Discrete Choice Models

Yuki Shiraito

POLSCI 798 Advanced Topics in Quantitative Methodology
University of Michigan

Winter 2019

Probit Models

- In social sciences, many variables are discrete
  - Likert-type survey items
  - Institutional features
  - Partisanship
  - etc...
- Probit regression model
  - Response of unit i is a Bernoulli random variable:

$$Y_i \overset{\text{indep.}}{\sim} \text{Bern}(p_i)$$

  - Generalized linear model approach:
    - Linear predictor and the probit link function:

$$p_i = \Phi\left(\mathbf{X}_i^\top \beta\right)$$

      where $\Phi(\cdot)$ is the standard Gaussian CDF
    - Parameter: β
- Prior distribution
  - Multivariate Gaussian prior: $\beta \sim \mathcal{N}\left(\beta_0, \Sigma_\beta\right)$

Posterior Density of Probit Model Parameters

- Posterior proportional to the prior times the likelihood:

$$
\begin{aligned}
p\left(\beta \mid \mathrm{Y}, \mathbf{X}\right) \propto\, & e^{-\frac{1}{2}\left(\beta-\beta_0\right)^\top \Sigma_\beta^{-1}\left(\beta-\beta_0\right)} \\
& \times \prod_{i=1}^{N}\left\{\Phi\left(\mathrm{X}_i^\top\beta\right)^{Y_i}\left(1-\Phi\left(\mathrm{X}_i^\top\beta\right)\right)^{1-Y_i}\right\} \\
\propto\, & e^{-\frac{1}{2}\left(\beta-\beta_0\right)^\top \Sigma_\beta^{-1}\left(\beta-\beta_0\right)} \\
& \times \left\{\prod_{i=1}^{N}\left(\int_{-\infty}^{\mathrm{X}_i^\top\beta}\frac{1}{\sqrt{2\pi}}e^{-\frac{p^2}{2}}dp\right)^{Y_i}\right. \\
& \left.\left(1-\int_{-\infty}^{\mathrm{X}_i^\top\beta}\frac{1}{\sqrt{2\pi}}e^{-\frac{p^2}{2}}dp\right)^{1-Y_i}\right\}
\end{aligned}
$$

- Computing posterior density of β
  - No simple form: β in integral bounds
  - Computationally expensive: Evaluation of integrals for values of β

Latent Variable Representation

- Latent variable representation of the probit regression model
  - Linear regression model for the latent response:

$$U_i = X_i^\top \beta + \varepsilon_i, \ \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$$

  - Observed response:

$$Y_i = \begin{cases} 0 & (U_i \leq 0) \\ 1 & (U_i > 0) \end{cases}$$

- Decision-theoretic interpretation
  - Individual's utility of two alternatives:

$$U_i(0) = X_i^\top \beta(0) + \eta_i, \ \eta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, .5)$$

$$U_i(1) = X_i^\top \beta(1) + \zeta_i, \ \zeta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, .5)$$

  - Utility difference:

$$\underbrace{U_i(1) - U_i(0)}_{=U_i} = X_i^\top \underbrace{(\beta(1) - \beta(0))}_{=\beta} + \underbrace{(\zeta_i - \eta_i)}_{\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)}$$

