Sampling Theory for Discrete Data

□ Economic survey data are often obtained from sampling protocols that involve stratification, censoring, or selection. Econometric estimators designed for random samples may be inconsistent or inefficient when applied to these samples.
□ When the econometrician can influence sample design, then the use of stratified sampling protocols combined with appropriate estimators can be a powerful tool for maximizing the useful information on structural parameters obtainable within a data collection budget.
□ Sampling of discrete choice alternatives may simplify data collection and analysis for MNL models.

Basics of Sampling Theory

Let z denote a vector of exogenous variables, and y denote an endogenous variable, or a vector of endogenous variables, such as choice indicators. The joint distribution of (z,y) in the population is $P(y|z,\beta_o)p(z) = Q(z|y,\beta_o)q(y,\beta_o)$; see Figure 1. $P(y|z,\beta_o)$, the conditional probability of y, given z, in a parametric family with true parameter vector β_o, is the structural model of interest. "Structural" means this conditional probability law is invariant in different populations or policy environments when the marginal distribution of z changes. If z causes y, then $P(y\mid z,\beta_o)$ is a structural relationship. Other notation:

p(z) marginal distribution of exogenous variables
Q(z|y) conditional distribution of z given y
q(y) marginal distribution of y

Figure 1. Population Probability Model

	y_1	y_2	...	y_J	Total
z_1	$P(y_1\|z_1,\beta_0)p(z_1)$	$P(y_2\|z_1,\beta_0)p(z_1)$		$P(y_J\|z_1,\beta_0)p(z_1)$	$p(z_1)$
z_2	$P(y_1\|z_2,\beta_0)p(z_2)$	$P(y_2\|z_2,\beta_0)p(z_2)$		$P(y_J\|z_2,\beta_0)p(z_2)$	$p(z_2)$
:					
z_K	$P(y_1\|z_K,\beta_0)p(z_K)$	$P(y_2\|z_K,\beta_0)p(z_K)$		$P(y_J\|z_K,\beta_0)p(z_K)$	$p(z_K)$
Total	$q(y_1,\beta_0)$	$q(y_2,\beta_0)$		$q(y_J,\beta_0)$	1

A *simple random sample* draws independent observations from the population, each with probability law $P(y|z,\beta_o)\cdot p(z)$. The kernel of the log likelihood of this sample depends only on the conditional probability $P(y|z,\beta)$, not on the marginal density p(z); thus, maximum likelihood estimation of the structural parameters β_o does not require that the marginal distribution p(z) be parameterized or estimated. β_o influences only how data are distributed within rows of the table above; and how the data are distributed across rows provides no additional information on β_o.

***Stratified Samples*. An (*exogenously*) *stratified random sample* samples among rows with probability weights different from p(z), but within rows the sample is distributed with the probability law for the population. Just as for a simple random sample, the sampling probabilities across rows do not enter the kernel of the likelihood function for β_o, so the exogenously stratified random sample can be analyzed in exactly the same way as a simple random sample.**

The idea of a stratified random sample can be extended to consider what are called *endogneous* or *choice-based* sampling protocols. Suppose the data are collected from one or more *strata*, indexed s = 1,..., S. Each stratum has a sampling protocol that determines the segment of the population that qualifies for interviewing.

Let R(z,y,s) = *qualification probability* that a population member with characteristics (z,y) will qualify for the subpopulation from which the stratum s subsample will be drawn. For example, a stratum might correspond to the northwest 2x2 subtable in Figure 1, where y is one of the values y1 or y2 and z is one of the values z1 or z2. In this case, R(z,y,s) equals the sum of the four cell probabilities. Qualification may be related to the *sampling frame,* which selects locations (e.g., census tracts, telephone prefixes), to *screening* (e.g., terminate interview if respondent is not a home-owner), or to *attrition* (e.g., refusals)

Examples:

1. *Simple random subsample,* with R(z,y,s) = 1.

2. *Exogenously stratified random subsample,* with R(z,y,s) = 1 if $z \in A_s$ for a subset A_s of the universe Z of exogenous vectors, R(z,y,s) = 0 otherwise. For example, the set A_s might define a geographical area. This corresponds to sampling randomly from one or more rows of the table.

