

Econ 673: Microeconometrics

Chapter 5: Binary Choices

Outline

- Introduction
- Model Specification
 - Linear Probability Model
 - Logit and Probit
- Estimation
 - Nonlinear Least Squares
 - Maximum Likelihood
 - Simulated Maximum Likelihood
 - The Expectation-Maximization (E-M) Algorithm
 - A Bayesian Perspective

Readings

Required

- Greene, W. H., *Econometric Analysis*, 4th edition, Upper Saddle River, New Jersey: Prentice-Hall, Inc., Sections 19.1 through 19.5.
- Train, K., (2003), *Discrete Choice Methods with Simulation*, Cambridge, MA: Cambridge University Press, Ch. 3.

Binary Choice Models

The simplest of discrete choice models is one between two alternatives

$$y_i = \begin{cases} 1 & \text{if alternative A is chosen} \\ 0 & \text{if alternative B is chosen.} \end{cases}$$

- Examples abound in studies of
 - Labor
 - Union Membership
 - Education
 - Housing
 - Voting
 - Crime
 - Purchases of Consumer Durables
 - Marriage
 - Transportation
 - Technology Adoption
- Useful not just directly, but indirectly as well in estimation

Model Specification

The statistical issue in modeling our binary choice is the specification of the probability of observing choice A, conditioning on attributes of

- The decision maker
- The choices themselves

$$\Pr[y_i = 1|x_i] = F(x_i, \beta)$$

$$\Pr[y_i = 0|x_i] = 1 - F(x_i, \beta)$$

Binary Specification

The binary specification for y_i is convenient, since then

$$\begin{aligned} E[y_i|x_i] &= 1 \cdot \Pr[y_i = 1] + 0 \cdot \Pr[y_i = 0] \\ &= F[x_i, \beta] \end{aligned}$$

This suggests a simple regression model, where

$$\begin{aligned} y_i &= E(y_i|x_i, \beta) + [y_i - E(y_i|x_i, \beta)] \\ &= F(x_i, \beta) + \xi_i \end{aligned}$$

where

$$\xi_i \equiv y_i - E(y_i|x_i, \beta)$$

Covariance Matrix in Binary Choice Model

The corresponding variance also has a convenient form:

$$\begin{aligned} \text{Var}[y_i|x_i] &= E[y_i^2|x_i] - E[y_i|x_i]^2 \\ &= (1^2 \cdot \Pr[y_i = 1] + 0^2 \cdot \Pr[y_i = 0]) - (F[x_i, \beta])^2 \\ &= F[x_i, \beta](1 - F[x_i, \beta]) \end{aligned}$$

The issue then is specifying a functional form for $F(x_i, \beta)$

The Linear Probability Model

Early efforts to model binary choice decisions employed a simple linear probability model:

$$F[x_i, \beta] = \beta' x_i$$

So that:

$$\begin{aligned} y_i &= E[y_i|x_i] + y_i - E[y_i|x_i] \\ &= \beta' x_i + \xi_i \end{aligned}$$

with

$$\xi_i \equiv y_i - E[y_i|x_i]$$

Limitations of LPM

- Heteroskedasticity

$$\begin{aligned} \text{Var}[\xi_i | x_i] &= \text{Var}[y_i | x_i] \\ &= F[x_i, \beta](1 - F[x_i, \beta]) \\ &= \beta' x_i (1 - \beta' x_i) \end{aligned}$$

- Fitted probabilities need not lie in unit interval
- At best, a reduced form approach to model a discrete event, typically involving little in the way economics.

Feasible GLS for Linear Probability Model

- Goldberger (1962) suggested using FGLS to correct for heteroskedasticity problem.
- Limitations
 - FGLS fitted choice probabilities still need not lie in unit interval.
 - FGLS is not efficient, since errors are clearly not normal
 - FGLS variance estimates themselves need not be positive.

Truncated LPM

- Ruud (2000, p.749) suggests a variant of the LPM based on the truncated uniform distribution

$$F(x_i, \beta) = \begin{cases} 0 & \beta'x_i < 0 \\ \beta'x_i & 0 \leq \beta'x_i < 1 \\ 1 & 1 < \beta'x_i \end{cases}$$

- Heteroskedasticity remains
- Conditional variances go to zero for extremes

Latent Variable Models

A more systematic approach, attributed to Golberger (1964), is that the decision is driven by y_i^* , an unobserved *latent* variable; i.e.,

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

where

$$y_i^* = h(x_i, \beta) - \eta_i$$

$h(x_i, \beta)$ is referred to as the *index function* and $F(\cdot)$ denotes the cdf for η_i

Motivation for the Index Function

- The most common motivation is McFadden's RUM hypothesis
- Individual i receives utility from choosing alternative j

$$U_{ij} = U_j(x_{ij}, z_{ij})$$

where x_{ij} and z_{ij} denote, respectively, the observed and unobserved characteristics of both the individuals themselves and the alternatives they are choosing from.

- *Observability* is from the perspective of the analyst.
- Individual i is assumed to choose alternative A if

$$U_A > U_B$$

RUM Model (Cont'd)

- Analyst specifies a functional form for the representative agent, segmenting individual's utility into two components

$$U_{ij} = V_j(x_{ij}; \beta) + \varepsilon_{ij}$$

where

$$\varepsilon_{ij} = U_{ij}(x_{ij}, z_{ij}) - V_j(x_{ij}; \beta)$$

- Errors can arise from
 - Omitted variables
 - Measurement errors
 - Specification errors due to functional form choice.

Choice Probabilities

Individual choices are random from the analyst's perspective

$$V_A(x_{iA}; \beta) + \varepsilon_{iA} > V_B(x_{iB}; \beta) + \varepsilon_{iB}$$

\Rightarrow

$$[V_A(x_{iA}; \beta) - V_B(x_{iB}; \beta)] + (\varepsilon_{iA} - \varepsilon_{iB}) > 0$$

or

$$y_i^* = h(x_i, \beta) - \eta_i > 0,$$

where

$$h(x_i, \beta) \equiv [V_A(x_{iA}; \beta) - V_B(x_{iB}; \beta)]$$

$$\eta_i = \varepsilon_{iB} - \varepsilon_{iA}$$

$$= [U_{iB}(x_{iB}, z_{iB}) - V_B(x_{iB}; \beta)] - [U_{iA}(x_{iA}, z_{iA}) - V_A(x_{iA}; \beta)]$$

Error Specification

- Most models assume that the errors are independent across individuals and choice occasions (panel setting)
- Correlation may arise across
 - individuals if an omitted variable is correlated across individuals, such as weather conditions.
 - choice occasions or alternatives because omitted key explanatory variable. Example: in recreation demand boat ownership.

Probit

- The two most common error specifications yield the logit and probit models.
- The probit model results if the ε_{ij} 's are distributed as normal variates, so that $\eta_i \sim N(0, \sigma^2)$

$$\begin{aligned}\Pr[y_i = 1] &= \Pr[h(x_i, \beta) - \eta_i \geq 0] \\ &= \Pr[\eta_i < h(x_i, \beta)] \\ &= \Phi\left[\frac{h(x_i, \beta)}{\sigma}\right]\end{aligned}$$

where $\Phi[\cdot]$ denotes the standard normal cdf.

Identification in Probit Model

In most probit models, the index function $h(\cdot)$ is linear in its parameters, so that β and σ cannot be separately identified

$$\begin{aligned}\Pr[y_i = 1] &= \Phi\left[\frac{\beta'x_i}{\sigma}\right] \\ &= \Phi\left[\left(\frac{\beta}{\sigma}\right)'x_i\right] \\ &= \Phi[\gamma'x_i] \\ \gamma_k &\equiv \frac{\beta_k}{\sigma}\end{aligned}$$

Typically normalize $\sigma = 1$

Exceptions to Identification Problem

- There are cases in which the full set of parameters is identified; e.g., contingent valuation.
- Individuals are asked if they are willing to pay B_i for an environmental improvement
- Respondents are assumed to have a value function associated with this change:

$$WTP_i = \beta'x_i + \eta_i$$

Contingent Valuation (cont'd)

- If the errors are $\eta_i \sim N(0, \sigma^2)$, then probability of responding “yes” to the CV question becomes:

$$\begin{aligned}\Pr[Yes] &= \Pr[WTP_i \geq B_i] \\ &= \Pr[\beta'x_i + \eta_i \geq B_i] \\ &= \Pr[\eta_i \geq B_i - \beta'x_i] \\ &= \Pr[-\eta_i \leq \beta'x_i - B_i] \\ &= \Phi\left[\frac{\beta'}{\sigma}x_i - \frac{1}{\sigma}B_i\right]\end{aligned}$$

- Variation in B_i across individuals is essential to distinguishing σ from the constant term.

Why Estimate σ ?

- We are often interested in not just the mean WTP, but also its distribution in the population.
- Kurkalova, Kling, and Zhao (2001) have employed a similar model to estimate in the context of adopting conservation tillage.
- Analysis provides distribution of returns or costs associated with conservation tillage and, hence, the potential costs of encouraging its adoption.

Logit Model

- The standard logit model results if the errors are *iid* extreme value variates, with

$$F(\varepsilon_{ij}) = \exp(\exp(-\varepsilon_{ij}))$$

- This in turn yields

$$\begin{aligned} F(\eta_i) &= \Lambda(\eta_i) \\ &\equiv \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \end{aligned}$$

- Thus

$$\begin{aligned} \Pr[y_i = 1] &= \Pr[\eta_i < h(x_i, \beta)] \\ &= \Lambda[h(x_i, \beta)] \end{aligned}$$

Identification in the Logit Model

- As with probit model, parameters are identified only up to a scalar factor.
- Rather than normalizing the error variance to 1, we normalize it to the standard deviation of a standard logit variate

$$\varphi \equiv \frac{\pi}{\sqrt{3}} \approx 1.8$$

- Thus, we are obtaining relative parameter estimates only

$$\gamma_k = \frac{\beta_k}{\sigma/\varphi} = \varphi \frac{\beta_k}{\sigma}$$

- Logit and probit parameter estimates will differ because a different normalization is being used.

