Ec. 2120 Spring 2010

G. Chamberlain

LECTURE NOTE 13

LIKELIHOOD

1. INTRODUCTION

As in Note 6, we shall assume random sampling:

$$(Y_i, Z_i) \overset{\text{i.i.d.}}{\sim} F \qquad (i = 1, \dots, n).$$

This joint distribution implies a conditional distribution for Y_i conditional on $Z_i = z$. We specify a set of conditional distributions, indexed by a parameter θ, that contains this conditional distribution:

$$\mathrm{Prob}(Y_i \in B \,|\, Z_i = z) = \int_B f(y \,|\, z, \theta)\, dm(y) \quad \text{for some} \quad \theta \in \Theta.$$

For each point θ in the parameter space Θ, there is a conditional density $f(\,\cdot \,|\, \cdot, \theta)$. The distribution of Y_i conditional on $Z_i = z$ has density $f(\,\cdot \,|\, z, \theta)$ for some θ in the parameter space. This density is with respect to the measure m. The function $f(\,\cdot \,|\, \cdot, \cdot)$ is given. It is known as the *likelihood function* (for a single observation).

For example, consider the normal linear model in Note 7:

$$Y_i \,|\, Z_i = z \sim \mathcal{N}(\beta' x, \sigma^2),$$

where $x = g(z)$ for a given, known function g. The parameter is $\theta = (\beta, \sigma^2)$, the parameter space is $\Theta = \mathcal{R}^K \times \mathcal{R}_+$, and the likelihood function is

$$f(y \,|\, z, (\beta, \sigma^2)) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp[-\frac{1}{2\sigma^2}(y - \beta' x)^2].$$

The model asserts that for some $\beta \in \mathcal{R}^K$ and $\sigma^2 \in \mathcal{R}_+$ (the true values),

$$\text{Prob}(Y_i \in [c, d] \,|\, Z_i = z) = \int_c^d f(y \,|\, z, (\beta, \sigma^2))\, dy.$$

(Here the measure m is Lebesgue measure on $\mathcal{R}$.) So we have a complete description of the possible conditional distributions for Y_i conditional on Z_i. (This is a partial likelihood, since we do not specify a set of distributions for Z_i; the marginal distribution of Z_i is left unrestricted.)

The normality assumption in the normal linear model is restrictive because there is a great variety of shapes for the density of a continuous distribution. Also the constant conditional variance assumption is restrictive. The conditional mean assumption that $E(Y_i \,|\, Z_i = z) = \beta' x$ need not be restrictive, because, for example, $\beta' x$ could represent a high-order polynomial.

If Y_i is a binary dependent variable, then a flexible specification for $E(Y_i \,|\, Z_i = z)$ implies a flexible specification for the conditional distribution of Y_i conditional on Z_i, and we can be more confident that our likelihood function is well specified. The next section considers a binary dependent variable.

2. BINARY DEPENDENT VARIABLE

2.1 *Probit Approximation*

Suppose that the dependent variable Y_i takes on only two values, which we shall denote by 0 and 1. Then the conditional expectation function is also a conditional probability function:

$$E(Y_i \,|\, Z_i) = 1 \cdot \text{Prob}(Y_i = 1 \,|\, Z_i) + 0 \cdot \text{Prob}(Y_i = 0 \,|\, Z_i) = \text{Prob}(Y_i = 1 \,|\, Z_i).$$

As in Note 2, let $r(\cdot)$ denote the regression function:

$$r(z) = E(Y_i \,|\, Z_i = z)$$

(which does not depend upon i due to random sampling). Since $r(\cdot)$ is also a conditional probability function, it only takes on values in the interval $[0,1]$:

$$0 \leq r(\cdot) \leq 1.$$

As in Note 2, we can approximate the regression function by a linear predictor:

$$r(Z_i) \cong E^*(Y_i \mid X_{i1}, \ldots, X_{iK}) = \beta_1 X_{i1} + \ldots + \beta_K X_{iK} = \beta' X_i, \tag{1}$$

where X_{ik} is a given function of Z_i: $X_{ik} = g_k(Z_i)$. For example, we could use polynomial approximation; if Z_i is a scalar, we would have $X_{ik} = Z_i^{k-1}$. But since we know $r(\cdot)$ is between 0 and 1, we hope to get a better approximation (for a given number K of terms) by imposing this restriction.