Exogenous stratified sampling can be generalized to variable sampling rates by permitting R(z,y,s) to be any function from (z,s) into the unit interval; a protocol for such sampling might be, for example, a screening interview that qualifies a proportion of the respondents that is a function of respondent age.

3. Endogenously stratified subsample: $R(z,y,s) = 1$ if $y \in B_s$, with B_s a subset of the universe of endogenous vectors Y, and $R(z,y,s) = 0$ otherwise. The set B_s might identify a single alternative or set of alternatives among discrete responses, with the sampling frame intercepting subjects based on their presence in B_s; e.g., buyers who register their purchase, recreators at national parks. Alternately, B_s might identify a range of a continuous response, such as an income category. Endogenous sampling corresponds to sampling randomly from one or more columns of the table. A choice-based sample for discrete response is the case where each response is a different stratum. Then $R(z,y,s) = 1(y = s)$.

Endogenous stratified sampling can be generalized to qualification involving both exogenous and endogenous variables, with B_s defined in general as a subset of Z×Y. For example, in a study of mode choice, a stratum might qualify bus riders (endogenous) over age 18 (exogenous). It can also be generalized to differential sampling rates, with a proportion R(z,y,s) between zero and one qualifying in a screening interview.

4. *Sample selection/attrition*, with R(z,y,s) giving the proportion of the population with variables (z,y) whose availability qualifies them for stratum s. For example, R(z,y,s) may give the proportion of subjects with variables (z,y) that can be contacted and will agree to be interviewed, or the proportion of subjects meeting an endogenous selection condition, say employment, that qualifies them for observation of wage (in z) and hours worked (in y).

The Sample Probability Law

The *population* probability law for (z,y) is $P(y|z,\beta_o)\cdot p(z)$. The qualification probability R(z,y,s) characterizes the sampling protocol for stratum s. Then, $R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)$ = joint probability that a member of the population will have variables (z,y) and will qualify for stratum s.

(2) $$r(s) = R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)$$

is the proportion of the population qualifying into the stratum, or *qualification factor*. The reciprocal of r(s) is called the *raising factor.*

(3) $G(z,y|s) = R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)/r(s)$,

the conditional distribution $G(z,y|s)$ of (z,y) given qualification, is the *sample probability law* for stratum s. The probability law $G(z,y|s)$ depends on the unknown parameter vector β, on p(z), and on the qualification probability R(z,y,s). In simple cases of stratification, R(z,y,s) is fully specified by the sampling protocol. The qualification factor r(s) may be known (e.g., stratification based on census tracts with known sizes); estimated from the survey (e.g.; qualification is determined by a screening interview); or estimated from an auxiliary sample. In case of attrition or selection, R(z,y,s) may be an unknown function, or may contain unknown parameters.

The precision of the maximum likelihood estimator will depend on the qualification factor. Suppose a random sample of size n_s is drawn from stratum s, and let N = $\sum_s n_s$ denote total sample size. Let n(z,y│s) denote the number of observations in the stratum s subsample that fall in cell (z,y). Then, the log likelihood for the stratified sample is

$$\textbf{(4)} \qquad \mathbf{L = \sum_{s=1}^{S} \sum_{z} \sum_{y}} \; \boldsymbol{n}\mathbf{(z,y|s)\cdot Log\ G(z,y|s).}$$

EXOGENOUS STRATIFIED SAMPLING

If R(z,y,s) is independent of y, the qualification factor r(s) = R(z,s)p(z) is independent of β_o, and the log likelihood function separates into the sum of a kernel

$$(5)\qquad L_1 = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot\text{Log } P(y|z,\beta)$$

and terms independent of β. Hence, the kernel is independent of the structure of exogenous stratification.

ENDOGENOUS STRATIFICATION

Suppose the qualification probability R(z,y,s) depends on y. Then the qualification factor (2) depends on β_o, and the log likelihood function (4) has a kernel depending in general not only on β, but also on the unknown marginal distribution p(z). Any unknowns in the qualification probability also enter the kernel.

There are four possible strategies for estimation under these conditions:

1. Brute force -- Assume p(z) and, if necessary, R(z,y,s), are in parametric families, and estimate their parameters jointly with β. For example, in multivariate discrete data analysis, an analysis of variance representation absorbs the effects of stratification.