Estimation: Nonlinear Least Squares

- Ruud (2001, 751-752) outlines Nonlinear Least Squares approach

$$\hat{\beta}_{NLS} = \arg \min_{\beta} \left\{ \sum_{i=1}^N [y_i - F(\beta'x_i)]^2 \right\}$$

- This ignores known heteroskedasticity, but can use 2-stage feasible GLS approach similar to standard weighted LS:

$$\hat{\beta}_{WNLS} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left[\frac{y_i - F(\beta'x_i)}{\sqrt{F(\hat{\beta}'_{NLS}x_i)[1 - F(\hat{\beta}'_{NLS}x_i)]}} \right]^2 \right\}$$

Maximum Likelihood

- The standard approach to estimating discrete choice models
- Assuming independence across observations, the log-likelihood function becomes

$$L(y, X, \beta) = \sum_i [(y_i) \ln(F[\beta'x_i]) + (1 - y_i) \ln(1 - F[\beta'x_i])]$$

Maximum Likelihood (cont'd)

- If the underlying distribution $F(\cdot)$ is symmetric, then

$$1 - F(\beta'x_i) = F(-\beta'x_i)$$

and

$$\begin{aligned} L(y, X, \beta) &= \sum_i [(y_i) \ln(F[\beta'x_i]) + (1 - y_i) \ln(F[-\beta'x_i])] \\ &= \sum_i \ln(F[q_i \beta'x_i]) \\ &= L(q, X, \beta) \end{aligned}$$

where $q_i \equiv 2y_i - 1$

ML First Order Conditions

- The resulting first order conditions will typically be nonlinear

$$\begin{aligned}
 0 &= \frac{\partial L(y, X, \beta)}{\partial \beta} \\
 &= \sum_i \left\{ \frac{y_i f[\beta' x_i]}{F[\beta' x_i]} + \frac{-(1-y_i) f[\beta' x_i]}{1-F[\beta' x_i]} \right\} x_i \\
 &= \sum_i \left\{ \frac{y_i(1-F[\beta' x_i]) - (1-y_i)F[\beta' x_i]}{F[\beta' x_i](1-F[\beta' x_i])} \right\} f[\beta' x_i] x_i \\
 &= \sum_i \left\{ \frac{y_i - F[\beta' x_i]}{F[\beta' x_i](1-F[\beta' x_i])} \right\} f[\beta' x_i] x_i \\
 &\text{where } f(\cdot) = \frac{dF_i}{d(\beta' x_i)}
 \end{aligned}$$

Notes on ML First Order Conditions

- Similar to those of WNLS, except $\hat{\beta}_{ML}$ replaces $\hat{\beta}_{NLS}$ in constructing the weights

$$\begin{aligned}
 0 &= \sum_i \left\{ \frac{y_i - F[\hat{\beta}'_{WNLS} x_i]}{F[\hat{\beta}'_{NLS} x_i](1-F[\hat{\beta}'_{NLS} x_i])} \right\} f[\hat{\beta}'_{WNLS} x_i] x_i \\
 0 &= \sum_i \left\{ \frac{y_i - F[\hat{\beta}'_{ML} x_i]}{F[\hat{\beta}'_{ML} x_i](1-F[\hat{\beta}'_{ML} x_i])} \right\} f[\hat{\beta}'_{ML} x_i] x_i
 \end{aligned}$$

Notes on ML First Order Conditions (cont'd)

2. When the distribution is symmetrical, FONC reduce to

$$0 = \frac{\partial L(y, X, \beta)}{\partial \beta} = \sum_i \left\{ \frac{q_i f[\beta' x_i]}{F[q_i \beta' x_i]} \right\} x_i$$

3. For the logit model

$$f(\cdot) = \frac{d\Lambda_i(\beta' x_i)}{d(\beta' x_i)} = \Lambda_i(\beta' x_i)[1 - \Lambda_i(\beta' x_i)]$$

Notes on ML First Order Conditions (cont'd)

3. For the logit model (cont'd)

$$\begin{aligned} 0 &= \sum_i \left\{ \frac{y_i \Lambda(\beta' x_i) [1 - \Lambda(\beta' x_i)]}{\Lambda[\beta' x_i]} + \frac{-(1 - y_i) \Lambda(\beta' x_i) [1 - \Lambda(\beta' x_i)]}{1 - \Lambda[\beta' x_i]} \right\} x_i \\ &= \sum_i \{ y_i [1 - \Lambda(\beta' x_i)] - (1 - y_i) \Lambda(\beta' x_i) \} x_i \\ &= \sum_i \{ y_i - \Lambda(\beta' x_i) \} x_i \end{aligned}$$

If one of the x_i 's is a constant, then

$$0 = \sum_i \{ y_i - \Lambda(\beta' x_i) \} \Rightarrow \bar{y} = \frac{1}{N} \sum_i \Lambda(\beta' x_i)$$

Hessians

The corresponding Hessians are given by:

$$H = \frac{\partial^2 L}{\partial \beta \partial \beta'} = \begin{cases} -\sum_i \Lambda_i (1 - \Lambda_i) x_i x_i' & \text{for the logit model} \\ -\sum_i \lambda_i (\lambda_i + \beta' x_i) x_i x_i' & \text{for the probit model} \end{cases}$$

where

$$\lambda_i = \begin{cases} \frac{-\phi_i}{1 - \Phi_i} & y_i = 0 \\ \frac{\phi_i}{\Phi_i} & y_i = 1 \end{cases}$$

Both log-likelihood functions are globally concave for linear index function

Asymptotic Covariance Matrix

$$\text{Asy.Var}[\hat{\beta}_{ML}] = (-E[H])^{-1} \quad \text{where}$$

$$-E(H) = \begin{cases} \sum_i \Lambda_i (1 - \Lambda_i) x_i x_i' & \text{for the logit model} \\ \sum_i \left[\frac{\phi_i^2}{\Phi_i (1 - \Phi_i)} \right] x_i x_i' & \text{for the probit model} \end{cases}$$

Robust Covariance Matrix Estimation

- Many packages will provide *robust covariance matrix estimates* based on White's (1982a) sandwich estimator

$$Est.Asy.Var[\hat{\beta}] = \hat{H}^{-1} \hat{B} \hat{H}^{-1}$$

where

$$B = \sum_{i=1} g_i^2 x_i x_i'$$

and $g_i = y - \Lambda_i$ for the logit model and λ_i for the probit model

Robust Covariance Matrix Estimation (cont'd)

Greene (2000, p. 823) notes that this result is of limited use

- The aim is to correct for error specifications, due to say heteroskedasticity.
- The problem is that probit MLE is not consistent in the presence of such errors. Thus, "...the sandwich estimator provides an appropriate covariance matrix for an estimator that is biased in an unknown direction." (p. 824).

Estimation: Maximum Simulated Likelihood

Consider the general binary choice maximum likelihood problem

$$\begin{aligned} L(y, X, \beta) &= \sum_i [(y_i) \ln(F[\beta'x_i]) + (1 - y_i) \ln(1 - F[\beta'x_i])] \\ &= \sum_i [(y_i) \ln(P_1[\beta'x_i]) + (1 - y_i) \ln(1 - P_1[\beta'x_i])] \end{aligned}$$

where

$$\begin{aligned} P_1[\beta'x_i] &= \Pr[y_i = 1 | x_i, \beta] \\ &= \int_{-\infty}^{\beta'x_i} f(\eta_i) d\eta_i \end{aligned}$$

Maximum Simulated Likelihood (cont'd)

In binary choice settings, $P_1[\beta'x_i]$ is readily computed, however this is often not the case for more than two choices

Maximum simulation likelihood maximizes

$$SL(y, X, \beta) = \sum_i [(y_i) \ln(\hat{P}_1[\beta'x_i]) + (1 - y_i) \ln(1 - \hat{P}_1[\beta'x_i])]$$

where $\hat{P}_1[\beta'x_i]$ relies on a simulated integral; e.g.,

$$\hat{P}_1[\beta'x_i] = \frac{1}{R} \sum_{r=1}^R 1(\eta^r < \beta'x_i); \eta^r \sim N(0,1)$$

Relative Merits of MSL

- Relatively easy to implement

$$E(\hat{P}_1) = P_1 \text{ and } \text{Var}(\hat{P}_1) \rightarrow 0 \text{ as } R \rightarrow \infty$$

- Limitations

- The crude simulator
 - Is not differentiable with respect to the parameters of interest.
 - Has an unnecessarily large variance
 - There is a strictly positive probability of $P_1=0$ or $=1$
 - Requires R to increase as N increases
- Börsch-Supan and Hajivassiliou (1993) use Monte Carlo experiments to show that with smooth strictly bounded simulators, SML have small asymptotic bias and root mean squared errors.

