_1\Big[\sigma^{-2}X'y + B_0^{-1}\beta_0\Big],$$

third line adds and subtracts $\bar{\beta}'B_1^{-1}\bar{\beta}$ inside the curly braces, and the last line follows by recognizing $\bar{\beta}'B_1^{-1}\bar{\beta}$ does not involve $\beta$ and can therefore be absorbed in the constant of proportionality. The result is the kernel of a Gaussian density and hence $\beta|\sigma^2, y \sim N(\bar{\beta}, B_1)$.

To derive the conditional posterior distribution for $\sigma^2$, we collect all terms involving $\sigma^2$ and derive as follows,

$$\begin{aligned}
\pi(\sigma^2|\beta,y) &\propto \left(\frac{1}{\sigma^2}\right)^{n/2}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right]\times\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\exp\left[-\frac{\delta_0}{2\sigma^2}\right].\\
&\propto \left(\frac{1}{\sigma^2}\right)^{(n/2+\alpha_0/2)+1}\exp\left[-\frac{1}{2\sigma^2}\big\{(y-X\beta)'(y-X\beta)+\delta_0\big\}\right]\\
&\propto \left(\frac{1}{\sigma^2}\right)^{(\alpha_1/2)+1}\exp\left[-\frac{\delta_1}{2\sigma^2}\right],
\end{aligned}$$

where in the last line we have introduced the following notations,

$$\alpha_1 = \alpha_0 + n, \quad \text{and} \quad \delta_1 = \delta_0 + (y-X\beta)'(y-X\beta).$$

The resulting expression for $\pi(\sigma^2|\beta, y)$ is recognized the kernel of an inverse Gamma distribution. Thus, we have $\sigma^2|\beta, y \sim IG(\alpha_1/2, \delta_1/2)$.

**Algorithm 6 (Gibbs Sampling for Normal Linear Regression)**

---

- Choose a starting value $\sigma^{2(1)}$
- At the g-th iteration, draw, $\beta \sim N_k(\bar{\beta}^{(g)}, B_1^{(g)})$, where

$$\begin{aligned}
B_1^{(g)} &= \big[\sigma^{-2(g-1)}X'X + B_0^{-1}\big]^{-1},\\
\bar{\beta}^{(g)} &= B_1^{(g)}\big[\sigma^{-2(g-1)}X'y + B_0^{-1}\beta_0\big],
\end{aligned}$$

  and then draw $\sigma^{2(g)} \sim IG\big(\alpha_1/2, \delta_1^{(g)}/2\big)$, where $\delta_1^{(g)} = \delta_0 + (y - X\beta^{(g)})'(y - X\beta^{(g)})$.
- Go to step (2) until g = B+G where B is burn-in sample and G is the desired sample size.

---

### Linear Regression with Conditionally Heteroskedastic Errors

In the Bayesian framework, heteroscedasticity can be introduced by mixing the variance of the error term with respect to gamma distribution. The linear regression model with conditionally heteroscedastic errors can be expressed as follows,

$$y_i = x_i'\beta + \epsilon_i, \qquad \epsilon_i|\lambda_i \sim N(0, \sigma^2\lambda_i^{-1}) \qquad \lambda_i \sim G(\nu/2, \nu/2).$$

We assume the following prior distributions: $\beta \sim N_k(\beta_0, B_0)$ and $\sigma^2 \sim IG\big(\alpha_0/2, \delta_0/2\big)$.

Once again, the joint posterior distribution is obtained as the product of the likelihood and prior distributions as,

$$\begin{aligned}\pi(\beta, \sigma^2, \lambda|y) &\sim \pi(\beta)\pi(\sigma^2)\prod_{i=1}^{n}\left\{\lambda_i^{\nu/2-1}\exp\left[-\frac{\nu\lambda_i}{2}\right]\right. \\ &\times \left.\left(\frac{\lambda_i}{\sigma^2}\right)^{1/2}\exp\left[-\frac{\lambda_i}{2\sigma^2}(y_i - x_i'\beta)^2\right]\right\}.\end{aligned}$$

To derive the conditional posteriors, we utilize the identity,

$$\sum_{i=1}^{n}\frac{\lambda_i}{2\sigma^2}\big(y_i - x_i'\beta\big)^2 = \frac{1}{2\sigma^2}\big(y - X\beta\big)'\Lambda\big(y - X\beta\big)$$

where $\Lambda = diag(\lambda_i)$. We next derive the conditional posterior distributions for the parameters $(\beta, \sigma^2, \{\lambda_i\})$.

(1) Conditional posterior distribution of $\beta$ is derived as follows:

$$\begin{aligned}\pi(\beta|\sigma^2, \lambda, y) &\propto \exp\left[-\frac{1}{2}(\beta - \beta_0)'B_0^{-1}(\beta - \beta_0)\right]\exp\left[-\frac{1}{2\sigma^2}\big(y - X\beta\big)'\Lambda\big(y - X\beta\big)\right] \\ &\propto \exp\left[-\frac{1}{2}\Big\{\beta'\big(\sigma^{-2}X'\Lambda X + B_0^{-1}\big)\beta - \beta'\big(\sigma^{-2}X'\Lambda y + B_0^{-1}\beta_0\big)\right. \\ &\qquad \left.-\big(y'\Lambda X + \beta_0'B_0^{-1}\big)\beta\Big\}\right] \\ &\propto \exp\left[-\frac{1}{2}\Big\{\beta'B_1^{-1}\beta - \beta B_1^{-1}\bar{\beta} - \bar{\beta}'B_1^{-1}\beta + \bar{\beta}'B_1^{-1}\bar{\beta} - \bar{\beta}'B_1^{-1}\bar{\beta}\Big\}\right] \\ &\propto \exp\left[-\frac{1}{2}\big(\beta - \bar{\beta}\big)'B_1^{-1}\big(\beta - \bar{\beta}\big)\right],\end{aligned}$$

where in the third line, we introduce the terms,

$$B_1^{-1} = \Big(\sigma^{-2}X'\Lambda X + B_0^{-1}\Big), \qquad \text{and} \qquad \bar{\beta} = B_1\Big(\sigma^{-2}X'\Lambda y + B_0^{-1}\beta_0\Big),$$

and add and subtract $\bar{\beta}'B_1^{-1}\bar{\beta}$ to complete the square in the fourth line. We recognize the last expression as the kernel of a multivariate normal distribution. As a result, we have $\beta|\sigma^2, \lambda, y \sim N(\bar{\beta}, B_1)$.