2. <u>Weighted Exogenous Sample Maximum Likelihood</u> (WESML) This is a quasi-maximum likelihood approach which starts from the likelihood function appropriate to a random sample, and reweights the data (if possible) to achieve consistency. A familiar form of this approach is the classical survey research technique of reweighting a sample so that it appears to be a simple random sample.

3. <u>Conditional Maximum Likelihood</u> (CML): This approach pools the observations across strata, and then forms the conditional likelihood of y given z in this pool. This has the effect of conditioning out the unknown density p(z).

4. <u>Full Information Maximum Likelihood</u> (FICLE): This approach formally maximizes the likelihood function in p(z) as a function of the data, the remaining parameters, and a finite vector of auxiliary parameters, and then concentrates the likelihood function by substituting in these formal maximum likelihood estimates for p(z).

□ Both the WESML and CML estimators are practical for many problems when auxiliary information is available that allows the raising factors to be estimated consistently. The FICLE estimator is computationally difficult and little used.

WESML

□ Recall the kernel of the log likelihood for exogenous sampling is given by (5). Suppose now endogenous sampling with true log likelihood (4), and consider a quasi-maximum likelihood criterion based on (5),

$$\textbf{(7)}\quad \mathbf{W(\beta)} = \sum_{s=1}^{S} \sum_{x} \sum_{y} \boldsymbol{n}\mathbf{(z,y \mid s)\cdot w(z,y,s)\cdot Log\,P(y \mid z,\beta)},$$

where w(z,y,s) = weight to achieve consistency.

□ Suppose r(s) is consistently estimated by f(s), from government statistics, survey frame data such as the average refusal rate, or an auxiliary sample. Consider the weights

$$(11) \qquad \mathbf{w(z,y) = 1/} \sum_{s=1}^{S} \left\lfloor R(z,y,s)n_s/Nf(s) \right\rfloor ;$$

these are well-defined if the bracketed expressions are positive and R(z,y,s) contains no unknown parameters. A classical application of WESML estimation is to a sample in which the strata coincide with the possible configurations of y, so that R(z,y,s) = 1(y = s). In this case, $w(z,y) = N\cdot f(y)/n_y$, the ratio of the population to the sample frequency. This is the *raising factor* encountered in classical survey research. Another application is to *enriched* samples, where a random subsample (s = 1) is enriched with an endogenous subsamples from one or more configurations of y; e.g., s = y = 2. Then, $w(z,1) = N/n_1$ and $w(z,2) = N\cdot f(2)/[n_1\cdot f(2) + n_2]$.

CML

Pool the observations from the different strata. Then, the data generation process for the pooled sample is

(15) $$\mathbf{Pr(z,y)} = \sum_{s=1}^{S} \mathbf{G(z,y,s)n_s/N,}$$

and the conditional probability of y given z from this pool is

(17) $$\mathbf{Pr(y|z)} = \frac{\sum_{s=1}^{S} R(z,y,s)\cdot P(y|z,\beta_o)\cdot n_s/N\cdot r(s)}{\sum_{y}\sum_{s=1}^{S} R(z,y,s)\cdot P(y|z,\beta_o)\cdot n_s/N\cdot r(s)} \quad .$$

The CML estimator maximizes the conditional likelihood of the pooled sample in β and any unknowns in R(z,y,s). When r(s) is known, or one wishes to condition on estimates f(s) of r(s) from auxiliary samples, (17) is used directly. More generally, given auxiliary sample information on the r(s), these can be treated as parameters and estimated from the joint likelihood of (17) and the likelihood of the auxiliary sample.

For discrete response in which qualification does not depend on z, the formula (17) simplifies to

$$\mathbf{Pr(y|z)} = \frac{P(y|z,\beta_o)\cdot\alpha_y}{\sum_y P(y|z,\beta_o)\cdot\alpha_y} \;,$$

where α_y = R(z,y,s)·n_s/N·r(s) can be treated as an alternative-specific constant. For multinomial logit choice models, Pr(y|z) then reduces to a *multinomial logit* formula with added alternative-specific constants. It is possible to estimate this model by the CML method using standard random sample computer programs for this model, obtaining consistent estimates for slope parameters, and for the sum of log α_y and alternative-specific parameters in the original model. What is critical for this to work is that the MNL model contain alternative-specific dummy variables corresponding to each choice-based stratum.