Estimation: Estimation- Maximization (EM) Algorithm

- EM algorithm is useful when latent variables y_i^* underlie the observed data, y_i , as in binary choice models

- Let

$f(y_i | x_i, \beta)$ denote the pdf for the observed variable, y_i

$f(y_i^* | x_i, \beta)$ denote the pdf for the unobserved variable y_i^*

- The log-likelihood function is given by:

$$L = \sum_{i=1}^N \ln [f(y_i | x_i, \beta)]$$

EM Algorithm

At the start of iteration $t+1$, we have $\hat{\beta}_t$

E-Step: Form

$$H(\beta | \hat{\beta}^t, y, x) = E \left[\sum_{i=1}^N \ln f(y_i^* | \beta, y_i, x_i) \right].$$

Essentially filling in lost (latent) information via expectation

M-Step: Solve

$$\hat{\beta}^{t+1} = \arg \max_{\beta} H(\beta | \hat{\beta}^t, y, x).$$

In a number of applications, the M-step is straightforward

Sources

- Boyles, R., (1983) "On the Convergence of the EM Algorithm," *Journal of the Royal Statistical Society, B* 45(1): 47-50.
- Greene, W. H. (2000) *Econometrics Analysis* (4th edition) New York: MacMillan, pp. 193-196.
- Ruud, P., (1991) "Extensions of Estimation Methods Using the EM Algorithm," *Journal of Econometrics* 49: 305-341.
- Wu, D., (1983) "On the Convergence Properties of the EM Algorithm," *Annals of Statistics* 11: 95-103.

EM Example: Binary Probit Greene (2000, p. 194)

Let

$$y_i^* = \beta' x_i - \eta_i$$

with

$$\eta_i \sim N(0,1)$$

Then

$$\ln f(y_i^* | x_i, \beta) = \frac{1}{2} \left[\ln(2\pi) + (y_i^* - \beta' x_i)^2 \right]$$

and

$$L^* = \frac{-N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^N (y_i^* - \beta' x_i)^2.$$

The E-Step for Binary Probit

$$\begin{aligned} H(\beta | \hat{\beta}^t, y, x) &= \frac{-N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^N E \left[(y_i^* - \beta' x_i)^2 \mid \hat{\beta}^t, y_i, x_i \right] \\ &= \frac{-N}{2} \ln(2\pi) - \frac{1}{2} \left\{ \sum_{y_i=1} E \left[(y_i^* - \beta' x_i)^2 \mid \hat{\beta}^t, y_i^* > 0, x_i \right] \right. \\ &\quad \left. + \sum_{y_i=0} E \left[(y_i^* - \beta' x_i)^2 \mid \hat{\beta}^t, y_i^* \leq 0, x_i \right] \right\} \end{aligned}$$

where $y_i^* | \beta, x_i \sim N(\beta' x_i, 1)$

We need to compute moments from a truncated normal distribution

The E-Step for Binary Probit (cont'd)

$$E\left[\left(y_i^* - \beta'x_i\right)^2 \mid \hat{\beta}'^t, y_i^* > 0, x_i\right] = \left\{ E\left[\left(y_i^* - \beta'x_i\right) \mid \hat{\beta}'^t, y_i^* > 0, x_i\right] \right\}^2 + \text{Var}\left[\left(y_i^* - \beta'x_i\right) \mid \hat{\beta}'^t, y_i^* > 0, x_i\right]$$

$$E\left[\left(y_i^* - \beta'x_i\right) \mid \hat{\beta}'^t, y_i^* > 0, x_i\right] = E\left[y_i^* \mid \hat{\beta}'^t, y_i^* > 0, x_i\right] - \beta'x_i$$

$$\text{Var}\left[\left(y_i^* - \beta'x_i\right) \mid \hat{\beta}'^t, y_i^* > 0, x_i\right] = \text{Var}\left[y_i^* \mid \hat{\beta}'^t, y_i^* > 0, x_i\right]$$

Moments of a Truncated Normal Greene (2000) Theorem 20.2

Given $x \sim N(\mu, \sigma^2)$

$$E(x|truncation) = \mu + \sigma\lambda(\alpha)$$

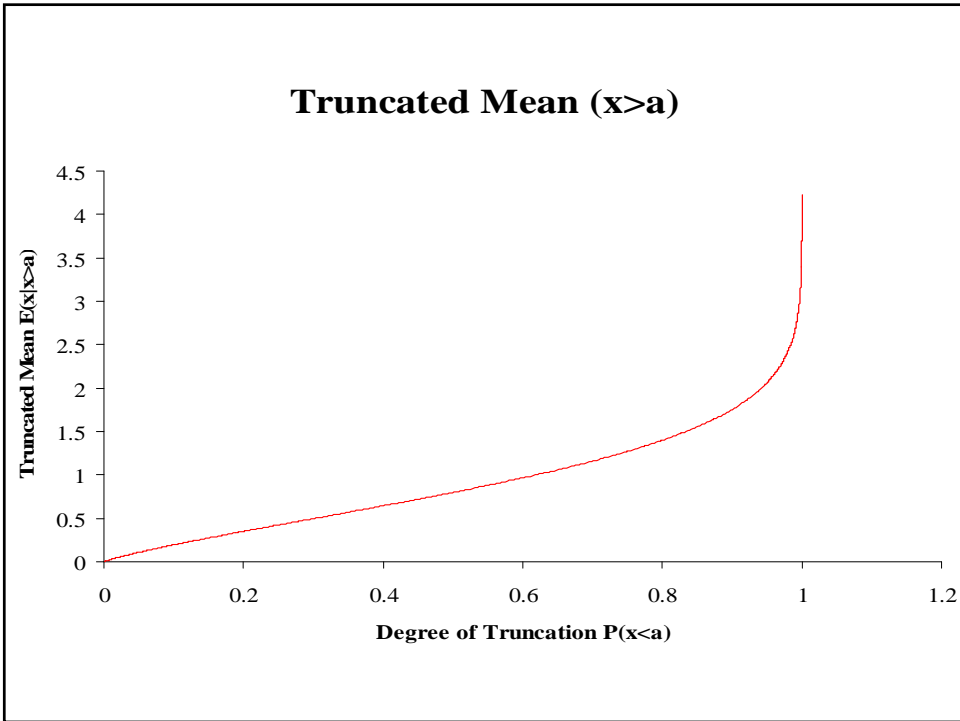
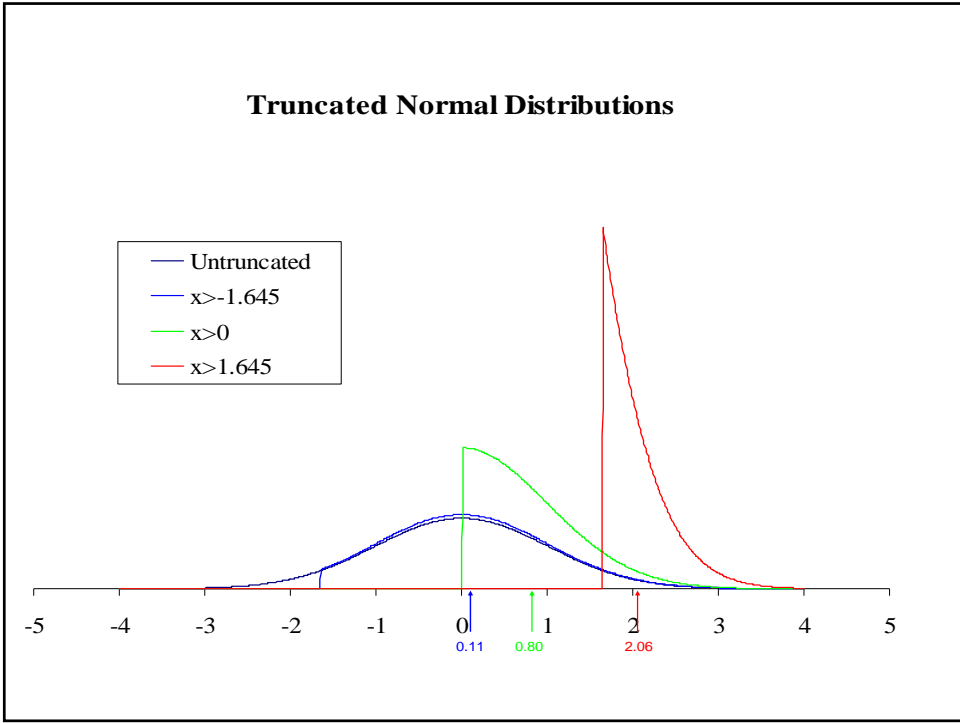
$$\text{Var}(x|truncation) = \sigma^2[1 - \delta(\alpha)]$$

where

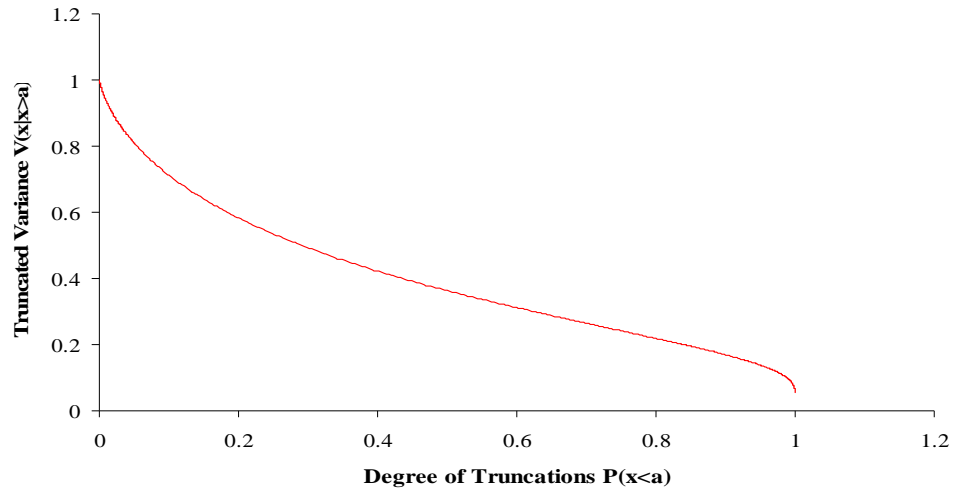
$$\alpha \equiv \frac{(a - \mu)}{\sigma}$$

$$\lambda(\alpha) = \begin{cases} \lambda^+(\alpha) \equiv \frac{\phi(\alpha)}{1 - \Phi(\alpha)} & \text{if truncation is } x > a \\ \lambda^-(\alpha) \equiv \frac{-\phi(\alpha)}{1 - \Phi(\alpha)} & \text{if truncation is } x < a \end{cases}$$