We can impose the restriction by putting the linear approximation inside a given, known function that is bounded between 0 and 1. A popular choice is the probit function

$$\Phi(s) = \text{Prob}(W \leq s) \quad \text{where} \quad W \sim \mathcal{N}(0,1).$$

So $\Phi(\cdot)$ is the cumulative distribution function (cdf) for the standard normal distribution. $\Phi(\cdot)$ is strictly monotonic with

$$\lim_{s \to -\infty} \Phi(s) = 0, \quad \lim_{s \to \infty} \Phi(s) = 1.$$

Now we can define a *probit approximation* to the regression function:

$$\gamma = \arg \min_{a \in \mathcal{R}^K} E[r(Z_i) - \Phi(a' X_i)]^2, \tag{2}$$

where the $K \times 1$ vector X_i is obtained from a given function of Z_i: $X_i = g(Z_i)$. As before, if Z_i is scalar, we could use $X_{ik} = Z_i^{k-1}$. Then our probit approximation is

$$r(Z_i) \cong \Phi(\gamma_1 X_{i1} + \ldots + \gamma_K X_{iK}) = \Phi(\gamma' X_i). \tag{3}$$

Its advantage over the linear predictor approximation in (1) is that it is guaranteed to stay between 0 and 1.

Define the prediction error

$$U_i = Y_i - E(Y_i \,|\, Z_i) = Y_i - r(Z_i),$$

and note that $E(U_i \,|\, Z_i) = 0$. Then

$$\begin{aligned} E[Y_i - \Phi(a'X_i)]^2 &= E[r(Z_i) + U_i - \Phi(a'X_i)]^2, \\ &= E[r(Z_i) - \Phi(a'X_i)]^2 + E(U_i^2), \end{aligned}$$

since

$$E[(r(Z_i) - \Phi(a'X_i))U_i] = E[E[(r(Z_i) - \Phi(a'X_i))U_i \,|\, Z_i]] = E[(r(Z_i) - \Phi(a'X_i))E(U_i \,|\, Z_i)] = 0.$$

So an equivalent definition of γ is

$$\gamma = \arg \min_{a \in \mathcal{R}^K} E[Y_i - \Phi(a'X_i)]^2. \tag{4}$$

The sample analog of the population definition of γ in (4) suggests the following estimator:

$$\hat{\gamma} = \arg \min_{a \in \mathcal{R}^K} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \Phi(a'X_i)]^2. \tag{5}$$

This is a *nonlinear least-squares* estimator.

2.2 *Partial Predictive Effect*

Once we have estimates for the probit approximation to the conditional expectation function, we can obtain predictive effects as in Section 4 of Note 2. For example, with two variables in Z_i, we have the following partial predictive effect from comparing the conditional expectation evaluated at $Z_{i1} = c$ and $Z_{i1} = d$, with Z_{i2} held constant at e:

$$E(Y_i \,|\, Z_{i1} = d, Z_{i2} = e) - E(Y_i \,|\, Z_{i1} = c, Z_{i2} = e) \cong \Phi(\gamma' g(d,e)) - \Phi(\gamma' g(c,e)), \tag{6}$$

with $X_i = g(Z_i)$. We obtain an estimate of this partial predictive effect of Z_1 on Y by replacing γ by the estimate $\hat{\gamma}$.

2.3 *Logit Approximation*

Another popular choice is the logit function

$$G(s) = \frac{\exp(s)}{1 + \exp(s)}.$$

Like the probit function, $G(\cdot)$ is strictly monotonic with

$$\lim_{s \to -\infty} G(s) = 0, \quad \lim_{s \to \infty} G(s) = 1.$$

$G(\cdot)$ is the cdf for the standard logistic distribution. This distribution is symmetric about 0 with the same general shape as a normal distribution, but its variance does not equal 1. We can define a *logit approximation* by replacing Φ by G in (2), (3), and (4). The coefficient vector γ will be different, but using the logit γ with G replacing Φ in (6) gives an approximation to the predictive partial effect which is usually quite similar to the probit approximation. Whether the probit or logit approximation will be better (for a given choice of $X_i = g(Z_i)$) will vary from one data set to another. Usually it does not matter and it is rarely if ever an important issue to focus on.