(2) The conditional posterior distribution of $\sigma^2$ is derived as follows:

$$\begin{aligned}\pi(\sigma^2|\beta,\lambda,y) &\propto \frac{\big(\delta_0/2\big)^{\alpha_0/2}}{\Gamma\big(\alpha_0/2\big)}\Big(\frac{1}{\sigma^2}\Big)^{\alpha_0/2+1}\exp\Big[-\frac{\delta_0}{2\sigma^2}\Big]\Big(\frac{1}{\sigma^2}\Big)^{n/2}\\ &\quad\times\exp\Big[-\frac{1}{2\sigma^2}\big(y-X\beta\big)'\Lambda\big(y-X\beta\big)\Big]\\ &\propto \Big(\frac{1}{\sigma^2}\Big)^{\alpha_0/2+n/2+1}\exp\Big[-\frac{1}{\sigma^2}\Big\{\frac{\big(y-X\beta\big)'\Lambda\big(y-X\beta\big)+\delta_0}{2}\Big\}\Big]\end{aligned}$$

The above expression is recognized as the kernel of an inverse gamma distribution. As such, we have,

$$\begin{aligned}\sigma^2|\beta,\lambda,y &\sim IG\Big(\alpha_1/2,\delta_1/2\Big), \qquad \text{where}\\ \alpha_1 &= \alpha_0+n,\\ \delta_1 &= \delta_0+\big(y-X\beta\big)'\Lambda\big(y-X\beta\big).\end{aligned}$$

(3) The conditional posterior distribution of $\lambda_i$ is derived as follows:

$$\begin{aligned}\pi(\lambda_i|\beta,\sigma^2,y) &\propto \lambda_i^{\nu/2-1}\exp\Big[-\frac{\nu\lambda_i}{2}\Big]\Big(\frac{\lambda_i}{\sigma^2}\Big)^{1/2}\exp\Big[-\frac{\lambda_i}{2\sigma^2}(y_i-x_i'\beta)^2\Big]\\ &\propto \lambda_i^{\nu/2+1/2-1}\exp\Big[-\lambda_i\Big\{\frac{\nu}{2}+\frac{(y_i-x_i'\beta)^2}{2\sigma^2}\Big\}\Big]\\ &\propto \lambda_i^{\nu_1/2-1}\exp\Big[-\lambda_i\Big\{\frac{\nu+\sigma^{-2}(y_i-x_i'\beta)^2}{2}\Big\}\Big].\end{aligned}$$

The above expression is recognized as the kernel of a Gamma distribution. Consequently, we have for $i=1,\cdots,n$,

$$\begin{aligned}\lambda_i|\beta,\sigma^2,y &\sim Gamma\Big(\nu_1/2,\nu_{2i}/2\Big), \qquad \text{where,}\\ \nu_1 &= \nu+1,\\ \nu_{2i} &= \nu+\sigma^{-2}(y_i-x_i'\beta)^2.\end{aligned}$$

To calculate the marginal likelihood for the linear regression with student $t$-error, we use the Chib method based on the identity,

$$f(y)=\frac{\prod f(y_i|\beta^*,\sigma^{2*})\pi(\beta^*)\pi(\sigma^{2*})}{\pi(\beta^*,\sigma^{2*}|y)},$$

where $f(y_i|\beta^*,\sigma^{2*})$ is the density of a student-$t$ distribution with mean $x_i'\beta^*$, scale parameter $\sigma^{2*}$ and $\nu_0$ degrees of freedom. To estimate the denominator, we use the decomposition,

$$\pi(\beta^*,\sigma^{2*}|y)=\pi(\beta^*|y)\pi(\sigma^{2*}|\beta^*,y).$$

The first term on the right hand side can be obtained as,

$$\begin{aligned}\hat{\pi}(\beta^*|y) &\approx \frac{1}{G}\sum_g N_k\big(\beta^*|\bar{\beta}^{(g)}, B_1^{(g)}\big), \qquad \text{where,}\\ B_1^{(g)} &= \Big[\sigma^{-2(g)}X'\Lambda^{(g)}X + B_0^{-1}\Big]^{-1},\\ \bar{\beta}^{(g)} &= B^{(g)}\Big[\sigma^{-2(g)}X'\Lambda^{(g)}y + B_0^{-1}\beta_0\Big],\end{aligned}$$

and $\sigma^{-2(g)}$ and $\Lambda^{(g)}$ are taken from the Gibbs output. The second term is estimated with the aid of reduced run. First generate a sample $\lambda^{(g)}$ from $\pi(\lambda|\beta^*, y)$ by successively sampling from $\pi(\lambda|\beta^*, \sigma^2, y)$ and $\pi(\sigma^2|\lambda, \beta^*, y)$. Then use the sample to compute,

$$\hat{\pi}(\sigma^{2*}|\beta^*, y) \approx \frac{1}{G}\sum IG\big(\sigma^{2*}|\alpha_1/2, \delta_1^{(g)}/2\big),$$

where $\delta_1^{(g)} = \delta_0 + (y - X\beta^*)'\Lambda^{(g)}(y - X\beta^*)$. The numerator can be computed directly. So we have our marginal likelihood.

### Tobit Model for Censored Data

In a censored data, the values of the response or dependent variable $y$ within a given range are repeated as a single value. For such data, we use the censored regression model, often called the *Tobit (Type I) model*. There are two primary examples of such data: *top coded data* and *corner solution outcomes*.

*Top Coded Data*: The values of $y_i$ are reported when $y_i \leq Y$; the value $Y$ is reported for observation $i$ if $y_i > Y$. This case arise as a result of the sampling scheme. An example is income data where income over some value, say of \$200,000 are reported as \$200,000. The observations are a mixture of data that are modeled continuously for $y_i \leq Y$ and a mass probability at the point $Y$. **In this model, it is assumed that the covariate vector $x_i$ is observed for all $i$.** Similarly, we can have *bottom coded censored data*.