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$$



Truncated Variance ($x > a$)



The E-Step for Binary Probit (cont'd)

Applying Theorem 20.2

$$\begin{aligned}
 E\left[\left(y_i^* - \beta'x_i\right) \mid \hat{\beta}', y_i^* > 0, x_i\right] &= E\left[y_i^* \mid \hat{\beta}', y_i^* > 0, x_i\right] - \beta'x_i \\
 &= \hat{\beta}'x_i + \lambda_i^+\left(\hat{\beta}'x_i\right) - \beta'x_i \\
 &= \hat{y}_i^* - \beta'x_i \\
 \text{Var}\left[\left(y_i^* - \beta'x_i\right) \mid \hat{\beta}', y_i^* > 0, x_i\right] &= 1 - \lambda^+\left(\hat{\beta}'x_i\right)\left[\hat{\beta}'x_i + \lambda^+\left(\hat{\beta}'x_i\right)\right] \\
 &\equiv 1 - \delta^+\left(\hat{\beta}'x_i\right)
 \end{aligned}$$

Similar terms emerge for $y_i = 0$ terms

The E-Step for Binary Probit (cont'd)

Substituting back in these truncated means and variances yields

$$H(\beta | \beta^r, y, x) = \frac{-N}{2} \ln(2\pi) - \frac{1}{2} \left\{ \sum_{i=1}^N (\hat{y}_i^* - \beta' x_i)^2 \right\} - \frac{1}{2} \left\{ \sum_{i=1}^N \left[1 - \delta^+ \left(\hat{\beta}' x_i \right) \right] \right\}$$

Only the middle term is a function of β

M-Step

$$\begin{aligned} \beta^{t+1} &= \arg \max_{\beta} H(\beta | \hat{\beta}^t, y, x) \\ &= \arg \max_{\beta} \left(\frac{-N}{2} \ln(2\pi) - \frac{1}{2} \left\{ \sum_{i=1}^N (\hat{y}_i^* - \beta' x_i)^2 \right\} - \frac{1}{2} \left\{ \sum_{i=1}^N \left[1 - \delta^+ \left(\hat{\beta}' x_i \right) \right] \right\} \right) \\ &= \arg \min_{\beta} \left\{ \sum_{i=1}^N (\hat{y}_i^* - \beta' x_i)^2 \right\} \end{aligned}$$

i.e., M-step consists in this case of least squares:

$$\hat{\beta}^{t+1} = (X'X)^{-1} X' \hat{y}^* (\hat{\beta}^t)$$

Bayesian Analysis of Binary Choice (Data Augmentation)

- The E-M algorithm provides a classical approach for dealing with latent variables
- The Data Augmentation combined with Gibbs sampling provides a similar approach within the Bayesian framework
- Source:

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

The Basic Bayesian Paradigm

Given a model of the form:

$$y = f(X; \beta)$$

If our prior beliefs about β take the form $p(\beta)$

and the likelihood function takes the form $p(y|X, \beta)$

Then available data is used to update the prior via Bayes rule :

$$p(\beta|y, X) \propto p(\beta)p(y|X, \beta)$$

Binary Choice Posterior

In our standard binary choice model we observe the discrete choice variable y_i , specifying

$$\Pr[y_i = 1 | x_i] = F(\beta' x_i)$$

The likelihood function becomes:

$$\prod_{i=1}^N F(\beta' x_i)^{y_i} [1 - F(\beta' x_i)]^{1-y_i}$$

Thus, the posterior density, given a prior of $p(\beta)$ is

$$p(\beta | y, X) \propto p(\beta) \prod_{i=1}^N F(\beta' x_i)^{y_i} [1 - F(\beta' x_i)]^{1-y_i}$$

which can be difficult to characterize

Albert and Chib (1993)

Suggest resolving the problem of characterizing posterior distribution by combining

- Data augmentation and
- Gibbs sampling

Generic Gibbs Sampling

- Suppose we have a parameter vector θ with a probability density function $p(\theta)$
- While the full pdf may be difficult to draw from, suppose that the parameter vector can be partitioned into K distinct subvectors, $\{\theta_1, \dots, \theta_K\}$, with known conditional density functions

$$p(\theta_k | \theta_j \forall j \neq k)$$

- The sequence

$$\left\{ (\theta_1^t, \theta_2^t, \dots, \theta_K^t) \text{ where } \theta_k^t \sim p(\theta_k | \theta_1^t, \dots, \theta_{k-1}^t, \theta_{k+1}^{t-1}, \dots, \theta_K^{t-1}) \right\}$$

$$\xrightarrow{d} p(\theta_1, \theta_2, \dots, \theta_K)$$

Gibbs Sampling – Step by Step

- Start with initial values, $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_K^{(0)}$
- Simulate in turn for $t=1, \dots, T$

$$\begin{array}{l} \theta_1^{(t)} \text{ drawn from } p(\theta_1 | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_K^{(t-1)}) \\ \theta_2^{(t)} \text{ drawn from } p(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_K^{(t-1)}) \\ \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ \theta_K^{(t)} \text{ drawn from } p(\theta_K | \theta_1^{(t)}, \theta_3^{(t)}, \dots, \theta_{K-1}^{(t)}) \end{array}$$

One can use

$$\theta^{(m)} = (\theta_1^{(T,m)}, \theta_2^{(T,m)}, \dots, \theta_K^{(T,m)}), m = 1, \dots, M$$

$$\theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \dots, \theta_K^{(m)}), m = T_0 + k\Delta T$$

Data Augmentation for the Probit Model

- Introduce latent variable as unknown variable, with

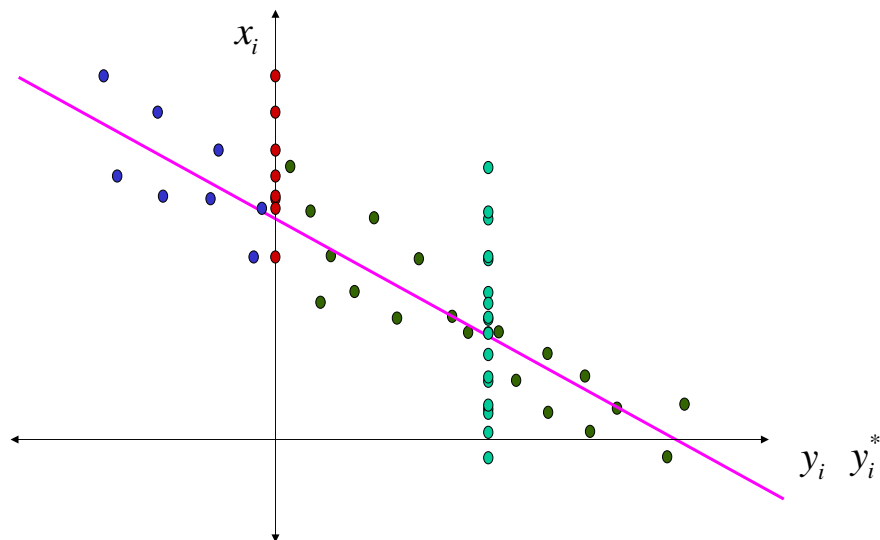
$$y_i^* = \beta' x_i - \eta_i; \eta_i \sim N(0,1)$$

which is unknown, much like the parameter vector β

- The joint posterior density for β and y_i^* is given by

$$\begin{aligned} p(\beta, y^* | y, X) &\propto p(\beta, y^*) p(y | y^*, \beta, X) \\ &= p(\beta) p(y^* | \beta, X) p(y | y^*, \beta, X) \\ &= p(\beta) \prod_{i=1}^N \{1(y_i^* \geq 0)1(y=1) + 1(y_i^* < 0)1(y=0)\} \phi(y_i^* | \beta, X) \end{aligned}$$

Basic Insight of Data Augmentation



Gibbs Sampling

While the joint posterior of β and y^* is complicated, the conditional posterior distributions are not

$p(y^*|\beta, y, X)$ is a truncated normal distribution

$p(\beta|y^*, y, X)$ depends upon the prior $p(\beta)$

If the prior is diffuse:

$$\beta|y^*, y, X \sim N(\hat{\beta}, [X'X]^{-1})$$

$$\hat{\beta} = [X'X]^{-1} X'y^*$$

Gibbs Sampling (cont'd)

If the prior is normal, i.e., $\beta \sim N(b, B^{-1})$

$$\beta|y^*, y, X \sim N(\tilde{\beta}, \tilde{B}^{-1})$$

then

$$\begin{aligned}\tilde{\beta} &= [X'X]^{-1} X'y^* \\ &= [B + X'X]^{-1} [Bb + X'y^*] \\ &= W_p b + W_d \hat{\beta}\end{aligned}$$

where

$$W_p = [B + X'X]^{-1} B$$

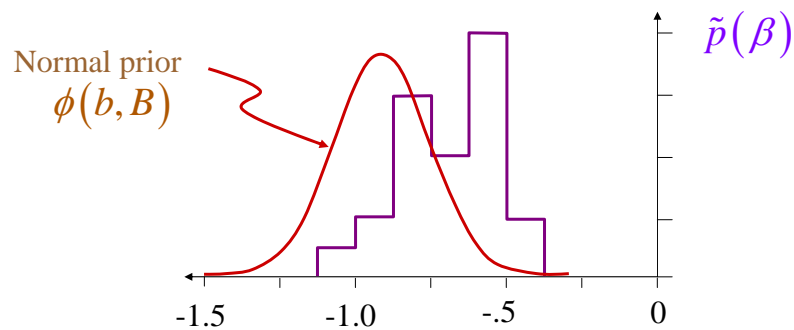
$$W_d = [B + X'X]^{-1} X'X$$

Alternative Priors

What if your information is not of the form

$$\beta \sim N(b, B^{-1})?$$

Example:



Posterior Reweighting

Given a normal prior:

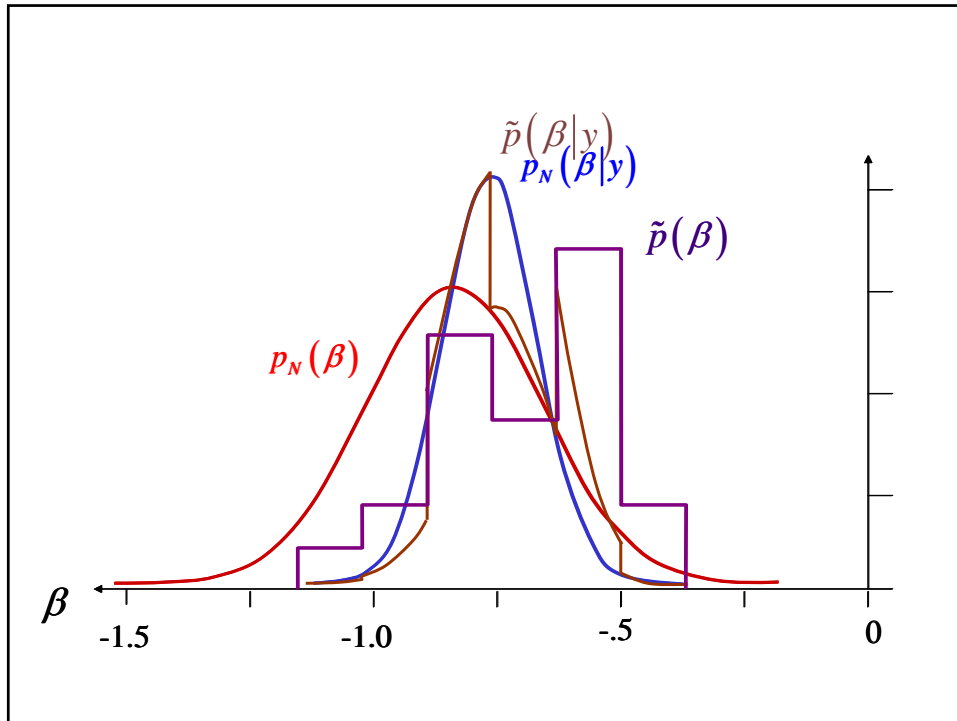
$$p_N(\theta|y, X) \propto p_N(\theta) p(y|X, \theta)$$

Given an alternative prior of $\tilde{p}(\theta)$, the posterior should be

$$\tilde{p}(\theta|y, X) \propto \tilde{p}(\theta) p(y|X, \theta)$$

Posterior reweighting simply takes advantage of the fact that

$$\begin{aligned} \tilde{p}(\theta|y, X) &\propto \tilde{p}(\theta) p(y|X, \theta) \\ &= \left[\frac{\tilde{p}(\theta)}{p_N(\theta)} \right] p_N(\theta) p(\theta|y, X) \\ &\propto w(\theta) p_N(\theta|y, X) \end{aligned}$$



Inference in Discrete Choice Models

- In the standard linear regression model we have

$$y_i = \beta'x_i + \varepsilon_i$$

with

$$E(y) = \beta'\bar{x}$$

and

$$\frac{\partial y_i}{\partial x_i} = \beta$$

- In discrete choice models, both prediction and interpretation of the parameters is complicated

Fitted Choice Probabilities

- The conditional choice probability of choosing alternative A is given by

$$\begin{aligned}P_1(x_i) &= \Pr[y_i = 1 | x_i, \beta] \\ &= \Pr[y_i^* > 0 | x_i, \beta] \\ &= \Pr[\eta_i < \beta'x_i | x_i, \beta] \\ &= F[\beta'x_i]\end{aligned}$$

- Issues
 - What are the statistical properties of fitted choice probabilities?
 - How does one aggregate over decision makers to make inference for the population?

Taylor Series Expansion

Suppose:

$$Asy.Var[\hat{\beta}] = V$$

A first order Taylor Series expansion of the fitted choice probability around β yields

$$\begin{aligned}\hat{P}_1(x) &= F[\hat{\beta}'x] \\ &\approx F[\beta'x] + (\beta - \hat{\beta})' \frac{\partial F[\beta'x]}{\partial \hat{\beta}'} \\ &= F[\beta'x] + (\beta - \hat{\beta})' f[\beta'x]x\end{aligned}$$

Taylor Series Expansion (cont'd)

Then

$$\begin{aligned} \text{Var}[\hat{P}_1(x)] &\approx \text{Var}\left\{F[\beta'x] + (\beta - \hat{\beta})' f[\beta'x]x\right\} \\ &= \text{Var}\left\{\hat{\beta}' f[\beta'x]x\right\} \\ &= f[\beta'x]^2 x' \text{Var}(\hat{\beta})x \end{aligned}$$

Thus

$$\text{Asy.Var}[\hat{P}_1(x)] \approx f(\beta'x)^2 x' Vx$$

which can be estimated using $f[\hat{\beta}'x]^2 x' \hat{V}x$

Simulation

- If one has a characterization of the asymptotic distribution for the estimated parameters, then simulation can also be used.
- For example, suppose, as in MLE

$$\hat{\beta} \stackrel{a}{\sim} N(\beta, V)$$

- Let β^r denote the r^{th} draw from the distribution
- A consistent estimator for the asymptotic variance for the fitted choice probabilities is given by

$$V_P^R = \frac{1}{R} \sum_{r=1}^R F(\beta^r' x)^2 - \left[\frac{1}{R} \sum_{r=1}^R F(\beta^r' x) \right]^2$$

Example: Chicago Transit Authority

- Let:

$$y_i = \begin{cases} 1 & \text{if individual } i \text{ commutes by car} \\ 0 & \text{if individual } i \text{ commutes by train} \end{cases}$$

- Explanatory variables
 - TW: walking time from nearest train stop to place of work (+)
 - AIVTSS: Difference between drive time and train ride time (-)
 - ACF: Difference between auto-parking charge and train fare (-)
 - AW: Number of vehicles in the household (+)

Estimated Parameters (MLE)

	Linear Probability Model	Logit	Probit
Constant	0.35** (0.05)	-5.74** (1.14)	-3.24** (0.61)
TW	0.0077* (0.0031)	0.18** (0.06)	0.11** (0.03)
AIVTSS	-0.0012 (0.0023)	-0.11* (0.04)	-0.06* (0.02)
ACF	-0.097** (0.004)	-2.45** (0.32)	-1.38** (0.17)
AW	0.12** (0.02)	4.37** (0.67)	2.45** (0.36)

Fitted Choice Probabilities

Let AIVTSS = 2 minutes, ACF = \$0.80, and AW = 3

	Linear Probability Model	Logit	Probit
TW=100	1.02 (0.20)	1.00 (<0.01)	1.00 (<0.01)
TW=25	0.51 (0.02)	0.44 (0.08)	0.45 (0.07)
TW=15	0.44 (0.04)	0.11 (0.07)	0.12 (0.08)

Aggregation

- Conditional probabilities are straightforward
- Extrapolating to the population can be more difficult

$$E[P_1(x)] \neq P_1[E(x)]$$

Train (2003) Example

- Suppose there are two types of individuals, a and b , equally represented in the population, with

$$V_a = \beta'x_a$$

$$V_b = \beta'x_b$$

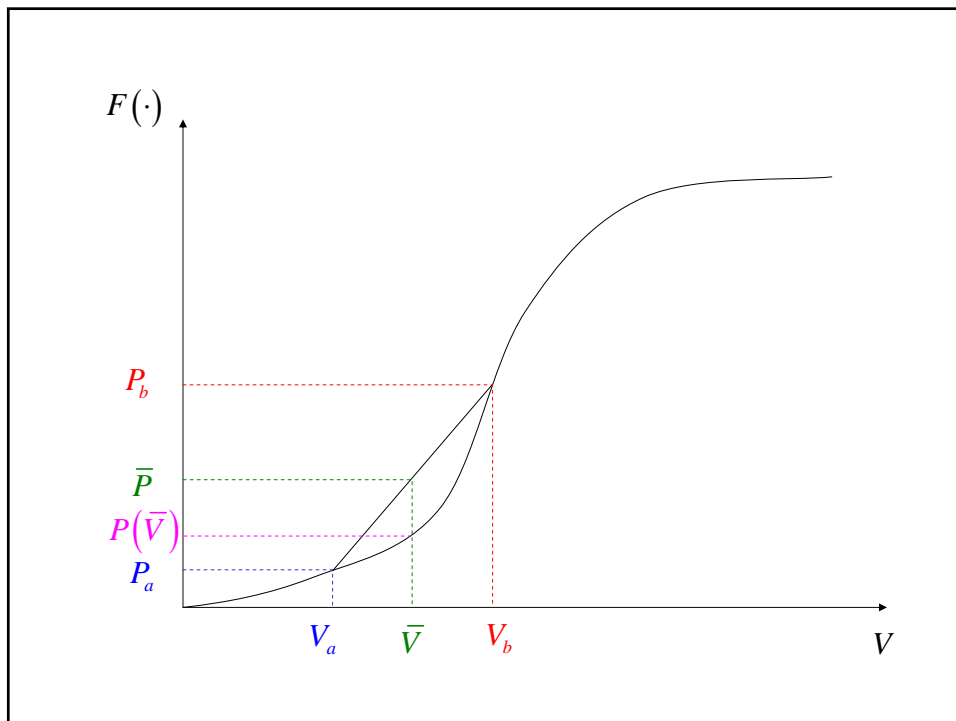
then

$$\begin{aligned} P_a &= \Pr[y_i = 1|x_a] \\ &= F[\beta'x_a] \end{aligned}$$

$$\begin{aligned} P_b &= \Pr[y_i = 1|x_b] \\ &= F[\beta'x_b] \end{aligned}$$

but

$$\bar{P} = \frac{1}{2}(P_a + P_b) \neq P(\bar{x}) = F[\beta'\bar{x}]$$



Errors in Aggregation

In general, $P(\bar{V})$ will tend to

- underestimate \bar{P} when probabilities are low
- overestimate \bar{P} when probabilities are high

