

## Lecture Note 9: Multinomial Probit and Related Models

Rossi, McCulloch, and Allenby:

Analyzing household purchases of a good (tuna).

Couponing: at checkout, machine generates coupons based on

- current purchases
- demographic characteristics
- past history of purchases

How to tailor or target coupons to maximize profits?

How useful is purchase history?

The model is complicated, so we'll build it up in steps:

- Probit and Multinomial Probit
- MNP with taste heterogeneity
- Panel MNP with taste heterogeneity

### Identification in the Probit Model

Consider the probit model in latent variable form:

$$y_i^* = x_i' \beta + \epsilon_i.$$
$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Now suppose we don't normalize the variance of  $\epsilon_i$  to 1:

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

Then

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

$$\begin{aligned}
&= \Pr(\epsilon_i > -x'_i\beta|x, \beta, \sigma) \\
&= \Pr(\epsilon_i < x'_i\beta|x, \beta, \sigma) \\
&= \Pr(\epsilon_i/\sigma < x'_i(\beta/\sigma)|x, \beta, \sigma) \\
&= \Phi(x'_i(\beta/\sigma)).
\end{aligned}$$

So the likelihood function is

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

Now consider another set of parameter values

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

for  $a > 0$ . Then

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

So  $\beta, \sigma$  are not identified: for any  $\beta, \sigma$ , 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  $\sigma$ ? Then there will not be a unique MLE (since for any  $\beta, \sigma$  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  $\beta$  and  $\sigma$ .

### MNP Model

Individuals:  $i = 1, \dots, n$ .

Choices:  $c = 1, \dots, C$

Measured characteristics of choices:  $x_{i1}, \dots, x_{iC}$ , where each  $x_{ic}$  is a  $k \times 1$  vector.

Examples of characteristics: prices, distance of choice to consumer.

Utilities:  $u_i = (u_{i1}, \dots, u_{iC})'$  is generated according to:

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

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

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

Choice rule: pick  $c$  if  $u_{ic} \geq u_{ic'} \quad \forall c' \in \{1, \dots, C\}$ .

(Assume no ties, and only one object chosen.)

Let  $d_{ic} = 1$  if choice  $c$  is chosen. Then

$$\begin{aligned} E[d_{ic} | X_i, \beta, \Sigma] &= Pr[d_c = 1 | X_i, \beta, \Sigma] \\ &= Pr[u_{ic} \geq u_{i1}, \dots, u_{ic} \geq u_{iC} | X, \beta, \Sigma] \\ &= \int \dots \int 1(u_{ic} \geq u_{i1}, \dots, u_{ic} \geq u_{iC}) dN(u_{i1}, \dots, u_{iC} | X_i \beta, \Sigma), \end{aligned}$$

where  $dN(\cdot | \cdot, \cdot)$  means integration with respect to the density of the multivariate normal distribution.

Likelihood for individual  $i$ :

$$\prod_{c=1}^C Pr(d_{ic} = 1 | X_i, \theta)^{d_{ic}}.$$

Full-sample likelihood:

$$\prod_{i=1}^n \prod_{c=1}^C Pr(d_{ic} = 1 | X_i, \theta)^{d_{ic}}.$$

Log likelihood:

$$L(\theta) = \sum_{i=1}^n \sum_{c=1}^C d_{ic} \log Pr(d_{ic} = 1 | X_i, \theta).$$

ML Estimator:

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta).$$

First order conditions:

$$0 = \frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n \sum_{c=1}^C d_{ic} \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1 | X_i, \theta).$$

Note that

$$\sum_{c=1}^C Pr(d_{ic} = 1 | X_i, \theta) = 1.$$

So

$$\sum_{c=1}^C \frac{\partial}{\partial \theta} Pr(d_{ic} = 1 | X_i, \theta) = 0.$$

Also,

$$\left[ \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1 | X_i, \theta) \right] \cdot Pr(d_{ic} = 1 | X_i, \theta) = \frac{\partial}{\partial \theta} Pr(d_{ic} = 1 | X_i, \theta),$$

So

$$0 = \sum_{c=1}^C \frac{\partial}{\partial \theta} Pr(d_{ic} = 1 | X_i, \theta) \cdot Pr(d_{ic} = 1 | X_i, \theta),$$

and we can write the FOC alternatively as

$$0 = \sum_{i=1}^n \sum_{c=1}^C \left[ \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1 | X_i, \theta) \right] \{d_{ic} - Pr(d_{ic} = 1 | X_i, \theta)\}.$$

This is somewhat intuitive: the second term  $\{d_{ic} - Pr(d_{ic} = 1 | X_i, \theta)\}$  is the outcome minus its conditional mean given  $X_i$ .

To solve for the MLE, we need to be able to calculate  $Pr(d_{ic} = 1 | X_i, \theta)$  (and possibly its derivatives) for each  $i, c$  and different possible parameter values  $\theta$ .

However, the integral defining  $Pr(d_{ic} = 1 | X_i, \theta)$  does not have a simple closed-form expression.

For relatively few choices ( $C \leq 4$ ), there exist deterministic (nonrandom) numerical integration routines that are efficient, but these cannot be used when the number of choices is large.