□ For an enriched sample, Pr(1|z) = P(1|z,β_o)·n_1/N·D and Pr(2|z) = P(2|z,β_o)·[n_1/N + n_2/N·r(2)]/D, where D = n_1/N + P(2|z,β_o)·n_2/N.

□ Example: Suppose y is a continuous variable, and the sample consists of a single stratum in which high income families are over-sampled by screening, so that

the qualification probability is $R(z,y,1) = \gamma < 1$ for $y \le y_o$ and $R(z,y,1) = 1$ for $y > y_o$. Then $Pr(y|z) = \gamma \cdot P(y|z,\beta_o)/D$ for $y \le y_o$ and $Pr(y|z) = P(y|z,\beta_o)/D$ for $y > y_o$, where $D = \gamma + (1-\gamma) \cdot P(y>y_o|z,\beta_o)$.

□ Both the WESML and CML estimators are computationally practical in a variety of endogenous sampling situations, and have been widely used.

Sampling of Alternatives for Estimation

Consider a simple or exogenously stratified random sample and discrete choice data to which one wishes to fit a multinomial logit model. Suppose the choice set is very large, so that the task of collecting attribute data for each alternative is burdensome, and estimation of the MNL model is difficult. Then, it is possible to reduce the collection and processing burden greatly by working with samples from the full set of alternatives. There is a cost, which is a loss of statistical efficiency. Suppose C_n is the choice set for subject n, and the MNL probability model is written

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j\in C_n} e^{V_{jn}}}$$

where i_n is the observed choice and $V_{in} = x_{in}.\beta$ is the systematic utility of alternative i.

Suppose the analyst selects a subset of alternatives A_n for this subject which will be used as a basis for estimation; i.e., if i_n is contained in A_n, then the subject is treated *as if* the choice were actually being made from A_n, and if i_n is not contained in A_n, the observation is discarded. In selecting A_n, the analyst may use the

information on which alternative i_n was chosen, and may also use information on the variables that enter the determination of the V_{in}. The rule used by the analyst for selecting A_n is summarized by a *probability* $\pi(A|i_n,V's)$ that subset A of C_n is selected, given the observed choice i_n and the observed V's (or, more precisely, the observed variables behind the V's).

The selection rule is a *uniform conditioning* rule if for the selected A_n, it is the case that $\pi(A_n | j,V's)$ is the same for all j in A_n. Examples of uniform conditioning rules are (1) select (randomly or purposively) a subset A_n of C_n *without* taking into account what the observed choice or the V's are, and keep the observation if and only if i_n is contained in the selected A_n; and (2) given observed choice i_n. select m-1 of the remaining alternatives at random from C_n, without taking into account the V's.

An implication of uniform conditioning is that in the sample containing the pairs (i_n,A_n) for which i_n is contained in A_n, the probability of observed response i_n conditioned on A_n is

$$P_{in|A_n} = \frac{e^{V_{in}}}{\sum_{j\in A_n} e^{V_{jn}}}$$

This is just a MNL model that treats choice *as if* it were being made from A_n rather than C_n, so that maximum likelihood estimation of this model for the sample of (i_n,A_n) with $i_n \in A_n$ estimates the same parameters as does maximum likelihood estimation on data from the full choice set. Then, this sampling of alternatives cuts down data collection time (for alternatives not in A_n) and computation size and time, but still gives consistent estimates of parameters for the original problem.

BIVARIATE SELECTION MODEL

(29) $$y^* = x\beta + \varepsilon\ ,$$
$$w^* = z\alpha + \sigma\nu\ ,$$

x, z	**vectors of exogenous variables, not necessarily all distinct,**
α, β	**parameter vectors, not necessarily all distinct,**
σ	**a positive parameter.**
y^*	**latent net desirability of work,**
w^*	**latent log potential wage.**

NORMAL MODEL

(30) $$\begin{bmatrix} \varepsilon \\ \nu \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)\ ,$$

correlation ρ.

Observation rule

"Observe $y = 1$ and $w = w^*$ if $y^* > 0$; observe $y = -1$ and do <u>not</u> observe w when $y^* \leq 0$".

The event of working ($y = 1$) or not working ($y = 0$) is observed, but net desirability is not, and the wage is observed only if the individual works ($y^* > 0$).

For some purposes, code the discrete response as $s = (y+1)/2$; then $s = 1$ for workers, $s = 0$ for non-workers.

The event of working is given by a *probit* model.

