CHAPTER 2. SAMPLING AND SELECTION

1. INTRODUCTION

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. Several strands in the econometrics literature have investigated estimators appropriate to such data: seminal papers of Heckman (1974) on sample selection, and Manski and Lerman (1977) on choice-based sampling; further work on endogenous stratification by Hausman and Wise (1977), Manski and McFadden (1981), Cosslett (1981), and Hsieh, Manski, and McFadden (1984); and related work on switching regression by Goldfeld and Quandt (1973, 1975), Madalla and Nelson (1974), and Lee and Porter (1984). This chapter synthesizes this literature, and provides machinery that can be used to crank out estimators for a variety of biased sampling problems.

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.[1]

The estimation problem facing an econometrician can be described, schematically, in terms of a contingency table relating a vector of exogenous variables z and a vector of endogenous variables y, as in the table below where each column and row corresponds to different values for the vector of variables. The joint distribution of (z,y) in the population is a probability

$$(1) \qquad p(z,y) \equiv P(y|z,\beta_o)p(z) \equiv Q(z|y)q(y),$$

where $P(y|z,\beta_o)$ is the *conditional probability* of the endogenous vector y, given the exogenous vector z, defined as a member of a parametric family with true parameter vector β_o; p(z) is the *marginal distribution* of the exogenous variables, obtained by a row sum in the table; q(y) is the *marginal distribution* of y, obtained by a column sum in the table; and $Q(z|y)$ is the *conditional distribution* of z given y, defined by Bayes law in equation (1).[2] We identify $P(y|z,\beta_o)$ as the *structural* model of econometric interest; where by "structural" we mean that this conditional probability law is *invariant* in different populations or policy environments where the marginal distribution of z is altered. A structural model will result if there is a *stable causal relationship* from z to y, with no contemporaneous feedback from y to z. One would expect this to be the case if z describes the environment of an economic agent (e.g., prices, income) and y describes the agent's

---

[1] Stratification may in itself be economical, permitting the contacting and interviewing of subjects at reduced cost. In addition, stratification may concentrate observations in areas yielding high information on the behavior of economic interest.

[2] In this chapter, we will treat the data vector (z,y) as discrete. There is no fundamental change if some components of (z,y) are continuous; it is merely necessary to replace summations with integrals with respect to appropriate continuous or counting measures. There are additional technical assumptions required to assure measurability and integrability when some components are continuous.

behavioral response (e.g., occupation choice, hours of labor supplied). However, there are many economic applications where it is a reasonable approximation for policy analysis to assume $P(y|z,\beta_o)$ is a "reduced form" with the needed invariance property, without invoking strict assumptions on causality.

	y_1	y_2	.....	y_J	
z_1	$P(y_1\|z_1,\beta_o)p(z_1)$	$P(y_2\|z_1,\beta_o)p(z_1)$	.....	$P(y_J\|z_1,\beta_o)p(z_1)$	$p(z_1)$
z_2	$P(y_1\|z_2,\beta_o)p(z_2)$	$P(y_2\|z_2,\beta_o)p(z_2)$	.....	$P(y_J\|z_2,\beta_o)p(z_2)$	$p(z_2)$
:	:	:		:	:
z_K	$P(y_1\|z_K,\beta_o)p(z_K)$	$P(y_2\|z_K,\beta_o)p(z_K)$	.....	$P(y_J\|z_K,\beta_o)p(z_K)$	$p(z_K)$
	$q(y_1)$	$q(y_2)$	.....	$q(y_J)$	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.[3] In this sample, z is *ancillary* to β_o, and the observation that it can be conditioned out without loss of information on β_o can be elevated to a general principle of statistical inference (Cox and Hinckley, 1974).