*Corner Solution Outcome*: the values of $y_i$ are bounded by a constraint. For example, expenditure on durable goods are non-negative and the demand for tickets at a ball game is limited by the capacity of the stadium. In the former case, a large number of households report zero expenditure on durable goods and in the latter case, the capacity attendance is reported on sellout days.

*Censored data is different from Truncated data.* The top coded data is an example of censored data. In censored data models, it is assumed that the covariate vector $x_i$ is observed for all $i$. *Truncated data*: Data sets in which neither $y_i$ nor $x_i$ are observed when $y_i > Y$ are called *truncated data*.

A third type of data structure is known as *incidentally truncated data*. In this setup, $y_i$ and the selection variable $s_i$ have a joint distribution, $y_i$ is observed only when $s_i > 0$, and at least some of the $x_i$ are observed for all $i$.

### Tobit Model with Censoring from Below

The Tobit model with censoring from below can be represented in the latent variable formulation as follows:

$$\begin{aligned} z_i &= x_i'\beta + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \\ y_i &= \begin{cases} z_i, & \text{if } z_i > 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

The threshold of 0 is arbitrary and could be changed in a given application. Note that the error variance is identified, although poorly, if the degree of censoring is high.

The probability of '0' outcome is derived as,

$$\begin{aligned} P(y_i = 0) &= P(z_i \le 0) \\ &= P\big(\epsilon_i/\sigma \le -x_i'\beta/\sigma\big) \\ &= \int_{-\infty}^{-x_i'\beta/\sigma} \frac{1}{\sqrt{2\pi}} \exp\Big[-\frac{\epsilon_i^2}{2\sigma^2}\Big] d\epsilon_i \\ &= \Phi\Big(-\frac{x_i'\beta}{\sigma}\Big). \end{aligned}$$

The likelihood for the left-censored regression model can be expressed as,

$$\begin{aligned} f(y|\beta, \sigma^2) &= \prod_{i, y_i=0} \Phi\Big(-\frac{x_i'\beta}{\sigma}\Big) \prod_{i, y_i>0} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\Big] \\ &= \prod_{i, y_i=0} \Phi\Big(-\frac{x_i'\beta}{\sigma}\Big) \prod_{i, y_i>0} \frac{1}{\sigma}\, \phi\Big(\frac{y_i - x_i'\beta}{\sigma}\Big), \end{aligned}$$

where $\phi$ and $\Phi$ denotes the *pdf* and *cdf* of a standard normal distribution.

Using the following prior distributions: $\beta \sim N(\beta_0, B_0)$ and $\sigma^2 \sim IG\big(\alpha_0/2, \delta_0/2\big)$, the joint posterior can be written as,

$$\pi(\beta, \sigma^2|y) \propto f(y|\beta, \sigma^2)\; N(\beta|\beta_0, B_0)\; IG\big(\sigma^2|\alpha_0/2, \delta_0/2\big).$$

The above joint posterior distribution does not lead to a tractable conditional posterior density. So, we augment the likelihood with latent data $z$ and write the complete or augmented joint posterior as,

$$\begin{aligned} \pi(\beta, \sigma^2, z|y) &\propto \pi(\beta, \sigma^2) f(y, z|\beta, \sigma^2) \\ &\propto \pi(\beta)\pi(\sigma^2) f(z|\beta, \sigma^2) f(y|\beta, \sigma^2, z). \end{aligned}$$

The form of the prior distributions $\pi(\beta)$ and $\pi(\sigma^2)$ are known. The distribution of $z$ conditional on $(\beta, \sigma^2)$ is normal i.e. $z \sim N(X\beta, \sigma^2 I_n)$. For the last term $f(y|\beta, \sigma^2, z)$, we note that given a latent observation $z_i$, the values of $y_i$ is determined with certainty, regardless of $(\beta, \sigma^2)$. Thus,

we can write,

$$f(y|\beta,\sigma^2,z) \;=\; f(y|z) = \prod_{i=1}^{n}\Big\{I(y_i=0)I(z_i\le 0) + I(y_i=z_i)I(z_i>0)\Big\}.$$

The equality $f(y|\beta,\sigma^2,z) = f(y|z)$ is the key. It decouples the likelihood function from the main model parameters $(\beta,\sigma^2)$. Consequently, the augmented joint posterior takes the following form:

$$\begin{aligned}\pi(\beta,\sigma^2,z|y) \;\propto\; & \exp\Big[-\frac{1}{2}(\beta-\beta_0)'B_0^{-1}(\beta-\beta_0)\Big]\Big(\frac{1}{\sigma^2}\Big)^{\alpha_0/2+1}\exp\Big(-\frac{\delta_0}{2\sigma^2}\Big)\Big(\frac{1}{\sigma^2}\Big)^{n/2} \\ & \times\exp\Big[-\frac{1}{2\sigma^2}(z-X\beta)'(z-X\beta)\Big]\prod_{i=1}^{n}\Big\{I(y_i=0)I(z_i\le 0)+I(y_i=z_i)I(z_i>0)\Big\}.\end{aligned}$$

The conditional posteriors can be derived from the above joint posterior by collecting terms involving one parameters at a time while holding the remaining fixed. Following this procedure, we can obtain the conditional posterior distributions of $\beta$, $\sigma^2$ and $z$.