Sample Enumeration

- In aggregating over individuals or projecting to the population as a whole, one needs to keep in mind
 - Degree to which sample is representative of target population
 - Endogeneities in sample selection
- Sample enumeration frequently used when sample is exogeneously determined
- Controlling for endogenous sampling is more difficult

Sample Enumeration (cont'd)

- Let $w(x)$ denote the probability of observing characteristics x in the sample
- Let $W(x)$ denote the probability of observing characteristics x in the population

$$P_1 = \frac{1}{N} \sum_i \left[\frac{W(x_i)}{w(x_i)} P_1(x_i) \right]$$

denotes an estimate of the population choice probability

- If $w(x) = W(x)$, then

$$P_1 = \frac{1}{N} \sum_i P_1(x_i)$$

Marginal Effects

It is important to keep in mind that parameters in discrete choice models rarely reflect marginal impact of the corresponding variable

In general:

$$\begin{aligned} \frac{\partial E[y_i | x_i]}{\partial x_i} &= \left\{ \frac{dF[\beta'x_i]}{d(\beta'x_i)} \right\} \beta \\ &= f(\beta'x_i) \beta \end{aligned}$$

so

$$\frac{\frac{\partial E[y_i | x_i]}{\partial x_{ik}}}{\frac{\partial E[y_i | x_i]}{\partial x_{ij}}} = \frac{\beta_k}{\beta_j}$$

Marginal Effects – Logit

For the logit model

$$\begin{aligned}\frac{\partial E[y_i|x_i]}{\partial x_i} &= \{\Lambda(\beta'x_i)[1-\Lambda(\beta'x_i)]\}\beta \\ &= h_L(x_i, \beta)\end{aligned}$$

The corresponding asymptotic variance is

$$\text{Asy.Var.}[h_L(x_i, \hat{\beta})] = [\Lambda(1-\Lambda)]^2 \Gamma'V\Gamma$$

where

$$\Gamma = I + (1-2\Lambda)x\beta'$$

Marginal Effects – Probit

For the logit model

$$\begin{aligned}\frac{\partial E[y_i|x_i]}{\partial x_i} &= \phi(\beta'x_i)\beta \\ &= h_p(x_i, \beta)\end{aligned}$$

The corresponding asymptotic variance is

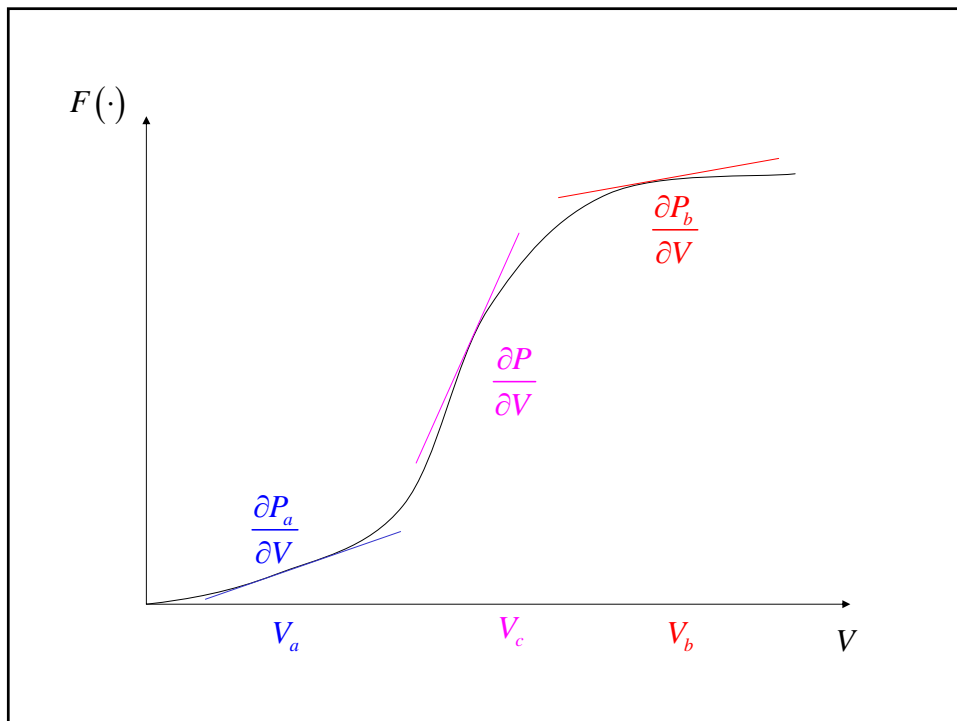
$$\text{Asy.Var.}[h_p(x_i, \hat{\beta})] = \phi^2\Gamma_p'V\Gamma_p$$

where

$$\Gamma_p = I - (\beta'x)x\beta'$$

Marginal Effects Continued

- Largest near center of distribution
- smallest in tail of distribution
- Sensible: says that exogenous factors have greatest impact for those close to choosing either alternative



Elasticities

In general

$$\frac{\partial E[y_i | x_i]}{\partial x_{ik}} \frac{x_{ik}}{E[y_i | x_i]} = \frac{f(\beta' x_i)}{F(\beta' x_i)} \beta_k x_{ik}$$

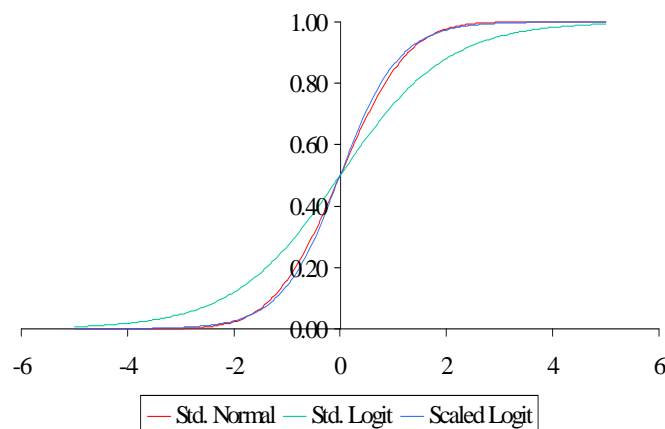
For logit

$$\begin{aligned} \frac{\partial E[y_i | x_i]}{\partial x_{ik}} \frac{x_{ik}}{E[y_i | x_i]} &= \frac{\{\Lambda(\beta' x_i)[1 - \Lambda(\beta' x_i)]\}}{\Lambda(\beta' x_i)} \beta_k x_{ik} \\ &= [1 - \Lambda(\beta' x_i)] \beta_k x_{ik} \end{aligned}$$

Probit does not reduce

Logit and Probit Yield Similar Results

Comparison of Logit and Probit



Example #1: Greene

- Comparison of LPM, logit, probit, and Weibull specifications
- dependent variable is whether or not a student's grade on an examination improved.
- explanatory variables include
 GPA: initial GPA,
 TUCE: pretest score,
 PSI: exposure to a new method of teaching economics

Variable	Coefficient Estimate				Marginal Impacts			
	Linear	Logit	Probit	Weibull	Linear	Logit	Probit	Weibull
Const.	-1.50	-13.02	-7.45	-10.63	--	--	--	--
GPA	.46	2.83	1.63	2.29	.46	.53	.53	.48
TUCE	.01	.10	.05	.04	.01	.02	.02	.01
PSI	.38	2.38	1.43	1.56	.38	.45	.47	.32

For Transportation Study: Marginal Effects

	Linear Probability Model	Logit	Probit
TW	0.0077*	0.0068	0.0070
AIVTSS	-0.0012	-0.0042	-0.0042
ACF	-0.097	-0.091	-.092
AW	0.12	0.16	0.16

Hypothesis Testing - General

Standard battery of tests can be applied

Wald test for $R\beta = q$: $(R\beta - q)' \{R\hat{V}R'\}^{-1} (R\beta - q) \sim \chi^2(r)$

Likelihood ratio tests: $LR = -2[l_{con} - l_{unc}] \sim \chi^2(r)$

Lagrange Multiplier test: $LM = g'\hat{V}g$

where g_i denotes the gradient of the unrestricted likelihood function evaluated at the restricted parameter estimates.

Specification Tests

Greene (2000) suggests considering two specification problems:

- Effects of omitted variables
- Heteroskedasticity

Often these are the same thing

$$\varepsilon_{ij} \equiv U(x_{ij}, z_{ij}, s_i, t_i) - V(x_{ij}, s_i)$$

Omitted Variables

- In linear regression model, omitted variables are not necessarily problematic if uncorrelated with included variables
- Problem is greater in binary choice models
 - Omitted variables lead to bias and inconsistency
 - MLE covariance matrix is inappropriate
- Omitted variable tests proceed in the obvious way, but require the identification of what is “omitted”.