We next introduce a notation for stratified or biased samples. Suppose the data are collected from one or more *strata*, indexed s = 1,..., S. Each stratum is characterized by a sampling protocol that determines the segment of the population that qualifies for interviewing. Define R(z,y,s) to be the *qualification probability* that a population member with characteristics (z,y) will qualify for the subpopulation from which the stratum s subsample will be drawn. Examples of sampling protocols and their characterizations in terms of qualification probabilities follow:

1. Simple random subsample, with $R(z,y,s) \equiv 1$.

2. Exogenous stratified sampling, 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. The set A_s might define a location, such as a census tract, or a socioeconomic characteristic such as race. The protocol for identifying the qualified subpopulation under locational stratification is typically to enumerate the response units at a location, and then sample randomly from this enumeration. In the contingency table, this corresponds to sampling randomly from one or more rows. The protocol for identifying the qualified subpopulation using a socioeconomic criterion is typically a screening interview. Exogenous stratified sampling can be generalized to differential rates by permitting R(z,y,s) to be any function from (z,s) into the unit interval;

[3] The log likelihood of an observation is $\log P(y|z,\beta) + \log p(z)$, and the kernel of this log likelihood is the part that depends on the parameter vector β.

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. Endogenous stratified sampling, with $R(z,y,s) = 1$ 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, such as the subpopulation whose appliance and energy consumption choices include an air conditioner. Alternately, B_s might identify a range of a continuous response, such as an income category. A classical choice-based sample for discrete response is the case where each response corresponds to a different stratum. In Figure 1, endogenous sampling corresponds to sampling randomly from one or more columns. Endogenous samples with strata corresponding to single columns are called pure choice-based samples. Endogenous stratified sampling can be generalized to qualification involving both exogenous and endogenous variables, with B_s defined in general as a subset of $Z \times Y$. 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 joint probability that a member of the population will have variables (z,y) and will qualify for stratum s is $R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)$. Then for stratum s, the proportion of the population qualifying into the stratum, or *qualification factor*[4], *is*

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

and the conditional distribution of (z,y) given qualification is

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

A sample from stratum s is governed by the probability law $G(z,y|s)$. Note that $G(z,y|s)$ depends on the unknown parameter vector β and on the distribution p(z) of the explanatory variables. In simple cases of stratification, such as Examples 1-3 above, R(z,y,s) is fully specified by the sampling protocol. The qualification factor r(s) may be known, for example when stratification is based on census tract with known sizes; estimated from the survey, for example when 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.

---

[4] The inverse of the qualification factor is called the *raising 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).[5] Then, the log likelihood for the stratified sample is

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

This likelihood does not include screening or auxiliary data on the qualification factors, which will be informative if these factors are unknown.

2. EXOGENOUS STRATIFIED SAMPLING

When the qualification probability R(z,y,s) is independent of y, the qualification factor $r(s) = \sum_z R(z,s)p(z)$ is independent of β_o, and the log likelihood function (4) 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. This implies that estimators designed for random samples will have the same properties in exogenously stratified samples. The information matrix for the likelihood function under exogenous stratification,

$$(6) \qquad J = \sum_{s=1}^{S} \mu_s \sum_{z} \frac{R(z,s)p(z)}{r(s)} \sum_{y} P(y|z,\beta_o)\cdot[\nabla_\beta \text{Log } P(y|z,\beta_o)]\cdot[\nabla_\beta \text{Log } P(y|z,\beta_o)]',$$

depends on the sample design. Then, exogenous stratification can be used to increase the information available in a sample of given size; this is precisely the objective of classical experimental design.

3. 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). Further, any unknowns in the qualification probability also enter the kernel. There are four possible strategies for estimation under these conditions:

---

[5] Note that $n(z,y|s)/n_s$ is the empirical probability measure for a random sample of size n_s from the population with law $G(z,y|s)$. In the case of discrete variables with a finite number of configurations, the $n(z,y|s)$ are simply cell counts. Nothing is changed for continuous variables, except that technically one must consider stochastic limits of empirical processes.

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, and allows one to back out the structural parameters. This approach is straightforward and needs no further discussion for small problems, but is burdensome or infeasible when the Z variables have many dimensions or categories, or are continuous.
2. Weighted Exogenous Sample Maximum Likelihood -- This is a pseudo-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 random.
3. Conditional Maximum Likelihood -- 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. Full Information Maximum Likelihood -- This approach estimates p(z) nonparametrically as a function of the remaining parameters, and substitutes to concentrate the likelihood as a function of the finite parameter vector.