Φ is the standard univariate cumulative normal.

The probability of working is

$$P(y=1|x) = P(\varepsilon > -x\beta) = \Phi(x\beta),$$

The probability of not working is

$$P(y=-1|x) = P(\varepsilon \leq -x\beta) = \Phi(-x\beta).$$

Compactly,

$$P(y|x) = \Phi(yx\beta).$$

In the bivariate normal, the conditional density of one component given the other is univariate normal,

$$\varepsilon \,|\, \nu \sim N(\rho\nu, 1\text{-}\rho^2) = \frac{1}{\sqrt{1-\rho^2}} \cdot \phi\left(\frac{\varepsilon - \rho\nu}{\sqrt{1-\rho^2}} \right)$$

and

$$\nu \,|\, \varepsilon \sim N(\rho\varepsilon, 1\text{-}\rho^2) = \frac{1}{\sqrt{1-\rho^2}} \cdot \phi\left(\frac{\nu - \rho\varepsilon}{\sqrt{1-\rho^2}} \right) \;.$$

The joint density: marginal times conditional,

$$(\varepsilon,\nu) \sim \phi(\nu)\cdot \frac{1}{\sqrt{1-\rho^2}}\cdot\phi\left(\frac{\varepsilon-\rho\nu}{\sqrt{1-\rho^2}}\right)$$

$$= \phi(\varepsilon)\cdot \frac{1}{\sqrt{1-\rho^2}}\cdot\phi\left(\frac{\nu-\rho\varepsilon}{\sqrt{1-\rho^2}}\right) \ .$$

The density of (y^*,w^*)
(31) $f(y^*,w^*)$

$$= \frac{1}{\sigma}\,\phi(\frac{w^*-z\alpha}{\sigma})\cdot \frac{1}{\sqrt{1-\rho^2}}\cdot\phi\left(\frac{y^*-x\beta-\rho(w^*-z\alpha)/\sigma}{\sqrt{1-\rho^2}}\right)$$

$$= \phi(y^*\text{-}x\beta)\cdot \frac{1}{\sigma\sqrt{1-\rho^2}}\cdot\phi\left(\frac{w^*-z\alpha-\rho\sigma(y^*-x\beta)}{\sigma\sqrt{1-\rho^2}}\right) \ .$$

Log likelihood of an observation, $l(\alpha,\beta,\sigma,\rho)$.

In the case of a non-worker (y = -1 and w = NA), the density (31) is integrated over $y^* < 0$ and all w^*. Using the second form in (31), this gives probability $\Phi(-x\beta)$.

In the case of a worker, the density (31) is integrated over y* ≥ 0. Using the first form in (31)

$$\mathbf{e^{l(\alpha,\beta,\sigma,\rho)}} = \begin{cases} \Phi(-x\beta) & if\, y = -1 \\ \dfrac{1}{\sigma}\phi(\dfrac{w-z\alpha}{\sigma})\cdot\Phi\left(\dfrac{x\beta+\rho\left(\dfrac{w-z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) & if\, y = 1 \end{cases} .$$

The log likelihood can be rewritten as the sum of the marginal log likelihood of the discrete variable y and the conditional log likelihood of w given that it is observed, $l(\alpha,\beta,\sigma,\rho) = l^1(\alpha,\beta) + l^2(\alpha,\beta,\sigma,\rho)$, with the marginal component,

(33) $$l^1(\beta) = \log \Phi(yx\beta),$$

and the conditional component (that appears only when y = 1),

(34) $$l^2(\alpha,\beta,\sigma,\rho) = -\log \sigma + \log \phi\left(\frac{w-z\alpha}{\sigma}\right) + \log \Phi\left(\frac{x\beta+\rho\left(\frac{w-z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) - \log \Phi(x\beta).$$

One could estimate this model by maximizing the sample sum of the full likelihood function *l*, by maximizing the sample sum of either the marginal or the conditional component, or by maximizing these components in sequence.

Note that asymptotically efficient estimation requires maximizing the full likelihood, and that not all the parameters are identified in each component; e.g., only β is identified from the marginal component.

Nevertheless, there may be computational advantages to working with the marginal or conditional likelihood. Maximization of l^1 is a conventional binomial probit problem. Maximization of l^2 could be done either jointly in all the parameters α, β, ρ, σ; or alternately in α, ρ, σ, with the estimate of β from a first-step binomial probit substituted in and treated as fixed.