2.4 *Heteroskedasticity*

When Y_i takes on only the values 0 and 1, the conditional variance is determined by the conditional expectation:

$$\mathrm{Var}(Y_i \,|\, Z_i) = E(Y_i^2 \,|\, Z_i) - [E(Y_i \,|\, Z_i)]^2 = E(Y_i \,|\, Z_i) - [E(Y_i \,|\, Z_i)]^2,$$

since $Y_i^2 = Y_i$. So

$$\mathrm{Var}(Y_i \,|\, Z_i) = r(Z_i)[1 - r(Z_i)].$$

This suggests an alternative estimator for γ. First get a preliminary estimate $\hat{\gamma}^{(1)}$ using nonlinear least squares in (5); then form

$$\hat{r}_i(Z_i) = \Phi(\hat{\gamma}^{(1)'} X_i);$$

then do *weighted* (nonlinear) least-squares, using the inverse of the estimated conditional variance as a weight:

$$\hat{\gamma} = \arg\min_{a \in \mathcal{R}^K} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \Phi(a' X_i])^2 / [\hat{r}(Z_i)(1 - \hat{r}(Z_i))]. \tag{7}$$

The intuition for (7) is that observations Y_i with higher variance (conditional on Z_i) are given less weight in the fitting criterion.

2.5 *Likelihood Function*

Assume that the probit approximation is exact: $\mathrm{Prob}(Y_i = 1 \,|\, Z_i) = \Phi(\gamma' X_i)$. Then the likelihood function is

$$f(y \,|\, z, \gamma) = \Phi(\gamma' x)^y [1 - \Phi(\gamma' x)]^{1-y} \quad \text{if} \quad y \in \{0, 1\},$$

with $x = g(z)$. If $y \notin \{0, 1\}$, then $f(y \,|\, z, \gamma) = 0$. The parameter space is $\Theta = \mathcal{R}^K$. The model asserts that for some $\gamma \in \mathcal{R}^K$ (the true value),

$$\mathrm{Prob}(Y_i \in B \,|\, Z_i = z) = \sum_{y \in B} f(y \,|\, z, \gamma),$$

where B is a subset of $\{0, 1\}$. (Here the measure m is counting measure.)

3. INFORMATION INEQUALITY

To simplify notation, let (Y, Z) be a random vector with the F distribution: $(Y, Z) \sim F$, so that $(Y, Z) \overset{d}{=} (Y_i, Z_i)$. Let $E_\theta(\cdot \,|\, z)$ denote conditional expectation based on the $f(\cdot \,|\, z, \theta)$ density. In particular,

$$E_\theta\big(\log[f(Y \,|\, z, \tilde{\theta})] \,|\, z\big) = \int \log[f(y \,|\, z, \tilde{\theta})] f(y \,|\, z, \theta)\, dm(y).$$

Claim 1. (Information Inequality) For all $\theta, \tilde{\theta} \in \Theta$,

$$E_\theta\big(\log[f(Y \,|\, z, \tilde{\theta})] \,|\, z\big) \le E_\theta\big(\log[f(Y \,|\, z, \theta)] \,|\, z\big).$$

Proof. Let Q denote the following random variable:

$$Q = f(Y \mid z, \tilde{\theta}) / f(Y \mid z, \theta).$$

By Jensen's inequality,

$$E_\theta\big(\log(Q) \mid z\big) \le \log\big(E_\theta(Q \mid z)\big).$$

Note that

$$\begin{aligned}
\log\big(E_\theta(Q \mid z)\big) &= \log \int \frac{f(y \mid z, \tilde{\theta})}{f(y \mid z, \theta)} f(y \mid z, \theta)\, dm(y) \\
&= \log \int f(y \mid z, \tilde{\theta})\, dm(y) \\
&= \log(1) = 0.
\end{aligned}$$

So

$$E_\theta\big(\log(Q) \mid z\big) = E_\theta\big(\log[f(Y \mid z, \tilde{\theta})] - \log[f(Y \mid z, \theta)] \mid z\big) \le 0,$$

which implies that

$$E_\theta\big(\log[f(Y \mid z, \tilde{\theta})] \mid z\big) \le E_\theta\big(\log[f(Y \mid z, \theta)] \mid z\big). \quad \diamond$$

4. MAXIMUM LIKELIHOOD IS CONSISTENT

The maximum-likelihood (ML) estimate of θ is

$$\hat{\theta} = \arg\max_{a \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i \mid Z_i, a). \tag{8}$$