4. WEIGHTED EXOGENOUS SAMPLE MAXIMUM LIKELIHOOD (WESML)

Recall that 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 pseudo- maximum likelihood criterion based on (5),

$$(7) \qquad W(\beta) = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot w(z,y,s)\cdot \text{Log } P(y|z,\beta),$$

where w(z,y,s) is a weight introduced to achieve consistency. Assume that $n_s/N \to \mu_s$ as $N \to \infty$. Then, using the notation "$\to_{as}$" to denote almost sure convergence,

$$(8) \qquad n(z,y|s)/N \equiv [n(z,y|s)/n_s]\cdot[n_s/N] \to_{as} G(z,y|s)\mu_s,$$

implying from (3) that

$$(9) \qquad W(\beta)/N \to_{as} \sum_{s=1}^{S} \mu_s \sum_{z} \sum_{y} G(z,y|s)\cdot w(z,y,s)\cdot \text{Log } P(y|z,\beta)$$

$$= \sum_{z} p(z)\cdot \sum_{y} \{ \sum_{s=1}^{S} R(z,y,s)w(z,y,s)\mu_s/r(s)\}\cdot P(y|z,\beta_o)\cdot \text{Log } P(y|z,\beta).$$

A sufficient condition for consistency of the pseudo-maximum likelihood estimator is that the bracketed term,

(10) $$\sum_{s=1}^{S} R(z,y,s)w(z,y,s)n_s/N\cdot r(s)$$

be independent of y. 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) $$w(z,y) = \left[\sum_{s=1}^{S} R(z,y,s)n_s/Nf(s)\right]^{-1};$$

these are well-defined if the bracketed expressions are positive, and R(z,y,s) contains no unknown parameters. These weights do not depend on the stratum from which the observation is drawn, but do depend generally on the endogenous variable y.

When the qualification probabilities R(z,y,s) are strictly positive for all (z,y) and all strata, and contain no unknowns, another set of possible weights is

(12) $$w(z,y,s) = 1/R(z,y,s).$$

These can be interpreted as reweighting observations in inverse proportion to the probability with which they qualify from the population, and are precisely the weighting most commonly used in classical survey research. When the weights (11) and (12) are both feasible, the weights (11) are more efficient.

A classical application of WESML estimation is to a sample in which the strata coincide with the possible configurations of y, so that $R(z,y,s) = \mathbf{1}(y = s)$. In this case, $w(z,y) = N\cdot f(y)/n_y$, the ratio of the population to the sample frequency. 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 $w(z,2) = N\cdot f(2)/[n_1\cdot f(2) + n_2]$.

When the r(s) are known, and f(s) ≡ r(s), the WESML estimator has an asymptotic covariance matrix $J_w^{-1}H_wJ_w^{-1}$, where

(13) $$J_w = -\sum_{s=1}^{S} (\mu_s/r(s)) \sum_{z} \sum_{y} w(z,y,s)R(z,y,s)P(y|z,\beta_o)p(z)\nabla_{\beta\beta}l\,,$$

(14) $$H_w = \sum_{s=1}^{S} \mu_s^2 \sum_{z} \sum_{y} w(z,y,s)^2[R(z,y,s)P(y|z,\beta_o)p(z)/r(s)]\cdot[\nabla_\beta l]\cdot[\nabla_\beta l]' - \sum_{s=1}^{S} q_s\cdot q_s'$$

where $l = \log P(y|x,\beta)$ and

$$q_s = \sum_{z} \sum_{y} \mu_s\, w(z,y,s)\cdot[R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)/r(s)]\nabla_\beta l,$$

and l and its derivatives are evaluated at β_o. These covariance terms come from a Taylor's expansion of the first-order conditions for maximization of $W(\beta)$, and can be estimated consistently by replacing terms with their sample analogs.