  - Decision: Choose 1 if $U_i(1) > U_i(0)$

Equivalence and Identifiability

- Symmetry of the standard Gaussian CDF: For any $a \in \mathbb{R}$,
$$\Phi(-a) = 1 - \Phi(a)$$
- Response probability in the latent variable representation:
$$\mathbb{P}(Y_i = 1 \mid \mathrm{X}_i) = \mathbb{P}(U_i > 0 \mid \mathrm{X}_i) = 1 - \Phi\left(-\mathrm{X}_i^\top \beta\right) = \Phi\left(\mathrm{X}_i^\top \beta\right)$$
- Equality of the likelihood functions:
$$p(Y_i \mid \mathrm{X}_i, \beta) = \Phi\left(\mathrm{X}_i^\top \beta\right)^{Y_i} \left(1 - \Phi\left(\mathrm{X}_i^\top \beta\right)\right)^{1-Y_i}$$
$$= \int_{-\infty}^{\infty} p(Y_i, U_i = u \mid \mathrm{X}_i, \beta)\, d\mathrm{u}$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{-\left(u - \mathrm{x}_i^\top \beta\right)^2}{2}} \left(1\{Y_i = 1 \wedge u > 0\} + 1\{Y_i = 0 \wedge u \leq 0\}\right) du$$
- Identification constraint: $\mathbb{V}(\varepsilon_i \mid X_i) = 1$
  - For any $r \in \mathbb{R}$, $U_i \leq 0 \Leftrightarrow rU_i \leq 0$
  - $U_i = \mathrm{X}_i^\top r\beta + \xi_i,\ \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, r^2) \rightsquigarrow$ identical likelihood
  - Can't estimate both β and $\mathbb{V}(\varepsilon_i \mid X_i)$

Data Augmentation

- Data augmentation
  - Model: $p\,(\text{Data} \mid \theta)$
  - Augmented variable: W with $p\,(\text{Data}, W \mid \theta)$ such that
    $$\int p\,(\text{Data}, W = w \mid \theta)\, dw = p\,(\text{Data} \mid \theta)$$
  - Posterior distribution:
    $$p\,(\theta \mid \text{Data}) \propto p(\theta) \underbrace{p\,(\text{Data} \mid \theta)}_{=\int p(\text{Data}, W=w \mid \theta) dw} \propto \int p\,(\theta, W = w \mid \text{Data})\, dw$$
  - Gibbs sampling with the augmented variable:
    1. Draw $W^{(s)}$ from $p(W \mid \text{Data}, \theta)$
    2. Draw $\theta^{(s)}$ from $p(\theta \mid \text{Data}, W)$

    $\rightsquigarrow$ MCMC sample from $p\,(\theta, W \mid \text{Data})$
- Gibbs sampler for the probit regression model

  $s = 0$ Set an arbitrary initial value of β

  $s = 1, 2, \ldots$ Repeat:
  1. Draw $U_i^{(s)}$ from the conditional posterior given $\beta^{(s-1)}$
  2. Draw $\beta^{(s)}$ from the conditional posterior given $U^{(s)}$

Gibbs Sampler for the Probit Model

- Joint posterior density:

$$p(\beta, \mathrm{U} \mid \mathrm{Y}, \mathbf{X}) \propto e^{-\frac{1}{2}(\beta-\beta_0)^\top \Sigma_\beta^{-1}(\beta-\beta_0)} \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}} e^{-\frac{-(U_i - \mathrm{x}_i^\top \beta)^2}{2}} \times \left(1\{Y_i = 1\}1\{U_i > 0\} + 1\{Y_i = 0\}1\{U_i \leq 0\}\right) \right\}$$

- Conditional posterior density of U_i given β:

$$p(U_i \mid \beta, Y_i = 0, \mathrm{X}_i) \propto \frac{1}{\sqrt{2\pi}} e^{-\frac{-(U_i - \mathrm{x}_i^\top \beta)^2}{2}} 1\{U_i \leq 0\}$$

$$p(U_i \mid \beta, Y_i = 1, \mathrm{X}_i) \propto \frac{1}{\sqrt{2\pi}} e^{-\frac{-(U_i - \mathrm{x}_i^\top \beta)^2}{2}} 1\{U_i > 0\}$$

  - Independent truncated Gaussians
    - Observations with $Y_i = 0$: $U_i^{(s)}$ drawn from $\mathcal{TN}_{(-\infty, 0]}\left(\mathrm{X}_i^\top \beta^{(s-1)}, 1\right)$
    - Observations with $Y_i = 1$: $U_i^{(s)}$ drawn from $\mathcal{TN}_{(0, \infty)}\left(\mathrm{X}_i^\top \beta^{(s-1)}, 1\right)$