By the information inequality, for any $a \in \Theta$,

$$\begin{aligned}
E\big(\log f(Y \mid Z, a)\big) &= E\Big(E_\theta\big(\log[f(Y \mid Z, a)] \mid Z\big)\Big) \\
&\le E\Big(E_\theta\big(\log[f(Y \mid Z, \theta)] \mid Z\big)\Big) \\
&= E\big(\log f(Y \mid Z, \theta)\big).
\end{aligned}$$

Under regularity conditions, we can obtain a *uniform law of large numbers*:

$$\sup_{a\in\Theta} \Big|\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i \mid Z_i, a) - E\big(\log f(Y \mid Z, a)\big)\Big| \xrightarrow{p} 0.$$

Then it can be shown that

$$\hat{\theta} = \arg\max_{a\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i \mid Z_i, a) \xrightarrow{p} \arg\max_{a\in\Theta} E\big(\log f(Y \mid Z, a)\big) = \theta.$$

5. LIMIT DISTRIBUTION FOR ML FROM GMM

The first-order condition for the maximization in (8) is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(Y_i \mid Z_i, \hat{\theta})}{\partial \theta} = 0.$$

Define

$$\psi((Y_i, Z_i), a) = \frac{\partial \log f(Y_i \mid Z_i, a)}{\partial \theta}.$$

This is known as the *score function*. We are going to use it as the moment function in GMM. Check the key condition:

$$\begin{aligned} E_\theta\Big(\frac{\partial \log f(Y_i \mid Z_i, \theta)}{\partial \theta} \mid Z_i = z\Big) &= \int [f(y \mid z, \theta)]^{-1} \frac{\partial f(y \mid z, \theta)}{\partial \theta} f(y \mid z, \theta)\, dm(y) \\ &= \frac{\partial}{\partial \theta} \int f(y \mid z, \theta)\, dm(y) \\ &= \frac{\partial}{\partial \theta} 1 = 0. \end{aligned}$$

So ψ is a a valid moment function.

Because $\dim(\psi) = \dim(\theta)$, we can set $\hat{D} = I$. So we have

Claim 1. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Lambda)$, with $\Lambda = \alpha\Sigma\alpha'$, and

$$\alpha = \Big[E[\frac{\partial \psi((Y_i, Z_i), \theta)}{\partial \theta'}]\Big]^{-1},$$

$$\Sigma = E[\psi((Y_i, Z_i), \theta)\psi((Y_i, Z_i), \theta)'].$$

6. INFORMATION EQUALITY

The argument used to show that the score function satisfies the key condition can be extended to show that

$$-E_\theta[\frac{\partial \psi((Y_i, Z_i), \theta)}{\partial \theta'} \mid Z_i = z] = E_\theta[\psi((Y_i, Z_i), \theta)\psi((Y_i, Z_i), \theta)' \mid Z_i = z].$$

This is known as the *information equality*. It implies that

$$-E[\frac{\partial \psi((Y_i, Z_i), \theta)}{\partial \theta'}] = E[\psi((Y_i, Z_i), \theta)\psi((Y_i, Z_i), \theta)'].$$

Hence $-\alpha = \Sigma^{-1}$. Combining this with Claim 1 gives

Claim 2. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$.

7. PANEL PROBIT

7.1 *Latent Variable Crossing a Threshold*

The cross-section probit model can be expressed in terms of a latent variable Y_i^* crossing a threshold: $Y_i = 1(Y_i^* \geq 0)$, with

$$Y_i^* = X_i'\beta + U_i, \quad U_i \mid Z_i \sim \mathcal{N}(0, \sigma^2).$$

Then we have

$$\begin{aligned} \mathrm{Prob}(Y_i = 1 \mid Z_i) &= \mathrm{Prob}(U_i/\sigma \geq -X_i'(\beta/\sigma) \mid Z_i) \\ &= 1 - \Phi(-X_i'(\beta/\sigma)) \\ &= \Phi(X_i'\gamma), \end{aligned}$$

with $\gamma = \beta/\sigma$; β and σ are not separately identified.