5. CONDITIONAL MAXIMUM LIKELIHOOD (CML)

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

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

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

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

Substituting (3) yields a formula independent of p(z),

$$(15) \qquad \Pr(y|z) = \frac{\sum_{s=1}^{S} R(z,y,s)\cdot P\ (y|z,\beta_o)\, n_{\ s}\ /N\, r\ (s)}{\sum_{y}\sum_{s=1}^{S} R(z,y,s)\cdot P\ (y|z,\beta_o)\, n_{\ s}\ /N\, r\ (s)} .$$

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, (15) is used directly. More generally, given auxiliary sample information on the r(s), these can be treated as parameters and estimated from the product of the likelihood (15) and the likelihood of the auxiliary sample.

For discrete response in which qualification does not depend on z, the formula (15) simplifies to $\Pr(y|z) = \frac{P(y|z,\beta_o)\cdot\alpha_y}{\sum_y P(y|z,\beta_o)\cdot\alpha_y}$, where $\alpha_y = \sum_{s=1}^{S} R(z,y,s)\cdot n_s/N\cdot r(s)$ 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 $\log \alpha_y$ and alternative-specific parameters in the original model. It remains necessary to use formulas for endogenous sampling to estimate the asymptotic covariance matrix consistently.

For the previous example of an enriched sample, one has $\Pr(1|z) = P(1|z,\beta_o)\cdot n_1/N\cdot D$ and $\Pr(2|z) = P(2|z,\beta_o)\cdot[n_1/N + n_2/N\cdot r(2)]/D$, where $D = n_1/N + P(2|z,\beta_o)\cdot n_2/N$. An example in a different context shows the breadth of application of (15). 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 $D = \gamma + (1-\gamma)\cdot P(y>y_o|z,\beta_o)$.

When the r(s) are known, the asymptotic covariance matrix of the CML estimator is $J_c^{-1}H_cJ_c^{-1}$, where

$$(16) \quad J_c = -\sum_{s=1}^{S} (\mu_s/r(s)) \sum_z \sum_y R(z,y,s)P(y|z,\beta_o)p(z)\nabla_{\beta\beta}c \ ,$$

$$(17) \quad H_w = \sum_{s=1}^{S} \mu_s^2 \sum_z \sum_y [R(z,y,s)P(y|z,\beta_o)p(z)/r(s)][\nabla_\beta c]\cdot[\nabla_\beta c] - \sum_{s=1}^{S} q_s q_s'$$

where $c = \log \Pr(y|z,\beta)$ and $q_s = \sum_z \sum_y \mu_s [R(z,y,s)P(y|z,\beta_o)p(z)/r(s)]\nabla_\beta c$, and c and its derivatives evaluated at β_o. Note that the structure of this covariance matrix is the same as that for WESML.

6. FULL INFORMATION CONCENTRATED MAXIMUM LIKELIHOOD (FICLE)

Formally, the likelihood (4) can be treated as a function of the unknown parameter vector β, any unknown parameters in the qualification probabilities, and the unknown multivariate density p(z), with this whole density treated as an unknown parameter, possibly infinite dimensional. This is a *semiparametric* estimation problem, in which a finite parameter vector is to be estimated in the presence of a possibly infinite-dimensional vector of nuisance parameters. In some applications, this can be done by direct formal maximization of the likelihood in p(z), given the remaining parameters, yielding a concentrated likelihood function of the finite parameter vector.
Let

$$(18) \quad L = \sum_{s=1}^{S} \sum_z \sum_y n(z,y|s)\cdot \text{Log } G(z,y|s)$$

$$+ \sum_{s=1}^{S} \lambda_s[r(s) - \sum_z \sum_y R(z,y,s)P(z,y,s)p(z)] + \lambda_o[1 - \sum_z p(z)]$$

be a Lagrangian for the formal maximization problem. Solving the first-order-condition for p(z) yields

$$(19) \quad p(z) = \left(\sum_{s=1}^{S} \sum_y n(z,y|s)\right) \Big/ \left(\sum_{s=1}^{S} \sum_y \lambda_s R(z,y,s)\cdot P(y|z,\beta_o) + \lambda_o\right) \ .$$