- The parameter $\beta|\sigma^2,z \sim N(\bar\beta, B_1)$, where,

$$B_1 \;=\; \Big[\sigma^{-2}X'X + B_0^{-1}\Big]^{-1}, \qquad \text{and} \qquad \bar\beta = B_1\Big[\sigma^{-2}X'z + B_0^{-1}\beta_0\Big].$$

- The parameter $\sigma^2|\beta,z \sim IG\Big(\alpha_1/2, \delta_1/2\Big)$, where,

$$\alpha_1 \;=\; \alpha_0 + n, \qquad \text{and} \qquad \delta_1 = \delta_0 + (z-X\beta)'(z-X\beta).$$

- To derive the conditional posterior for $z$, we consider the last two expressions of the joint posterior at the individual level i.e.,

$$f(z_i|\beta,\sigma^2,y) \;\propto\; \exp\Big[-\frac{1}{2\sigma^2}(z_i-x_i'\beta)^2\Big]\Big\{I(y_i=0)I(z_i\le 0)+I(y_i=z_i)I(z_i>0)\Big\},$$

  which immediately implies,

$$\begin{aligned} f(z_i|\beta,\sigma^2,y_i=0) \;&\propto\; \exp\Big[-\frac{1}{2\sigma^2}(z_i-x_i'\beta)^2\Big]I(z_i\le 0) \\ f(z_i|\beta,\sigma^2,y_i=z_i) \;&=\; 1.\end{aligned}$$

  Thus, we leave $z_i > 0$ cases untouched and draw $z_i \le 0$ cases from the normal density truncated from above at zero:

$$z_i|\beta,\sigma^2,y_i=0 \;\sim\; TN_{(-\infty,0]}(x_i'\beta,\sigma^2).$$

**Algorithm 7 (Gibbs Sampling for Tobit Model)**

---

- Draw $\beta^{(g)}$ from $N(\bar{\beta}^{(g)}, B_1^{(g)})$, where,

$$\begin{aligned} B_1^{(g)} &= \left[\sigma^{-2(g-1)} X'X + B_0^{-1}\right]^{-1}, \\ \bar{\beta}^{(g)} &= B_1^{(g)}\left[\sigma^{-2(g-1)} X' z^{(g-1)} + B_0^{-1}\beta_0\right]. \end{aligned}$$

- Draw $\sigma^{2(g)} \sim IG\big(\alpha_1/2, \delta_1^{(g)}/2\big)$, where,

$$\begin{aligned} \alpha_1 &= \alpha_0 + n, \\ \delta_1^{(g)} &= \delta_0 + (z^{(g-1)} - X\beta^{(g)})'(z^{(g-1)} - X\beta^{(g)}) \end{aligned}$$

- For $i = 1, 2, \ldots, n$, draw $z_i^{(g)} | \beta, \sigma^2, y_i = 0 \sim TN_{(-\infty,0]}\big(x_i'\beta^{(g)}, \sigma^{2(g)}\big)$.

---

## Binary Probit Model

The binary probit model is designed to deal with situations in which the outcome $y$ can take only one of two values, typically coded as 1 for success and 0 for failure.

Example 1: If $y$ denotes the status of union membership, then $y_i = 1$ can be used to denote union membership and $y_i = 0$ may denote person $i$ does not belong to a union.

Example 2: If $y$ denotes receiving a medical treatment, then $y_i = 1$ can be used to denote that subject $i$ receives a medical treatment and 0 if not.

The binary probit model can be expressed in terms of a latent variable as follows:

$$\begin{aligned} z_i &= x_i'\beta + \epsilon_i, \qquad \epsilon_i \sim N(0,1), \\ y_i &= \begin{cases} 1, & \text{if } z_i > 0, \\ 0, & \text{if } z_i \leq 0. \end{cases} \end{aligned}$$

The binary probit model has two identification requirements: anchoring the location and fixing the scale of the distribution. The location is fixed by setting the cut-point $\gamma_1 = 0$ (or at some constant $c$), and the scale restriction is achieved by assuming a fixed variance; for convenience $V(\epsilon) = \sigma^2 = 1$. Next, we derive the probability of each outcomes:

$$\begin{aligned} P(y_i = 0) &= P(z_i \leq 0) = P\big(\epsilon_i \leq -x_i'\beta\big) \\ &= \Phi(-x_i'\beta) = 1 - \Phi(x_i'\beta) \\ P(y_i = 1) &= P(z_i \geq 0) = P\big(\epsilon_i \geq -x_i'\beta\big) \\ &= P\big(\epsilon_i \leq x_i'\beta\big) \\ &= \Phi(x_i'\beta). \end{aligned}$$

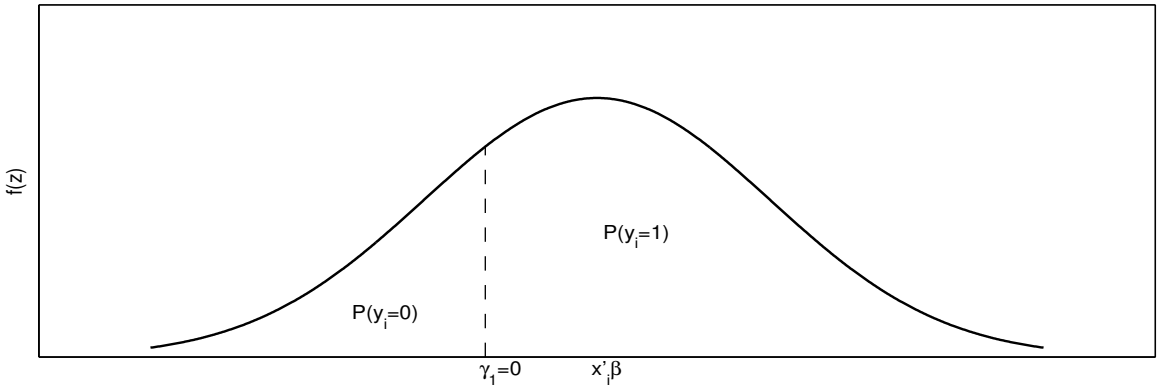


Figure 1: The cut-point $\gamma_1$ divides the area under the curve into two parts, the probability of failure and probability of success. Note that for each individual $i$ the mean $x_i'\beta$ and hence the probabilities, $\mathrm{P}(y_i = 0)$ and $\mathrm{P}(y_i = 1)$, will be different. Source: authors' creation.

The likelihood for the Probit model can therefore be expressed as,

$$f(y|\beta, X) = \prod_{i=1}^{n} \Big\{ \Big[\Phi\big(x_i'\beta\big)\Big]^{y_i} \Big[1 - \Phi\big(x_i'\beta\big)\Big]^{1-y_i} \Big\}.$$

The expression automatically picks the correct $\Phi$ term for each of the two possible outcomes.