When $\rho = 0$, the case of "exogenous" selection in which there is no correlation between the random variables determining selection into the observed population and the level of the observation, l^2 reduces to the log likelihood for a regression with normal disturbances. When $\rho \neq 0$, selection matters and regressing of w on z will not give consistent estimates of α and σ.

An alternative to maximum likelihood estimation is a GMM procedure based on the moments of w. Using the property that the conditional expectation of ν given y = 1 equals the conditional expectation of ν given ε, integrated over the conditional density of ε given y = 1, plus the property of the normal that $d\phi(\varepsilon)/d\varepsilon = -\varepsilon\cdot\phi(\varepsilon)$, one has

$$
\begin{aligned}
(35) \qquad \mathrm{E}\{w|z,y=1\} &= z\alpha + \sigma \mathrm{E}\{\nu|y=1\} \\
&= z\alpha + \sigma \int_{-x\beta}^{+\infty} \mathrm{E}\{\nu|\varepsilon\}\phi(\varepsilon)d\varepsilon/\Phi(x\beta) \\
&= z\alpha + \sigma\rho \int_{-x\beta}^{+\infty} \varepsilon\phi(\varepsilon)d\varepsilon/\Phi(x\beta) \\
&= z\alpha + \sigma\rho\phi(x\beta)/\Phi(x\beta) \\
&\equiv z\alpha + \lambda M(x\beta),
\end{aligned}
$$

where $\lambda = \sigma\rho$ and $M(c) = \phi(c)/\Phi(c)$ is called the inverse Mill's ratio.

Further,

$$E(\nu^2|\varepsilon) = \mathrm{Var}(\nu|\varepsilon) + \{E(\nu|\varepsilon)\}^2$$
$$= 1 - \rho^2 + \rho^2\varepsilon^2,$$

$$\int_{-c}^{+\infty} \varepsilon^2\phi(\varepsilon)d\varepsilon = - \int_{-c}^{+\infty} \varepsilon\phi'(\varepsilon)d\varepsilon$$

$$= -c\phi(c) + \int_{-c}^{+\infty} \phi(\varepsilon)d\varepsilon$$

$$= -c\phi(c) + \Phi(c),$$

(36) $$E\{(w-z\alpha)^2|z,y=1\} = \sigma^2E\{\nu^2|y=1\}$$

$$= \sigma^2 \int_{-x\beta}^{+\infty} E\{\nu^2|\varepsilon\}\phi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= \sigma^2 \int_{-x\beta}^{+\infty} \{1 - \rho^2 + \rho^2\varepsilon^2\}\phi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= \sigma^2\{1 - \rho^2 + \rho^2 - \rho^2x\beta\phi(x\beta)/\Phi(x\beta)\}$$
$$= \sigma^2\{1 - \rho^2x\beta\phi(x\beta)/\Phi(x\beta)\}$$
$$= \sigma^2\{1 - \rho^2x\beta\cdot M(x\beta)\}.$$

Then,

(37) E $\left\{[w - z\alpha - \boldsymbol{E}\{w - z\alpha \,|\, z, y = 1\}]^2 | z, y = 1\right\}$ =

$E\{(w-z\alpha)^2 | z,y=1\} - [E\{w-z\alpha | z,y=1\}]^2$
$= \sigma^2\{1 - \rho^2 x\beta\phi(x\beta)/\Phi(x\beta) - \rho^2\phi(x\beta)^2/\Phi(x\beta)^2\}$

$= \sigma^2\{1 - \rho^2 M(x\beta)[x\beta + M(x\beta)\}.$

A GMM estimator for this problem can be obtained by applying NLLS, for the observations with y = 1, to the equation

$$w = z\alpha + \sigma\rho M(x\beta) + \zeta, \tag{38}$$

where ζ is a disturbance that satisfies $E\{\zeta | y=1\} = 0$. This ignores the heteroskedasticity of ζ, but it is nevertheless consistent. This regression estimates only the product $\lambda \equiv \sigma\rho$.