Heteroskedasticity Test

One can specify:

$$\text{Var}(\varepsilon_i) = [\exp(\gamma'z)]^2$$

The log-likelihood function becomes

$$\begin{aligned} L(y, X, \beta, \gamma, z) \\ = \sum_i \left[(y_i) \ln \left(F \left[\frac{\beta'x_i}{\exp(\gamma'z)} \right] \right) + (1 - y_i) \ln \left(1 - F \left[\frac{\beta'x_i}{\exp(\gamma'z)} \right] \right) \right] \end{aligned}$$

Test $\gamma = 0$

Goodness of Fit Measure

McFadden's pseudo-R²

$$LRI = 1 - \frac{L}{L_0}$$

L denotes the unconstrained log likelihood function

L_0 denotes the log likelihood function when only a constant is included in the model

$$L_0 = n [P \ln P + (1 - P) \ln(1 - P)]$$

where P denotes the percentage of observations in the sample with $y_i=1$.

Merits of Pseudo R²

- Lies in the unit interval
- If the model provides no predictive power, then $L = L_0$ and $LRI=0$
- As the model's fit improves, then $L \rightarrow 0$ and $LRI \rightarrow 1$
- However, no interpretation to scale between 0 and 1

Cross Tabulation of Hits and Misses

Let

$$\hat{y}_i = \begin{cases} 1 & \hat{F}_i \geq 0.5 \\ 0 & \hat{F}_i < 0.5 \end{cases}$$

		Predicted	
		$\hat{F}_i \geq 0.5$	$\hat{F}_i < 0.5$
Actual	$y_i = 1$		
	$y_i = 0$		

Problems with Success Table

- The prediction rule is arbitrary.
 - No linkage made to the costs of the individual errors made. It may be more costly to make an error to classify a “yes” as a “no” than to misclassify a “no” as a “yes”.
 - Some loss function would be more helpful in this case.
- There is no way to judge departures from a diagonal table.

Recent Developments

- Fixed effects models
- Random effects models
- Mixed logit
- Reduced parameterization
 - Maximum Score Estimator
 - semi-nonparametric
 - Maximum Entropy
 - Nonparametric, semi-parametric,....

Panel Data – Linear Regression Context

- The standard linear regression model in a panel data setting has the form:

$$y_{it}^* = \beta' x_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

where

α_i capture unobserved individual effects

δ_t capture unobserved time effects

ε_{it} denotes the residual error

- Two principle sets of assumptions are used to deal with these effects
 - Assume fixed effects
 - Assume random effects

Panel Data – Fixed Effects (cont'd)

- For simplicity, consider the case in which there are no time effects; i.e.,:

$$y_{it}^* = \beta' x_{it} + \alpha_i + \varepsilon_{it}$$

- The fixed effects are typically nuisance parameters and are averaged out using

$$\bar{y}_{i\cdot}^* \equiv \frac{1}{T} \sum_{t=1}^T y_{it}^* = \beta' \bar{x}_{i\cdot} + \alpha_i + \bar{\varepsilon}_{i\cdot}$$

So that

$$\begin{aligned} \tilde{y}_{it}^* &\equiv y_{it}^* - \bar{y}_{i\cdot}^* \\ &= \beta' (x_{it} - \bar{x}_{i\cdot}) + (\varepsilon_{it} - \bar{\varepsilon}_{i\cdot}) \\ &= \beta' \tilde{x}_{it} + \tilde{\varepsilon}_{it} \end{aligned}$$

- OLS can be run on this transformed model

Panel Data – Fixed Effects Binary Choice Models

- In a binary choice setting, we only observe:

$$y_{it} = \begin{cases} 1 & y_{it}^* > 0 \\ 0 & y_{it}^* \leq 0 \end{cases}$$

- Averaging out the nuisance parameters is not generally possible
- Estimating N α_i 's is not feasible for large N

Chamberlain (1980) Conditional MLE for logit

- Chamberlain (1980) suggested that rather than focus on the unconditional likelihood,

$$l = \prod_{i=1}^N \prod_{t=1}^T F(\beta' x_{it})^{y_{it}} [1 - F(\beta' x_{it})]^{(1-y_{it})}$$

One should focus on the conditional likelihood

$$l^c = \prod_{i=1}^N \Pr \left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iT} = y_{iT} \mid \sum_{t=1}^T y_{it} \right]$$

i.e., we are conditioning on the number of 1's over time for each individual
analogous to conditioning on group means in continuous case

Example: T=2

$$l^c = \prod_{i=1}^N \Pr[Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \mid y_{i1} + y_{i2}]$$

Case 1: $y_{i1} + y_{i2} = 0 \Rightarrow \Pr[Y_{i1} = 0, Y_{i2} = 0 \mid y_{i1} + y_{i2} = 0] = 1$

Case 2: $y_{i1} + y_{i2} = 2 \Rightarrow \Pr[Y_{i1} = 1, Y_{i2} = 1 \mid y_{i1} + y_{i2} = 2] = 1$

Case 3: $y_{i1} + y_{i2} = 1 \Rightarrow$

Two possible outcome: (0,1) or (1,0)

$$\Pr[Y_{i1} = 0, Y_{i2} = 1 \mid y_{i1} + y_{i2} = 1]$$

$$= \frac{\Pr[Y_{i1} = 0, Y_{i2} = 1]}{\Pr[Y_{i1} = 0, Y_{i2} = 1] + \Pr[Y_{i1} = 1, Y_{i2} = 0]}$$

$$= \frac{\frac{1}{1 + e^{\beta'x_{i1} + \alpha_i}} \frac{e^{\beta'x_{i2} + \alpha_i}}{1 + e^{\beta'x_{i2} + \alpha_i}}}{\frac{1}{1 + e^{\beta'x_{i1} + \alpha_i}} \frac{e^{\beta'x_{i2} + \alpha_i}}{1 + e^{\beta'x_{i2} + \alpha_i}} + \frac{e^{\beta'x_{i1} + \alpha_i}}{1 + e^{\beta'x_{i1} + \alpha_i}} \frac{1}{1 + e^{\beta'x_{i2} + \alpha_i}}}$$

$$= \frac{e^{\beta'x_{i2} + \alpha_i}}{e^{\beta'x_{i1} + \alpha_i} + e^{\beta'x_{i2} + \alpha_i}}$$

$$= \frac{e^{\beta'x_{i2}}}{e^{\beta'x_{i1}} + e^{\beta'x_{i2}}}$$

Similarly:

$$\begin{aligned}
 & \Pr[Y_{i1} = 1, Y_{i2} = 0 \mid y_{i1} + y_{i2} = 1] \\
 &= \frac{\Pr[Y_{i1} = 1, Y_{i2} = 0]}{\Pr[Y_{i1} = 0, Y_{i2} = 1] + \Pr[Y_{i1} = 1, Y_{i2} = 0]} \\
 &= \frac{\frac{e^{\beta'x_{i1} + \alpha_i}}{1 + e^{\beta'x_{i1} + \alpha_i}} \frac{1}{1 + e^{\beta'x_{i2} + \alpha_i}}}{\frac{1}{1 + e^{\beta'x_{i1} + \alpha_i}} \frac{e^{\beta'x_{i2} + \alpha_i}}{1 + e^{\beta'x_{i2} + \alpha_i}} + \frac{e^{\beta'x_{i1} + \alpha_i}}{1 + e^{\beta'x_{i1} + \alpha_i}} \frac{1}{1 + e^{\beta'x_{i2} + \alpha_i}}} \\
 &= \frac{e^{\beta'x_{i1} + \alpha_i}}{e^{\beta'x_{i1} + \alpha_i} + e^{\beta'x_{i2} + \alpha_i}} \\
 &= \frac{e^{\beta'x_{i1}}}{e^{\beta'x_{i1}} + e^{\beta'x_{i2}}}
 \end{aligned}$$

Each element of the conditional log-likelihood is independent of α_i 's

Panel Data – Random Effects Regression Context

- The curse of the fixed effects specification is that the number of parameters grows linearly with N
- The random effects specification avoids this curse by assuming α_i 's are drawn from some underlying distribution; e.g.,

$$\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2)$$

- One then need only estimate $\bar{\alpha}$ and σ_α
Increases in N makes this easier to do

Panel Data – Random Effects Error Components

- The random effects specification has an error components interpretation:

$$\begin{aligned}y_{it}^* &= \beta' x_{it} + \alpha_i + \varepsilon_{it} \\ &= \beta' x_{it} + \eta_{it}\end{aligned}$$

with

$$\eta_{it} \equiv \alpha_i + \varepsilon_{it}$$

- This is appealing in the binary setting, where we assume a portion of the error represents unobserved individual characteristics

Random Effects Discrete Choice Sources

- Butler, J., and R. Moffitt (1982), “A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit,” *Econometrica* **50**: 761-764.
- Greene, W. H., *Econometric Analysis*, 4th edition, Upper Saddle River, New Jersey: Prentice-Hall, Inc., Sections 19.5.1.
- Tauchen, H., A. D. Witte, and H. Griensinger (1994), “Criminal Deterrence: Revisiting the Issue with a Birth Cohort,” *The Review of Economics and Statistics*, **76**: 399-412.