Assuming a normal prior distribution $\beta \sim N(\beta_0, B_0)$, we arrive at the joint posterior distribution as the product of the likelihood and prior distribution as follows:

$$\pi(\beta|y) \quad \propto \quad \exp\Big[ -\frac{1}{2}(\beta - \beta_0)' B_0^{-1} (\beta - \beta_0) \Big] \prod_{i=1}^{n} \Big\{ \Big[\Phi\big(x_i'\beta\big)\Big]^{y_i} \Big[1 - \Phi\big(x_i'\beta\big)\Big]^{1-y_i} \Big\}.$$

It would be awkward to draw $\beta$ from the above expression. We can simplify things by using the technique of data augmentation. Therefore, we utilize the 'complete' or 'augmented data density' $f(y, z|\beta)$. The augmented joint posterior can be written as,

$$\begin{aligned}
\pi(\beta, z|y) &\propto \pi(\beta) f(y, z|\beta) \\
&\propto \pi(\beta) f(z|\beta) f(y|z, \beta) \\
&\propto \pi(\beta) f(z|\beta) f(y|z) \\
&\propto \exp\Big[ -\frac{1}{2}(\beta - \beta_0)' B_0^{-1} (\beta - \beta_0) \Big] \exp\Big[ -\frac{1}{2}\big(z - X\beta\big)'\big(z - X\beta\big) \Big] \\
&\quad \times \prod_{i=1}^{n} \Big\{ I(y_i = 0) I(z_i \leq 0) + I(y_i = 1) I(z_i > 0) \Big\}.
\end{aligned}$$

The conditional posteriors can be derived from the joint posterior distribution by collecting terms for one parameter while holding the remaining fixed.

(1) First, we derive the conditional posterior for $\beta$ as follows:

$$\pi(\beta|z,y) \quad \propto \quad \exp\Big[-\frac{1}{2}(\beta-\beta_0)'B_0^{-1}(\beta-\beta_0)\Big]\exp\Big[-\frac{1}{2}\big(z-X\beta\big)'\big(z-X\beta\big)\Big].$$

Opening the quadratic expressions, collecting terms involving $\beta$ and then completing the squares for $\beta$, yields the kernel of a normal distribution. Therefore, we have

$$\begin{aligned}\beta|z,y &\sim N(\bar{\beta},B_1), \qquad \text{where,}\\ B_1^{-1} &= \Big[X'X+B_0^{-1}\Big]\\ \bar{\beta} &= B_1\Big[X'z+B_0^{-1}\beta_0\Big].\end{aligned}$$

(2) We next consider the conditional posterior for $z$. For each observation $i$, we have the expression,

$$\pi(z_i|\beta,y_i)\propto\exp\Big[-\frac{1}{2}\big(z_i-x_i'\beta\big)^2\Big]\Big\{I(y_i=0)I(z_i\leq 0)+I(y_i=1)I(z_i>0)\Big\},$$

which immediately implies,

$$\begin{aligned}\pi(z_i|\beta,y_i=0) &\propto \exp\Big[-\frac{1}{2}\big(z_i-x_i'\beta\big)^2\Big]I(z_i\leq 0)\\ \pi(z_i|\beta,y_i=1) &\propto \exp\Big[-\frac{1}{2}\big(z_i-x_i'\beta\big)^2\Big]I(z_i>0).\end{aligned}$$

The above expressions are recognized as kernels of the truncated normal distribution. Thus, we can write,

$$\begin{aligned}z_i|\beta,y_i=0 &\sim TN_{(-\infty,0]}(x_i'\beta,1)\\ z_i|\beta,y_i=1 &\sim TN_{(0,\infty)}(x_i'\beta,1)\end{aligned}$$

*Posterior quantities of interest in the Probit Model*: There are two main posterior predictive constructs of interest in a Probit framework.

(1) The probability of 1 (or positive) outcome given a specific vector of explanatory variable $x_p$ can be written as,

$$\begin{aligned}\Pr(y_p=1) &= \Pr(z_p>0)\\ &= \int_\beta \Pr(z_p>0|\beta)\;\pi(\beta|z)\;d\beta\\ &= \int_\beta \Phi(x_p'\beta)\;\pi(\beta|z)\;d\beta.\end{aligned}$$

(2) In a nonlinear model, such as the binary probit model, the marginal effects need to be calculated as the coefficients do not represent the marginal effect. For the $j$th continuous

**Algorithm 8 (Gibbs Sampling for Binary Probit Model)**

---

- Choose a starting value $\beta^{(1)}, z^{(1)}$.
- At the $g$-th iteration, draw $\beta^{(g)}|z^{(g-1)}$ from $N(\bar{\beta}^{(g)}, B_1^{(g)})$, where,

$$\begin{aligned} B_1 &= \Big[X'X + B_0^{-1}\Big]^{-1}, \\ \bar{\beta}^{(g)} &= B_1\Big[X'z^{(g-1)} + B_0^{-1}\beta_0\Big]. \end{aligned}$$

- For $i = 1, 2, \ldots, n$, draw $z_i^{(g)}$ as follows:

$$z_i^{(g)}|\beta, y \sim \begin{cases} TN_{(-\infty,0]}(x_i'\beta^{(g)}, 1), & \text{if } y_i = 0, \\ TN_{(0,\infty)}(x_i'\beta^{(g)}, 1), & \text{if } y_i = 1. \end{cases}$$

---

regressor, the average marginal effect is calculated as,

$$\begin{aligned} \left\{\frac{\partial \Pr(y_p = 1)}{\partial x_{pj}}\right\} &= \int\int \beta_j \; \phi(x_p'\beta) \; \pi(x_{p,-j})\pi(\beta|y,z) \; d(x_{p,-j}) \, d\beta \\ &\simeq \frac{1}{nG}\sum_{g=1}^{G}\sum_{p=1}^{n}\Big[\beta_j^{(g)} \; \phi(x_p'\beta^{(g)})\Big], \end{aligned}$$

where $G$ is the number of Gibbs iteration post burn-in and $n$ is the number of observations in the sample. For the $j$th binary (dummy or indicator) variable, the average marginal effect is calculated as,

$$\begin{aligned} &\Big\{\Pr\Big(y_p = 1|x_{p,-j}, x_{p,j} = 1\Big) - \Pr\Big(y_p = 1|x_{p,-j}, x_{p,j} = 0\Big)\Big\} \\ &= \int\int \Big[\Phi\big(x_p'\beta|x_{p,j} = 1\big) - \Phi\big(x_p'\beta|x_{p,j} = 0\big)\Big] \; \pi(x_{p,-j})\pi(\beta|y,z) \; d(x_{p,-j}) \, d\beta \\ \simeq \; &\frac{1}{nG}\sum_{g=1}^{G}\sum_{p=1}^{n}\Big[\Phi\big(x_p'\beta^{(g)}|x_{p,j} = 1\big) - \Phi\big(x_p'\beta^{(g)}|x_{p,j} = 0\big)\Big]. \end{aligned}$$

Note that this calculation of marginal effect accounts for uncertainty in parameters and the covariates. Please read Jeliazkov and Vossmeyer (2018) for calculation of marginal effects.