Consistent estimates of σ and ρ could be obtained in a second step:

(39) $V\{\zeta \mid x,z,y=1\}$

$$= \sigma^2\{1 - \rho^2 M(x\beta)[x\beta + M(x\beta)]\},$$

Estimate σ^2 by regressing the square of the estimated residual, $\zeta_e^{\,2}$,

(40) $\zeta_e^{\,2} = a + b\{M(x\beta_e)[x\beta_e + M(x\beta_e)]\} + \xi$

provide consistent estimates of σ^2 and $\sigma^2\rho^2$, respectively.

There is a two-step estimation procedure, due to Heckman, that requires only standard computer software, and is widely used:

[1] Estimate the binomial probit model,

$$P(y|x,\beta) = \Phi(yx\beta) \,, \tag{42}$$

by maximum likelihood.

[2] Estimate the linear regression model,

$$w = z\alpha + \lambda M(x\beta_e) + \zeta, \tag{43}$$

where $\lambda = \sigma\rho$ and the inverse Mill's ratio M is evaluated at the parameters estimated from the first stage.

To estimate σ and ρ, and increase efficiency, one can do an additional step,

[3] Estimate σ^2 using the procedure described in (40), with estimates λ_e from the second step and β_e from the first step.

One limitation of the bivariate normal model is most easily seen by examining the regression (43). Consistent estimation of the parameters α in this model requires that the term M(xβ|) be estimated consistently. This in turn requires the assumption of normality, leading to the first-step probit model, to be exactly right. Were it not for this restriction, estimation of α in (43) would be consistent under the much more relaxed requirements for consistency of OLS estimators. To investigate this issue further, consider the bivariate selection model (29) with the following more general distributional assumptions: (i) ε has a density $f(\varepsilon)$ and associated CDF $F(\varepsilon)$; and (ii) ν has $E(\nu|\varepsilon) = \rho\varepsilon$ and a second moment $E(\nu^2|\varepsilon) = 1 - \rho^2$ that is independent of ε. Define the <u>truncated</u> <u>moments</u>

$$J(x\beta) = E(\varepsilon|\varepsilon>-x\beta)$$

$$= \int_{-x\beta}^{\infty} \varepsilon f(\varepsilon)d\varepsilon/[1 - F(-x\beta)]$$

and

$$K(x\beta) = E(1 - \varepsilon^2|\varepsilon>-x\beta)$$

$$= \int_{-x\beta}^{\infty} [1 - \varepsilon^2]f(\varepsilon)d\varepsilon/[1 - F(-x\beta)]\,.$$

Then, given the assumptions (i) and (ii),

$$E(w|z,y=1) = z\alpha + \sigma\rho E(\varepsilon|\varepsilon>-x\beta)$$
$$= z\alpha + \sigma\rho J(x\beta),$$
$$E((w - E(w|z,y=1))^2|z,y=1)$$
$$= \sigma^2\{1 - \rho^2[K(x\beta) + J(x\beta)^2]\}.$$

Thus, even if the disturbances in the latent variable model were not normal, it would nevertheless be possible to write down a regression with an added term to correct for self-selection that could be applied to observations where y = 1:

(45) $$w = z\alpha + \sigma E\{\nu|x\beta+\varepsilon>0\} + \zeta$$
$$= z\alpha + \sigma\rho J(x\beta) + \zeta,$$

where ζ is a disturbance that has mean zero and the heteroskedastic variance

$$E(\zeta^2|z,y=1)) = \sigma^2\{1 - \rho^2[K(x\beta) + J(x\beta)^2]\}.$$

Now suppose one runs the regression (37) with an inverse Mill's ratio term to correct for self-selection, when in fact the disturbances are not normal and (44) is the correct specification. What bias results? The answer is that the closer $M(x\beta)$ is to $J(x\beta)$, the less the bias. Specifically, when (44) is the correct model, regressing w on z and $M(x\beta)$ amounts to estimating the misspecified model

$$w = z\alpha + \lambda M(x\beta) + \{\zeta + \lambda[J(x\beta) - M(x\beta)]\} .$$

The bias in NLLS is given by

$$\begin{bmatrix} \hat{\alpha} - \alpha \\ \lambda_e - \lambda \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{Ez}'z & \boldsymbol{Ez}'M \\ \boldsymbol{EMz} & \boldsymbol{EM}^2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{Ez}(J-M) \\ \boldsymbol{EM}(J-M) \end{bmatrix}$$

this bias is small if $\lambda = \sigma\rho$ is small or the covariance of J - M with z and M is small.