- Conditional posterior density of β given U

$$p(\beta \mid \mathrm{U}, \mathrm{Y}, \mathbf{X}) \propto e^{-\frac{1}{2}(\beta-\beta_0)^\top \Sigma_\beta^{-1}(\beta-\beta_0)} \frac{1}{\sqrt{2\pi}} e^{-\frac{-(U_i - \mathrm{x}_i^\top \beta)^2}{2}}$$

  - Bayesian linear regression of U_i on X_i with $\sigma^2 = 1$

Ordered Probit Models

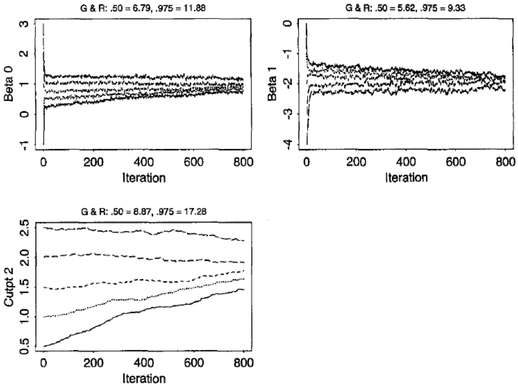
- E.g., Survey question about political efficacy (King et al., 2004):

  How much say do you have in getting the government to address issues that interest you?
  (5) Unlimited say, (4) A lot of say, (3) Some say,
  (2) Little say, (1) No say at all.

- Ordered probit regression model
  - Ordered response: $Y_i = 1, \dots, J$
  - Cumulative link model for Y_i
    - $\mathbb{P}\left(Y_i \leq j \mid \mathrm{X}_i, \beta, a\right) = \Phi\left(a_j - \mathrm{X}_i^\top \beta\right)$
    - Special case with $J = 2, a_1 = 0$: Probit model
  - Latent variable representation
    - $U_i = X_i^\top \beta + \varepsilon_i, \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
    - $Y_i = j \Leftrightarrow a_{j-1} < U_i \leq a_j$
    - *Identification constraint*: $a_0 = -\infty, a_1 = 0$, and $a_J = \infty$
  - Noninformative uniform prior on cutpoints: $p(a_j) \propto 1$

Slow Convergence

- Gibbs sampler for the ordered probit model
  - Additional Gibbs steps for $a_j, j = 2, \ldots, J$:
    - Conditional distribution of a_j given $\mathrm{U}, \beta, a_0, \ldots, a_{j-1}, a_{j+1}, \ldots, a_J$
    - $\text{Unif}\left[\max(\max\{U_i^{(s)} \mid Y_i = j\}, a_{j-1}^{(s)}), \min(\min\{U_i^{(s)} \mid Y_i = j+1\}, a_{j+1}^{(s)})\right]$
  - Problem: Slow mixing due to high autocorrelation



Cowles (1996) Fig. 2. Three-bin ordinal probit, univariate full conditionals, 800 iterations.

The Metropolis-Hastings Algorithm

- Caveats of the Gibbs sampler:
  - Slow convergence if parameters have high posterior correlation
  - Inefficiency if the model is not conditionally conjugate
- The Metropolis-Hastings (M-H) algorithm
  - Draws multiple parameters jointly
    $\rightsquigarrow$ reduce correlation across MCMC iterations
  - Incorporates an accept-reject step within a Markov chain
    $\rightsquigarrow$ Draws efficiently from a non-standard posterior distribution
- M-H update in iteration s:
  1. Draw a *proposal* θ_p from a *proposal distribution* $g(\theta^{(p)} \mid \theta^{(s-1)})$
  2. Calculate the *acceptance ratio*:
  $$\rho \equiv \frac{p(\theta^{(p)} \mid \text{Data})/g(\theta^{(p)} \mid \theta^{(s-1)})}{p(\theta^{(s-1)} \mid \text{Data})/g(\theta^{(s-1)} \mid \theta^{(p)})}$$
  3. Accept $\theta^{(p)}$ as $\theta^{(s)}$ with prob. $\min(r, 1)$; if reject, $\theta^{(s)} = \theta^{(s-1)}$
- Stationary distribution
$$\int \rho g(\theta^{(p)} \mid \theta^{(s-1)} = \theta) p(\theta \mid \text{Data}) d\theta = p(\theta^{(p)} \mid \text{Data})$$
- Gibbs as a special case of M-H with $\rho = 1$