### Binary Logit Model

The *binary logit model* can be expressed in the latent variable formulation as,

$$\begin{aligned} z_i &= x_i'\beta + \epsilon_i, \qquad \epsilon_i \sim \mathcal{L}(0,1), \\ y_i &= \begin{cases} 1, & \text{if } z_i > 0, \\ 0, & \text{if } z_i \leq 0, \end{cases} \end{aligned}$$

where $\mathcal{L}(0,1)$ denotes a logistic distribution with mean 0 and variance $\pi^2/3$. For the logit model,

$$P(y_i = 1) = P_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}},$$

which implies the link function is $G^{-1}(\cdot)$ where $G(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$. The model has the interesting property that,

$$\text{logit}(P_i) \equiv \log\left(\frac{P(y_i = 1)}{P(y_i = 0)}\right) = \log\left(\frac{P_i}{1 - P_i}\right) = x_i'\beta;$$

that is, the logarithm of the odds ratio is a linear function of the covariates.

The binary logit model can be estimated using the MCMC sampling algorithm proposed by Jeliazkov and Rahman (2012). In this method, we express the logistic distribution as a scale mixture of normals with respect to the Kolmogorov distribution (Andrews and Mallows, 1974; Poirier, 1978), we can write that,

$$f_L(s|\mu) = \int f_N(s|\mu, 4\kappa^2) f_K(\kappa) d\kappa \tag{2}$$

where $f_L(s|\mu)$ denotes the density of a random variable that has a logistic distribution around $\mu$ and variance $\pi^2/3$, and $f_K(\kappa)$ represents the Kolmogorov density,

$$f_K(\kappa) = 8\kappa \sum_{j=1}^{\infty} (-1)^{j+1} j^2 e^{-2j^2\kappa^2}.$$

This implies that if $\kappa_i$ has a Kolmogorov distribution, then conditionally on $\kappa_i$, $z_i|\kappa_i \sim N(x_i'\beta, 4\kappa_i^2)$, then marginally of $\kappa_i$, $z_i$ has logistic density $f_L(z_i|x_i'\beta)$.

Therefore, letting $\kappa = (\kappa_1, \ldots, \kappa_n)'$, we can consider the augmented posterior,

$$\begin{aligned} \pi(\beta, z, \kappa|y) &\propto f(y|\beta, z, \kappa) f(\beta, z, \kappa) \\ &= f(y|\beta, z, \kappa) f(z|\beta, \kappa) \pi(\beta)\pi(\kappa) \\ &= \left\{\prod_{i=1}^{n} f(y_i|z_i)\right\} f(z|\beta, \kappa) \pi(\beta)\pi(\kappa). \end{aligned} \tag{3}$$

where $f(y_i|z_i) = 1\{z_i \in \mathcal{B}_i\}$, $f(z|\beta, \kappa) = f_N(z|X\beta, K)$ with $K = \text{diag}(4\kappa^2)$, $\pi(\beta) = f_N(\beta|\beta_0, B_0)$, and $\pi(\kappa) = \prod_{i=1}^{n} f_K(\kappa_i)$.

The resulting Gibbs sampler for simulating from $\pi(\beta, z, \kappa|y)$ is constructed by sequentially drawing from the following full conditionals,

$$\beta|z, \kappa \sim N(\bar{\beta}, B_1),$$

with $B_1 = (B_0^{-1} + X'K^{-1}X)^{-1}$ and $\bar{\beta} = B_1(B_0^{-1}\beta_0 + X'K^{-1}z)$,

$$z_i|y, \beta, \kappa_i \sim TN_{\mathcal{B}_i}(x_i'\beta, 4\kappa_i^2), \qquad i = 1, \ldots, n,$$

and

$$\kappa_i|y,\beta,z_i \sim f(\kappa_i|z_i,\beta), \qquad i=1,\ldots,n, \tag{4}$$

where $f(\kappa_i|z_i,\beta)$ does not belong to a known family of distributions. However, a very convenient result can be obtained by representing this distribution in terms of Bayes' formula as,

$$\begin{aligned} f(\kappa_i|z_i,\beta) &= \frac{f(z_i|\beta,\kappa_i)f(\kappa_i)}{\int f(z_i|\beta,\kappa_i)f(\kappa_i)d\kappa_i} \\ &= \frac{f_N(z_i|x_i'\beta,4\kappa_i^2)f_K(\kappa_i)}{f_L(z_i|x_i'\beta)}. \end{aligned} \tag{5}$$

The last line in (5) follows by recognizing that the numerator densities are Gaussian and Kolmogorov, and the denominator, by equation (2), is simply the logistic density. Therefore, the unknown $f(\kappa_i|z_i,\beta)$ can now be represented very simply in terms of other well-known densities.

The fact that $f(\kappa_i|z_i,\beta)$ can be evaluated explicitly means that one can also evaluate the corresponding cdf,

$$F_{\kappa|z,\beta}(\kappa_i|z_i,\beta) = \int_0^{\kappa_i} f(s|z_i,\beta)ds.$$

In turn, $F_{\kappa|z,\beta}(\kappa_i|z_i,\beta)$ can be utilized to produce the draws needed in (4) by solving $\kappa_i = F^{-1}_{\kappa|z,\beta}(u)$, where $u \sim U(0,1)$ is a uniform random variable on the unit interval. The latter technique is known as the inverse cdf method and follows because

$$\Pr(\kappa_i \le a) = \Pr(F^{-1}_{\kappa|z,\beta}(u) \le a) = \Pr(u \le F_{\kappa|z,\beta}(a)) = F_{\kappa|z,\beta}(a).$$

This completes the proposed Gibbs sampling scheme for logit models.