7.2 *Random Effects*

Now suppose we have panel data:

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iT} \end{pmatrix}, \quad Z_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{iT} \end{pmatrix},$$

with $Y_{it} = 0$ or 1. We assume random sampling for the cross-section units: (Y_i, Z_i) i.i.d. for $i = 1, \dots, N$. The probit random-effects model can be obtained from a normal random-effects model for a latent variable Y_{it}^*.

$$
\begin{aligned}
Y_{it}^* &= X_{it}'\beta + U_{it}, \\
U_{it} &= V_i + \epsilon_{it} \qquad (i = 1, \dots, N;\, t = 1, \dots, T),
\end{aligned}
$$

where, conditional on Z_i, V_i is independent of $(\epsilon_{i1}, \dots, \epsilon_{iT})$ and

$$
V_i \sim \mathcal{N}(0, \sigma_v^2), \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \quad (t = 1, \dots, T).
$$

(X_{it} is a given, known function of Z_i: $X_{it} = g_{it}(Z_i)$.)

Let $U_i' = (U_{i1}, \dots, U_{iT})$. Then

$$
U_i \,|\, Z_i \;\sim \mathcal{N}(0, \Omega)
$$

with

$$
\Omega = \sigma_v^2 1_T 1_T' + \sigma_\epsilon^2 I_T.
$$

(1_T is a $T \times 1$ vector of ones.)

We observe $Y_{it} = 1$ if $Y_{it}^* \geq 0$; otherwise we observe $Y_{it} = 0$. Since only the sign of Y_{it}^* is observed, we can just as well work with $\tilde{Y}_{it}^* = Y_{it}^*/\sigma_\epsilon$:

$$
\begin{aligned}
\tilde{Y}_{it}^* &= X_{it}'\frac{\beta}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon}V_i + \frac{1}{\sigma\epsilon}\epsilon_{it} \\
&= X_{it}'\alpha + \tilde{V}_i + \tilde{\epsilon}_{it},
\end{aligned}
$$

with

$$
\alpha = \frac{\beta}{\sigma_\epsilon}, \quad \sigma_{\tilde{v}}^2 = \frac{\sigma_v^2}{\sigma_\epsilon^2}, \quad \sigma_{\tilde{\epsilon}}^2 = 1,
$$

and

$$
Y_{it} = \begin{cases} 1, & \text{if } \tilde{Y}_{it}^* \geq 0; \\ 0, & \text{otherwise.} \end{cases}
$$

7.3 *Partial Effect*

In evaluating a partial effect, one possibility is to set $\tilde{V}_i = 0$, which is the mean, median, and mode of the $\tilde{V}_i$ distribution:

$$\text{Prob}(Y_{it} = 1 \,|\, X_{it} = s, \tilde{V}_i = 0) = \Phi(s'\alpha). \tag{9}$$

Another possibility is to average over the distribution of $\tilde{V}_i$:

$$E[\text{Prob}(Y_{it} = 1 \,|\, X_{it} = s, \tilde{V}_i)] = \int \Phi(s'\alpha + \tilde{v})h(\tilde{v})\, d\tilde{v}, \tag{10}$$

where h is the density for a $\mathcal{N}(0, \sigma^2_{\tilde{v}})$ distribution. Because $\tilde{V}_i$ is independent of X_{it}, we can use iterated expectations to evaluate (10):

$$\begin{aligned} E[E(Y_{it} \,|\, X_{it} = s, \tilde{V}_i)] &= E[E(Y_{it} \,|\, X_{it} = s, \tilde{V}_i) \,|\, X_{it} = s] \\ &= E(Y_{it} \,|\, X_{it} = s) \\ &= \text{Prob}((\tilde{V}_i + \tilde{\epsilon}_{it})/(\sigma^2_{\tilde{v}} + 1)^{1/2} \geq -s'\alpha/(\sigma^2_{\tilde{v}} + 1)^{1/2}) \\ &= \Phi(s'\alpha/(\sigma^2_{\tilde{v}} + 1)^{1/2}). \end{aligned} \tag{11}$$

Whether we use $\Phi(s'\alpha)$ from (9) or $\Phi(s'\alpha/(\sigma^2_{\tilde{v}} + 1)^{1/2})$ from (10) and (11) can make a big difference, because $\sigma_{\tilde{v}}$ can be arbitrarily large.

I think it is better to average over the distribution of $\tilde{V}_i$. When $\sigma_{\tilde{v}}$ is large, there is only a small fraction of the population with $\tilde{V}_i$ near 0.