Panel Data – Random Effects Binary Probit

- Suppose:

$$\begin{aligned} y_{it}^* &= \beta' x_{it} + \alpha_i - \varepsilon_{it} \\ &= \beta' x_{it} - \eta_{it} \end{aligned}$$

where

$$\varepsilon_{it} \sim N(0,1)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

Notes: $\bar{\alpha}$ is subsumed into the intercept term of $\beta' x_{it}$
 α_i and ε_{it} are assumed to be independent

Contribution of Individual i to Likelihood Function

$$\begin{aligned} \Pr[y_{i1}, \dots, y_{iT}] &= \Pr[\eta_{i1} < q_{i1} \beta' x_{i1}, \dots, \eta_{iT} < q_{iT} \beta' x_{iT}] && ; q_{it} \equiv 2y_{it} - 1 \\ &= \Pr[\varepsilon_{i1} < q_{i1} \beta' x_{i1} + \alpha_i, \dots, \varepsilon_{iT} < q_{iT} \beta' x_{iT} + \alpha_i] \\ &= \int_{-\infty}^{\infty} \Pr[\varepsilon_{i1} < q_{i1} \beta' x_{i1} + \alpha_i, \dots, \varepsilon_{iT} < q_{iT} \beta' x_{iT} + \alpha_i | \alpha_i] f(\alpha_i) d\alpha_i \\ &= \int_{-\infty}^{\infty} \left\{ \Pr[\varepsilon_{i1} < q_{i1} \beta' x_{i1} + \alpha_i | \alpha_i] \cdots \Pr[\varepsilon_{iT} < q_{iT} \beta' x_{iT} + \alpha_i | \alpha_i] \right\} f(\alpha_i) d\alpha_i \\ &= \int_{-\infty}^{\infty} \left\{ \prod_{t=1}^T \Phi[q_{it} \beta' x_{it} + \alpha_i] \right\} f(\alpha_i) d\alpha_i \end{aligned}$$

Random Effects Probit (cont'd)

- The trick here is to do the integration in steps, taking advantage of the independence of ε_{it} and α_i
- Similar steps can be used to introduce time effects

$$y_{it}^* = \beta' x_{it} + \alpha_i + \delta_t - \varepsilon_{it} = \beta' x_{it} - \eta_{it}$$

$$\delta_t \sim N(0, \sigma_\delta^2)$$

$$l = \int_{-\infty}^{\infty} \prod_{i=1}^N \left[\int_{-\infty}^{\infty} \left\{ \prod_{t=1}^T \Phi[q_{it} \beta' x_{it} + \alpha_i + \delta_t] \right\} f(\alpha_i) d\alpha_i \right] f(\delta_1) \cdots f(\delta_T) d\delta_1 \cdots d\delta_T$$

Example: Tauchen, White, and Griesinger (1994)

Focus on efficacy of criminal deterrence

$$c_{it}^* = \beta' x_{it} + \alpha_i + \delta_t - \varepsilon_{it}$$

where c_{it}^* denotes latent variable determining individual i 's propensity to be arrested in year t

Observe:

$$c_{it} = \begin{cases} 1 & c_{it}^* > 0 \\ 0 & c_{it}^* \leq 0 \end{cases}$$

Simplest Model Results

Variable	Coefficient	P-value
Real Police Budget	-0.019	(0.01)
IQ	-0.016	(0.01)
Age	-0.003	(0.93)
US born parents	0.314	(0.34)
Occupational status of Head of Household	-0.005	(0.22)
Number of Addresses during school	0.110	(0.03)
Attended parochial school	-0.390	(0.08)
Average income of neighborhood	0.077	(0.52)
Caucasian	-0.560	(0.02)
Variance on individual effects	0.97	(<0.01)
Variance of time effects	<0.01	(>0.99)

Random Effects: Alternative Assumptions

- The random effects model above uses normal assumptions throughout
- Other distributional assumptions are possible
- Suppose that ε_{it} is drawn from a logistic distribution and

$$\begin{aligned}
 y_{it}^* &= \beta' x_{it} + \alpha_i - \varepsilon_{it} & \alpha_i &\sim N(0, \sigma_\alpha^2) \\
 &= \beta' x_{it} - \eta_{it}
 \end{aligned}$$

then

$$= \prod_{i=1}^N \int_{-\infty}^{\infty} \left\{ \prod_{t=1}^T \Lambda[q_{it} \beta' x_{it} + \alpha_i] \right\} f(\alpha_i) d\alpha_i$$

Mixed Logit

- Mixed logit is essentially a generalization of this idea

- Let

$$y_{it}^* = \beta' x_{it} - \varepsilon_{it}$$

where ε_{it} is drawn from a logistic distribution and

$$\beta \sim N(\bar{\beta}, \Sigma_{\beta})$$

- Conditional on the parameter vector β , the likelihood function is given by

$$l(y|X, \beta) = \prod_{i=1}^N \prod_{t=1}^T \Lambda(\beta' x_{it})$$

Mixed Logit (cont'd)

- The unconditional likelihood function is given by

$$\begin{aligned} l(y|X) &= \int l(y|X, \beta) f(\beta) d\beta \\ &= \int \left\{ \prod_{i=1}^N \prod_{t=1}^T \Lambda(\beta' x_{it}) \right\} f(\beta) d\beta \end{aligned}$$

- Simulation methods are typically used to do the above integration
- Error components, for example, emerge if x_{it} includes individual specific dummy variables.

Mixed Logit Sources

- Brownstone, D., and K. E. Train, "Forecasting New Product Penetration with Flexible Substitution Patterns," *Journal of Econometrics*, **89**(1999): 109-129.
- McFadden, D., and K. E. Train, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, **15**(2000): 447-470.
- Revelt, D., and K. E. Train, "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level," *Review of Economics and Statistics*, **80**(1998): 647-657.
- Train, K. E., "Recreation Demand Models with Taste Differences Over People," *Land Economics* **74**(1998): 230-39.
- Train, K. E., "Mixed Logit Models for Recreation Demand." in *Valuing Recreation and the Environment: Revealed Preference Methods in Theory and Practice* (J. A. Herriges and C. L. Kling, eds.) Edward Elgar Publishing Ltd., Aldershot (1999a).

Relaxing/Avoiding Parametric Assumptions

- The problem with all of the models/methods thus far is that they rely on correctly specifying the underlying distributions
- A variety of alternative methods have been suggested to relax or avoid these assumptions:
 - Maximum Score Estimator (MSCORE)
 - Semi-Nonparametric Estimation
 - Non-Parametric Estimation
 - Maximum Entropy

MSCORE

- The maximum score estimator solves the fitting rule

$$\underset{\beta \ni \beta' \beta = 1}{\text{Max}} S_{\alpha}(\beta) = \frac{1}{N} \sum_{i=1}^N [q_i - (1 - 2\alpha)] \text{sgn}(\beta' x_i)$$

$$q_i = 2y_i - 1$$

- Consider the case in which $\alpha = 0.5$

$$S_{0.5}(\beta) = \frac{1}{N} \sum_{i=1}^N q_i \text{sgn}(\beta' x_i) = \frac{1}{N} \sum_{i=1}^N q_i \hat{q}_i(\beta)$$

$$q_i \hat{q}_i = \begin{cases} 1 & (y_i, \hat{y}_i) = (1, 1) \text{ or } (0, 0) \\ -1 & (y_i, \hat{y}_i) = (0, 1) \text{ or } (1, 0) \end{cases}$$

Maximizes number of correct predictions

MSCORE

- Key limitation: no standard errors
- Bootstrapping is used to assess sample variability of estimate

$$MSD(b) = \frac{1}{B} \sum_{b=1}^B [b_M(b) - b_n][b_M(b) - b_n]'$$

where

$$b_n = \underset{\beta \ni \beta' \beta = 1}{\text{arg Max}} S_{\alpha}(\beta)$$

and $b_M(b)$ is the MSCORE estimator from a bootstrapped sample

Semi-nonparametric Estimation

Suppose

$$y_i^* = f(x_i, \beta) - \eta_i$$

where $\eta_i \sim F(\eta)$ is from an unknown distribution.

The heart of the semi-nonparametric procedure is that there is an unknown transformation function, $h(\cdot)$, such that

$$\Gamma[h(\eta)] = F(\eta)$$

where $\Gamma(\cdot)$ is a known distribution (e.g., logistic)

Both $f(\cdot)$ and $h(\cdot)$ are then approximated by flexible forms

Semi-nonparametric Estimation (cont'd)

Example

$$h_r(u) = \gamma_0 + \int_0^u (\gamma_1 + \gamma_2\eta + \cdots + \gamma_r\eta^{r-1})^2 d\eta.$$

The polynomial:

$$\gamma_1 + \gamma_2\eta + \cdots + \gamma_r\eta^{r-1}$$

Provides an approximation to

$$\sqrt{h'(\eta)}$$

and insures monotonicity of the transformation

Semi-nonparametric Estimation (cont'd)

A Fourier Flexible Form (FFF) can be used to model $f(\cdot)$

$$f(x_i; \delta_{JA}) \\ = \mu_0 + b'x_i + \frac{1}{2}x_i'Cx_i + \sum_{\alpha=1}^A \left(\mu_{0\alpha} + 2 \sum_{j=1}^J \{ \mu_{j\alpha} \cos(jk'_{\alpha}x_i) - v_{j\alpha} \sin(jk'_{\alpha}x_i) \} \right)$$

Sources

- MSCORE
 - Greene, W. H., *Econometric Analysis*, 4th edition, Upper Saddle River, New Jersey: Prentice-Hall, Inc., Sections 19.5.3
 - Manski, C. (1975), “The Maximum Score Estimator of the Stochastic Utility Model of Choice,” *Journal of Econometrics* **3**: 205-228
- Semi-nonparametric
 - Chen, H., and A. Randall. (1997) “Semi-Nonparametric Estimation of Binary Response Models with An Application To Natural Resource Valuation.” *Journal of Econometrics* **76**: 323-340.
 - Gerfin, M., (1996) “Parametric and Semi-parametric Estimation of the Binary Response model of Labour Market Participation,” *Journal of Applied Econometrics*