### Binary *t*-link Model

The binary models can be extended to accommodate a Student-$t$ error distribution or *robit* model. The $t$-link or *robit* model can be written in the latent formulation as,

$$\begin{aligned} z_i &= x_i'\beta + \epsilon_i, \qquad \epsilon_i \sim T_\nu(0,1), \\ y_i &= \begin{cases} 1, & \text{if } z_i > 0, \\ 0, & \text{if } z_i \le 0, \end{cases} \end{aligned}$$

where $T_\nu(0,1)$ denotes a Students-$t$ distribution with mean 0 (for $\nu > 1$), variance $\frac{\nu}{\nu-2}$ (for $\nu > 2$), and $\nu$ degrees of freedom.

The $t$-link (robit) model can be estimated by extending the data augmentation approach. The discussion follows Albert and Chib (1993) and rests on the result that the $t$ distribution can be represented as a scale mixture of normals (Andrews and Mallows, 1974). Specifically, if for $i=1,\ldots,n$, $\lambda_i$ has a gamma distribution,

$$\lambda_i \sim G(\tau/2,\tau/2), \tag{6}$$

then conditionally on $\lambda_i$, we have

$$z_i|\lambda_i \sim N(x_i'\beta, 1/\lambda_i), \tag{7}$$

then marginally of $\lambda_i$, $z_i$ is distributed,

$$z_i \sim T_\tau(x_i'\beta, 1).$$

Therefore, letting $\lambda = (\lambda_1, \ldots, \lambda_n)'$, we can consider the augmented posterior,

$$\begin{aligned} \pi(\beta, z, \lambda|y) &\propto f(y|\beta, z, \lambda) f(\beta, z, \lambda) \\ &= f(y|\beta, z, \lambda) f(z|\beta, \lambda) \pi(\beta)\pi(\lambda) \\ &= \left\{\prod_{i=1}^{n} f(y_i|z_i)\right\} f(z|\beta, \lambda) \pi(\beta)\pi(\lambda), \end{aligned} \tag{8}$$

where $f(y_i|z_i) = 1\{z_i \in \mathcal{B}_i\}$ as before, $f(z|\beta, \lambda) = f_N(z|X\beta, \Lambda^{-1})$ with $\Lambda = \text{diag}(\lambda)$ which follows from (7), $\pi(\beta) = f_N(\beta|\beta_0, B_0)$ is the prior on $\beta$, and $\pi(\lambda)$ is given by the product of $n$ independent gamma densities stemming from (6)

$$\pi(\lambda) = \prod_{i=1}^{n} f_G(\lambda_i|\tau/2, \tau/2).$$

It is then quite straightforward to show that the Gibbs sampler for simulating from $\pi(\beta, z, \lambda|y)$ can be constructed by sequentially drawing from the full conditional distributions. The vector $\beta$ is sampled as,

$$\beta|z, \lambda \sim N(\bar{\beta}, B_1),$$

with $B_1 = (B_0^{-1} + X'\Lambda X)^{-1}$ and $\bar{\beta} = B_1(B_0^{-1}\beta_0 + X'\Lambda z)$. While the remaining parameters/variables are sampled element-wise as follows:

$$z_i|y_i, \beta, \lambda_i \sim TN_{\mathcal{B}_i}(x_i'\beta, \lambda_i^{-1}), \qquad i = 1, \ldots, n,$$

and

$$\lambda_i|y_i, \beta, z_i \sim G\left(\frac{\tau+1}{2}, \frac{\tau + (z_i - x_i'\beta)^2}{2}\right), \qquad i = 1, \ldots, n.$$

## Ordinal Probit Model

Ordinal models arise when the response variable has three or more categories which have a natural order. For example, a response to a question may range from 'strongly disagree' to 'strongly agree'. They have a natural order and we may assign numbers such as 1: strongly disagree, 2: disagree, 3: neither agree or disagree, 4: Agree, and 5: strongly agree. However, no cardinal interpretation can be attached to the numbers.

The ordinal probit model can be written as,

$$\begin{aligned} z_i &= x_i'\beta + \epsilon_i, \qquad \epsilon_i \sim N(0,1), \qquad \text{for } i = 1, \ldots, n \\ y_i &= j, \quad \text{if } \gamma_{j-1} < z_i \leq \gamma_j, \qquad \text{for } j = 1, 2, \ldots, J. \end{aligned}$$

where $J$ is number of categories, $\gamma_0 = -\infty$, and $\gamma_j = \infty$. We need two restrictions to identify the parameters of the ordinal model. The location is anchored by setting $\gamma_1 = 0$ and scale is fixed by assuming that the variance of the error term is 1. For alternative identification conditions, please see Jeliazkov et al. (2008) and Jeliazkov and Rahman (2012).

In the ordinal model, the probability of outcome $j$ can be derived to have the following expression:

$$P(y_i = j) = \Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta).$$

So the likelihood for the ordinal probit model can be expressed as follows:

$$f(y|\beta, \gamma) = \prod_{i=1}^{n} \prod_{j=1}^{J} \Big[\Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta)\Big]^{I(y_i=j)}.$$

While the parameters can be sampled easily, the sampling of $\gamma = (\gamma_1, \ldots, \gamma_{J-1})$ is not straightforward because of the ordering restrictions $\gamma_1 < \gamma_2 < \ldots < \gamma_{J-1}$. So, we utilize the