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  $\beta, \Sigma$  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  $\sigma_{11}$ , the (1,1) element of  $\Sigma$ , 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  $\beta$  and  $\Sigma$ . Let the prior for  $\beta$  and  $\Sigma$  be independent:

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

with

$$\begin{aligned} \beta &\sim N(\bar{b}, A^{-1}), \\ \Sigma^{-1} &\sim \text{Wishart}(v, V). \end{aligned}$$

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

Let  $u$  denote the vector of all latent utilities  $\{u_i : i = 1, \dots, n\}$ , 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  $\beta$  is multivariate normal; we can draw for  $\Sigma$  by drawing  $\Sigma^{-1}$  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).<sup>1</sup>

---

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

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  $\epsilon_{i1}$  (the first component of the vector  $\epsilon_i$ ) is

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

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

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

where  $\gamma$  is the vector of covariances between  $\epsilon_{i1}$  and the elements of  $(\epsilon_{i2}, \dots, \epsilon_{iC})'$  and  $\Sigma_2$  is the joint covariance matrix of  $(\epsilon_{i2}, \dots, \epsilon_{iC})'$ .

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

So we can rewrite

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

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

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

So our new parameters are  $\gamma$  (a vector) and  $\Phi$  (a symmetric PD matrix). Plus we have the parameter  $\beta$  as before.

We'll use the following priors:

$$\begin{aligned} \gamma &\sim N(\bar{\gamma}, B^{-1}), \\ \Phi^{-1} &\sim \text{Wishart}(\kappa, K), \end{aligned}$$

and for  $\beta$  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  $i = 1, \dots, n$  and  $c = 1, \dots, C$ , 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  $\gamma$  and  $\Phi$ , we form  $\Sigma$  (which now incorporates the normalization) and proceed as in the previous case.

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

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

### Rossi, McCulloch, and Allenby

Next we want to build in heterogeneity in consumer “tastes,” and we observe multiple observations for consumer. Let  $z_i$  be taste predictors, such as demographic characteristics.

$$u_{it}|X_{it}, z_i, \beta_i, \Pi, V_\beta \sim N(X_{it}\beta_i, I),$$

(Variance matrix is set to identity for simplicity, but it would be desirable to relax this.)

$$\beta_i|X_i, z_i, \Pi, V_\beta \sim N(\Pi z_i, V_\beta).$$

Use a flat prior on  $\pi$  and a Wishart prior on  $V_\beta^{-1}$ .

Other aspects of the model are defined as before. This is now a hierarchical model.

Gibbs: simulate the full conditionals of

- $\beta$
- $u$
- $\Pi, V_\beta$ .

Target couponing: will discuss in lecture.

**Geweke, Gowrisankaran, and Town (2003)**

$i = 1, \dots, n$ : patients with pneumonia in LA County

$j = 1, \dots, J$ : hospitals in LA County

$x_i$ : a  $k \times 1$  vector of patient characteristics

$z_{ij}$ : a  $q \times 1$  vector of patient-hospital characteristics, including distance of patient  $i$  to hospital  $j$ .

$m_i$ : mortality indicator, =1 if patient dies.

$c_i$ :  $J \times 1$  vector of indicators for whether patient  $i$  was admitted to a hospital.

Model:

$$m_i^* = c_i' \beta + x_i' \gamma + \epsilon_i,$$

$$\epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1).$$

Here  $\beta = (\beta_1, \beta_2, \dots, \beta_J)'$ .

Interpretation: if patient  $i$  were randomly assigned to hospital  $j$ , then

$$Pr(m_i = 1) = \Phi(\beta_j + x_i' \gamma).$$

However, we suspect that hospital choice is not random, so that  $c_i$  is correlated with  $\epsilon_i$ .

Let

$$Z_i = \begin{pmatrix} z_{i1}' \\ \vdots \\ z_{iJ}' \end{pmatrix},$$

$$c_i^* = Z_i \alpha + \eta_i.$$

The hospital indicators are formed from the latent  $J \times 1$  vector  $c_i^*$  by

$$c_i = \begin{pmatrix} c_{i1} \\ \vdots \\ c_{iJ} \end{pmatrix}, \quad \text{with } c_{ij} = 1(c_{ij}^* \geq c_{ik}^* \forall k).$$

Normalize  $c_{iJ}^* = 0$ , so there are  $J - 1$  latent variables. Assume

$$\begin{pmatrix} \epsilon_i \\ \eta_{i1} \\ \vdots \\ \eta_{i,J-1} \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & \pi' \\ \pi & \Sigma \end{bmatrix}\right).$$

This allows correlation between the elements of  $\eta$  and  $\epsilon$ , which makes  $\eta$  and  $c_i$  correlated.

Exclusion restriction: variables like distance to hospital are assumed to affect hospital choice, but not have a direct effect or mortality given hospital choice. Thus there is “exogenous variation” in hospital choice.

Here  $J$  is large:  $J = 114$ . This means that  $\Sigma$  has 6441 free parameters. In order to obtain a tractable model, the authors make some simplifying assumptions on the form of  $\Sigma$ .

Since  $J$  is large, this also means that  $\beta = (\beta_1, \dots, \beta_J)'$  is high dimensional. It is useful to impose some simplifying structure and relate the  $\beta_j$  to characteristics of hospitals.

Let

$$\beta_j = \beta_0 + w_j' \lambda + u_j,$$

where the  $u_j$  are IID  $N(0, \sigma_\beta^2)$ , and  $w_j$  contains dummy variables for hospital size and type of ownership (public, private nonprofit, private for-profit, private teaching).

Some findings:

- smallest and largest hospitals have higher quality on average
- more seriously ill patients tend to go to higher-quality hospitals