Lecture Note 11: Bayesian Estimation of the MNP Model

Identification in the Probit Model

Continue with the probit model in latent variable form:

$$y_i^* = x_i'\beta + \epsilon_i.$$

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases}$$

Now suppose we don't normalize the variance of ϵ_i to 1:

$$\epsilon_i | X \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Then

$$\begin{aligned} Pr(y_i = 1|x, \beta, \sigma) &= Pr(y_i^* > 0|x, \beta, \sigma) \\ &= Pr(x_i'\beta + \epsilon_i > 0|x, \beta, \sigma) \\ &= Pr(\epsilon_i > -x_i'\beta|x, \beta, \sigma) \\ &= Pr(\epsilon_i < x_i'\beta|x, \beta, \sigma) \\ &= Pr(\epsilon_i/\sigma < x_i'(\beta/\sigma)|x, \beta, \sigma) \\ &= \Phi(x_i'(\beta/\sigma)). \end{aligned}$$

So the likelihood function is

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^{n} \Phi(x_i'(\beta/\sigma))^{y_i} \left[1 - \Phi(x_i'(\beta/\sigma))\right]^{1-y_i}.$$

Now consider another set of parameter values

$$(\tilde{\beta}, \tilde{\sigma}) = (a\beta, a\sigma)$$

for $a > 0$. Then

$$\mathcal{L}(\tilde{\beta}, \tilde{\sigma}) = \prod_{i=1}^{n} \Phi(x_i'(\tilde{\beta}/\tilde{\sigma}))^{y_i} \left[1 - \Phi(x_i'(\tilde{\beta}/\tilde{\sigma}))\right]^{1-y_i} = \prod_{i=1}^{n} \Phi(x_i'(\beta/\sigma))^{y_i} \left[1 - \Phi(x_i'(\beta/\sigma))\right]^{1-y_i} = \mathcal{L}(\beta, \sigma).$$

So β, σ are not identified: for any β, σ, there are "imposter" parameter values that give the same value for the likelihood.

Another way to see this is to note that for y_i^* defined as above, we could also multiply y_i^* by $a > 0$ and get the same implication for the observed choices. But $ay_i^* = x_i'(a\beta) + a\epsilon_i$, so we could equivalently consider a model with slope coefficient $a\beta$ and innovation variance $a^2\sigma^2$.

The usual approach is to normalize $\sigma = 1$. This removes the identification problem.

What if we don't normalize σ? Then there will not be a unique MLE (since for any β, σ that maximizes the likelihood, so will $c\beta, c\sigma$).

If we do a Bayesian analysis with a proper prior, then even though the likelihood has these "flat" regions, the posterior will still be a well-defined probability distribution. This is because a proper prior implies a proper joint distribution

$$p(\beta, \sigma, z) = p(\beta, \sigma) p(z|\beta, \sigma),$$

and hence a proper conditional distribution $p(\beta, \sigma|z)$.

In practice, though, the posterior distribution will be sensitive to the choice of the prior, because as we noted above, the likelihood function cannot distinguish between observationally equivalent sets of parameter values.

So it would be preferable to normalize $\sigma = 1$, unless you had very strong prior information about β and σ.

MNP Model

Now let's turn to the multinomial probit model. We'll work with a slightly simpler version of the model:

$$X_i = \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{iC} \end{pmatrix}.$$

$$u_i|X_i \sim N(X_i\beta, \Sigma).$$

Σ is a symmetric positive definite matrix, and all the distinct components of Σ are treated as free parameters.

As before, d_i is the $C \times 1$ vector of choice indicators.

Identification: there is a similar identification problem as in the probit model. Since scaling the vector u_i by $a > 0$ does not change the implications for the choice outcomes, and $au_i|X_i \sim N(X_i a\beta, a^2\Sigma)$, the likelihood function under β, Σ has the same value as the likelihood function under $a\beta, a^2\Sigma$:

$$\mathcal{L}(\beta, \Sigma) = \mathcal{L}(a\beta, a^2\Sigma).$$

The usual normalization is to set σ_{11}, the (1,1) element of Σ, equal to 1.

For Bayesian analysis, there are a couple of ways to proceed.

One way is to skip the normalization, and put a proper prior distribution on β and Σ. Let the prior for β and Σ be independent:

$$p(\beta, \Sigma) = p(\beta)p(\Sigma),$$

with

$$\beta \sim N(\bar{b}, A^{-1}),$$

$$\Sigma^{-1} \sim Wishart(v, V).$$

(As before, we are conditioning throughout on the X_i, so these should be thought of as priors conditional on the X_i.)

Wishart distribution: the $Wishart(v, V)$ distribution is a distribution on the space of symmetric, positive definite $C \times C$ matrices with density

$$p(G) = d|G|^{(v-C-1)/2} \exp\left(tr(-1/2GV\right),$$

where d is a normalizing constant. Here $v > 0$ and V is a symmetric positive definite matrix.