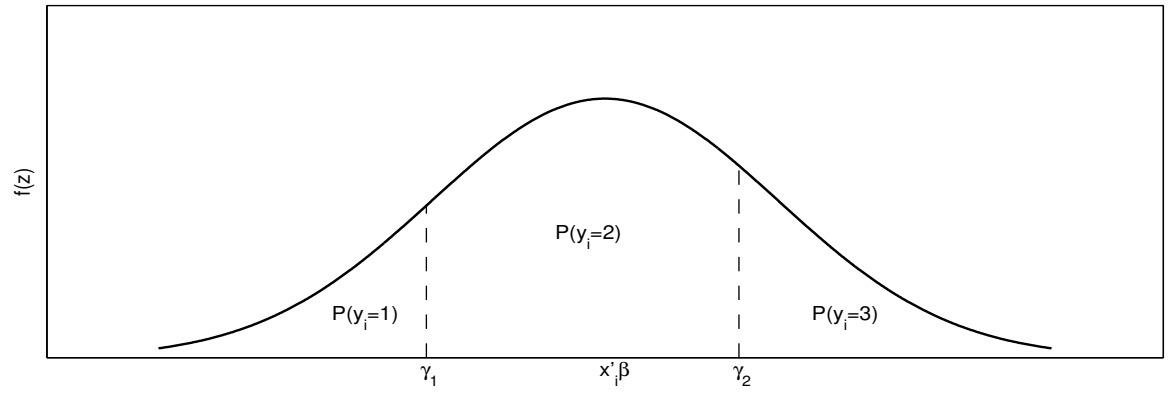


Figure 2: The two cut-points $(\gamma_1, \gamma_2)$ divide the area under the curve into three parts, with each part representing the probability of a response falling in the three response categories. Note that for each individual $i$ the mean $x_i'\beta$ will be different and so will be the category probabilities. Source: authors' creation.

**Algorithm 9 (MCMC Sampling for Ordinal Probit Model)**

---

1. Sample $\delta, z|y, \beta$ in one block as follows:

   (a) Sample $\delta|y, \beta$ marginally of $z$ by drawing $\delta' \sim q(\delta|y, \beta)$ from a student-$t$ proposal density $q(\delta|y, \beta) = f_T(\delta|\hat{\delta}, \hat{D}, \nu)$ where $\hat{\delta} = \arg\max f(y|\beta, \delta)\pi(\delta)$ and $\hat{D}$ is the inverse of the negative Hessian of $\ln\{f(y|\beta, \delta)\pi(\delta)\}$ evaluated at $\hat{\delta}$ and $\nu$ is the degrees of freedom parameter.

   Given the current value $\delta$ and the proposed draw $\delta'$, return $\delta'$ with probability,

   $$\alpha_{MH}(\delta, \delta') = \min\left\{1, \frac{f(y|\beta, \delta')\pi(\beta, \delta')f_T(\delta|\hat{\delta}, \hat{D}, v)}{f(y|\beta, \delta)\pi(\beta, \delta)f_T(\delta'|\hat{\delta}, \hat{D}, v)}\right\}$$

   otherwise, repeat the old value $\delta$. The parameter $\nu$ is set to a low number.

   (b) Sample $z|y, \beta, \delta$ by drawing $z_i|\beta, \gamma, y \sim TN_{(\gamma_{j-1}, \gamma_j)}(x_i'\beta, 1)$ for $i = 1, \ldots, n$ where $\gamma$ is obtained by the one-to-one mapping used to relate $\gamma$ and $\delta$.

2. Sample $\beta|z \sim N(\bar{\beta}, B_1)$ where $B_1 = (B_0^{-1} + X'X)^{-1}$ and $\bar{\beta} = B_1(B_0^{-1}\beta_0 + X'z)$

---

transformation,

$$\delta_j = \log(\gamma_j - \gamma_{j-1}) \qquad 2 \leq j \leq J-1,$$

which implies $\gamma_j = \sum e^{\delta_j}$. The original cut points can then be obtained by a one-to-one mapping between $\delta = (\delta_2, \delta_3, \ldots, \delta_{J-1})$ and $\gamma = (\gamma_1, \ldots, \gamma_{J-1})$, where recall $\gamma_1 = 0$.

Assuming normal prior distributions on $\beta \sim N(\beta_0, B_0)$ and $\delta \sim N(\delta_0, D_0)$, the augmented joint posterior distribution can be written as,

$$\begin{aligned}
\pi(\beta, \delta, z|y) &\propto f(y|\beta, \delta, z)\pi(z|\beta, \delta)\pi(\beta, \delta, ) \\
&\propto f(y|\beta, \delta, z)f(z|\beta)\pi(\beta)\pi(\delta) \\
&\propto \Big\{\prod_{i=1}^{n} f(y_i|z_i, \delta)\Big\}f(z|\beta)\pi(\beta)\pi(\delta) \\
&\propto \left\{\prod_{i=1}^{n} 1\big\{\gamma_{y_{i-1}} < z_i \leq \gamma_{y_{i-1}}\big\}\right\} N(z|X\beta, I_n)\, N(\beta|\beta_0, B_0)\, N(\delta|\delta_0, D_0).
\end{aligned}$$

The conditional posteriors can then be derived for $\beta$ and $z$ from the augmented joint posterior distribution. However, the cut-points $\delta$ do not have any tractable conditional posterior and needs to be sampled using the MH algorithm. The MCMC sampling for ordinal probit model is presented in Algorithm 9.

In step 1(a) instead of tailored MH, one may also use random walk as $\delta' = \delta + u$, $u \sim N(0_{J-2}, \iota^2\hat{D})$ where $\iota$ is a tuning parameter and $\hat{D}$ denotes negative Inverse-Hessian obtained by maximizing the log-likelihood with respect to $\delta$.

## References


Albert, J. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Andrews, D. F. and Mallows, C. L. (1974), "Scale Mixture of Distributions," *Journal of the Royal Statistical Society – Series B*, 36, 99–102.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S. and Jeliazkov, I. (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Jeliazkov, I. and Rahman, M. A. (2012), "Binary and Ordinal Data Analysis in Economics: Modeling and Estimation," in *Mathematical Modeling with Multidisciplinary Applications*, ed. X. S. Yang, pp. 123–150, John Wiley & Sons Inc., New Jersey.

Jeliazkov, I. and Vossmeyer, A. (2018), "The Impact of Estimation Uncertainty on Covariate Effects in Nonlinear Models," *Statistical Papers*, 59, 1031–1042.

Jeliazkov, I., Graves, J., and Kutzbach, M. (2008), "Fitting and Comparison of Models for Multivariate Ordinal Outcomes," *Advances in Econometrics: Bayesian Econometrics*, 23, 115–156.

Poirier, D. J. (1978), "A Curious Relationship between Probit and Logit Models," *Southern Economic Journal*, 40, 640–641.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1728.