M-H Algorithm for the Ordered Probit Model

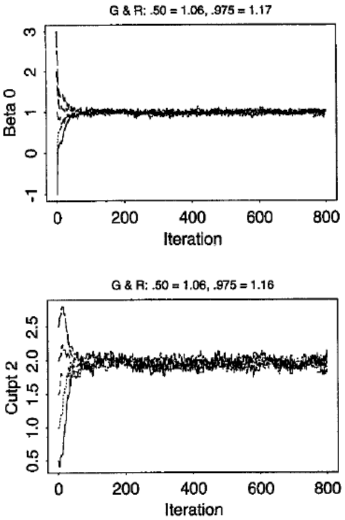
1. Set an arbitrary initial value of β and $a_2, \ldots, a_{J-1}$
2. Repeat for $s = 1, 2, \ldots$:
   1. Draw $U_i^{(s)}$ and $a_2^{(s)}, \ldots, a_{J-1}^{(s)}$ via an M-H step
      1. Propose $a_2^{(p)}, \ldots, a_{J-1}^{(p)}$ recursively by $\mathcal{TN}_{\left(a_{j-1}^{(p)}, a_{j+1}^{(s-1)}\right)}\left(a_j^{(s-1)}, \sigma_a^2\right)$
      2. Compute the acceptance ratio:

$$\rho = \prod_{j=2}^{J-1} \frac{\Phi\left(\frac{a_{j+1}^{(s-1)} - a_j^{(s-1)}}{\sigma_a}\right) - \Phi\left(\frac{a_{j-1}^{(p)} - a_j^{(s-1)}}{\sigma_a}\right)}{\Phi\left(\frac{a_{j+1}^{(p)} - a_j^{(p)}}{\sigma_a}\right) - \Phi\left(\frac{a_{j-1}^{(s-1)} - a_j^{(p)}}{\sigma_a}\right)}$$

$$\times \prod_{i=1}^{N} \frac{\Phi\left(a_{Y_i}^{(p)} - X_i^\top \beta^{(s-1)}\right) - \Phi\left(a_{Y_i-1}^{(p)} - X_i^\top \beta^{(s-1)}\right)}{\Phi\left(a_{Y_i}^{(s-1)} - X_i^\top \beta^{(s-1)}\right) - \Phi\left(a_{Y_i-1}^{(s-1)} - X_i^\top \beta^{(s-1)}\right)}$$

      3. Accept $a^{(p)}$ as $a^{(s)}$ and draw $U_i^{(s)} \sim \mathcal{TN}_{\left(a_{Y_i-1}^{(s)}, a_{Y_i}^{(s)}\right)}\left(X_i^\top \beta^{(s-1)}, 1\right)$ with probability $\min(\rho, 1)$; otherwise, do not update a and U_i
   2. Draw $\beta^{(s)}$ as if Bayesian linear regression of $U^{(s)}$ on X

Faster Convergence



Cowles (1996) Fig. 3. Three-bin ordinal probit, multivariate Hastings, 800 iterations.

Summary

- Binary and ordered probit regression models
  - Data augmentation with the latent response
  - Bayesian linear regression conditional on the augmented variable
- The Metropolis-Hastings Algorithm
  - Joint draws from non-standard joint posterior distributions
  - More efficient when posterior correlation is high

- Readings for review
  1. Data augmentation:
     - Albert and Chib (1993) "Bayesian Analysis of Binary and Polychotomous Response Data"
     - (Optional) van Dyk and Meng (2001) "The Art of Data Augmentation"
  2. The M-H algorithm:
     - **BDA3** Sections 11.2–3
     - (Optional) **BDA3** Chapter 12
     - (Optional) Cowles (1996) "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models"