This has expectation

$$E[G|v, V] = vV^{-1}.$$

The parameter v roughly controls the precision of the distribution: smalller values of v lead to more diffuse distributions. So if we have some idea of a plausible value of Σ^{-1}, we can choose v and V to reflect the prior mean and then pin down v to reflect how precise our prior is. It can be useful to simulate values from the prior for a given v and V to check that it seems a good approximation to our prior beliefs.

The Wishart distribution is the conjugate prior for the variance in a multivariate normal model.

Usually we work with v an integer. To draw from the Wishart distribution when v is an integer and $v > C$, draw v draws $\alpha_1, \ldots, \alpha_v$ from the $N(0, V^{-1})$ distribution, and form $G = \sum_{j=1}^{v} \alpha_j \alpha_j'$.

back to the MNP model

Let u denote the vector of all latent utilities $\{u_i : i = 1, \ldots, n\}$, let d denote all the choice vectors $\{d_i\}$, and let X denote all the covariate matrices $\{X_i\}$.

The Gibbs sampler then cycles through the following conditional distributions:

1. Draw $\beta|\Sigma, u, d, X$.
2. Draw $\Sigma|\beta, u, d, X$.

3. For $i = 1, \dots, n$ and $c = 1, \dots, C$, draw

$$u_{ic} | u_{-ic}, \beta, \Sigma, d, X.$$

   Here u_{-ic} denotes all the latent utilities other than u_{ic}.

This approach turns out to work nicely because our choice of prior distributions. In particular, the full conditional for β is multivariate normal; we can draw for Σ by drawing Σ^{-1} from a certain Wishart distribution, and the draws for u_{ic} are truncated univariate normal. The exact form of these conditional distributions is given in McCulloch and Rossi (1994).[1]

However, this approach is not ideal, because it requires strong prior distributions to work well in practice. A better approach is to normalize $\sigma_{11} = 1$. But then it is less clear how to construct the prior in a way that leads to tractable full conditional distributions.

McCulloch, Polson, and Rossi suggest a clever reparametrization of the variance parameters.

Write

$$u_i = X_i \beta + \epsilon_i,$$

where $\epsilon_i \sim N(0, \Sigma)$ is a $C \times 1$ multivariate normal disturbance.

By the properties of the multivariate normal distribution, the marginal distribution of ϵ_{i1} (the first component of the vector ϵ_i) is

$$\epsilon_{i1} \sim N(0, \sigma_{11}),$$

and the conditional distribution of $(\epsilon_{i2}, \dots, \epsilon_{iC})'$ is

$$\begin{pmatrix} \epsilon_{i2} \\ \vdots \\ \epsilon_{iC} \end{pmatrix} | \epsilon_{i1} \sim N(\gamma/\sigma_{11} \cdot u_{i1}, \Sigma_2 - \gamma\gamma'/\sigma_{11}),$$

where γ is the vector of covariances between ϵ_{i1} and the elements of $(\epsilon_{i2}, \dots, \epsilon_{iC})'$ and Σ_2 is the joint covariance matrix of $(\epsilon_{i2}, \dots, \epsilon_{iC})'$.

Let $\Phi = \Sigma_2 - \gamma\gamma'/\sigma_{11}$.

So we can rewrite

$$\Sigma = \begin{bmatrix} \sigma_{11} & \gamma' \\ \gamma & \Phi + \gamma\gamma'/\sigma_{11} \end{bmatrix}.$$

[1] McCulloch, R., and P. Rossi, 1994, "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics* 64, 207-240.

Now normalize $\sigma_{11} = 1$. Then we get

$$\Sigma = \begin{bmatrix} 1 & \gamma' \\ \gamma & \Phi + \gamma\gamma' \end{bmatrix}.$$

So our new parameters are γ (a vector) and Φ (a symmetric PD matrix). Plus we have the parameter β as before.

We'll use the following priors:

$$\gamma \sim N(\bar{\gamma}, B^{-1}),$$

$$\Phi^{-1} \sim Wishart(\kappa, K),$$

and for β we can either use a normal prior or an improper uniform prior.

A Gibbs sampler can then be set up, which will cycle through the following steps:

1. Draw $\beta | \gamma, \Phi, u, d, X$.
2. For $i = 1, \ldots, n$ and $c = 1, \ldots, C$, draw

$$u_{ic} | u_{-ic}, \beta, \gamma, \Phi, d, X.$$

3. Draw $\gamma | \beta, \Phi, u, d, X$.
4. Draw $\Phi | \gamma, \beta, u, d, X$.

Steps 1 and 2 work exactly the same as before: given γ and Φ, we form Σ (which now incorporates the normalization) and proceed as in the previous case.

For step 3, notice that γ is the vector of regression coefficients in the multivariate regression model $(\epsilon_{i2}, \ldots, \epsilon_{iC})' | \epsilon_{i1} \sim N(\gamma \cdot u_{i1}, \Phi)$. It can be shown that γ therefore has a multivariate normal distribution.

For step 4, it can be shown that Φ^{-1} has a Wishart distribution.

For the specific forms of these distributions, see McCulloch, Polson, and Rossi.