Substituting (19) into (18), simplifying, and dropping terms independent of the unknowns, yields

$$(20) \quad L_1 = \sum_{s=1}^{S} \sum_z \sum_y n(z,y|s)\cdot \text{Log} \ \frac{R(z,y,s)\cdot P(y|z,\beta)/r(s)}{N + \sum_{s=1}^{S} \lambda_s[\sum_y R(z,y,s)\cdot P(y|z,\beta) - r(s)]}$$

$$+ \sum_z \left[\sum_{s=1}^{S} \sum_y n(z,y|s)\right] \cdot \frac{\sum_{s=1}^{S} \lambda_s[r(s) - \sum_y R(z,y,s)\cdot P(y|z,\beta)]}{N + \sum_{s=1}^{S|} \lambda_s[\sum_y R(z,y,s)\cdot P(y|z,\beta) - r(s)]}$$

A joint critical point of this concentrated function in β and the λ_s gives the FICLE estimator. Cosslett (1981) has shown that estimators in this class are fully efficient. Since this is a semiparametric problem, Cosslett's argument required calculation by variational methods of the least information contained in the parametric part of the problem; this method in its general form provides what are now called the Wellner efficiency bounds. The asymptotic covariance matrix of the FICLE estimators has the same general structure as the previous estimators, but the specifics are complicated by the presence of the finite vector of nuisance parameters λ_s. For straightforward response-based endogenous samples, with y used to define non-overlapping strata, the FICLE criteria and the CML criteria can be manipulated into almost the same form, with $n_s/Nf(s)$ and λ_s/N appearing in analogous positions and converging to the same limit.

7. ENDOGENOUS SAMPLING IN THE MNL CASE

An important simplification of the CML method occurs for what in biostatistics is termed *logistic regression* in case-control designs. Suppose that the vector of covariates is partitioned into components $z = (v,x)$ with v discrete. (In biostatistics, v will often include variables such as age and gender whose distributions are matched between cases and controls.) Suppose that $P(y|v,x)$ has a *multinomial logit* form, $P(y|v,x) = \exp(\alpha_y+\gamma_{yv}+x\beta_y)/\sum_{y'}\exp(\alpha_{y'}+\gamma_{y'v}+x\beta_{y'})$. In this model, the β_y are slope coefficients for the covariates x, and α_y and γ_{yv} are response-specific effects and interactions of response-specific and v-specific effects. Suppose that the qualification probability $R(v,x,y,s)$ does not depend on x, but does depend on y and v through a sample design that first draws a stratum of *cases* for a specified y, and then draws strata of controls that are screened so that their distribution of v's "matches" the distribution of v's among the cases. For example, the controls may be matched in distribution with the cases on age and gender. For identification, a normalization such as $\alpha_1 = \gamma_{1v} = \beta_1 = 0$ is imposed. The conditional probability $g(y|z)$ is again of multinomial logit form, with the same β_y parameters but with the remaining parameters shifted from their population values by sampling factors,

$$g(y|v,x) = \exp(\alpha^*_y+\gamma^*_{yv}+x\beta_y)/\sum_{y'}\exp(\alpha^*_{y'}+\gamma^*_{y'v}+x\beta_{y'}),$$

with $\alpha^*_y +\gamma^*_{yv}= \alpha_y +\gamma_{yv} +\log(\sum_s R(v,y,s)\cdot f(s)/r(s))$. Note that consistent estimation of this model requires the inclusion of all the alternative-specific effects and interactions that are modified by sampling factors. However, if these variables are included, then the slope parameters β_y are estimated consistently without further adjustments for endogenous sampling. (If the raising factors are estimated rather than known, there is an additional contribution to the asymptotic covariance matrix; see Hsieh, Manski, and McFadden (1985). If the model already contained v-effects and v-y interaction effects, then no modification to a random sampling likelihood is needed to estimate the β_y parameters consistently. (Of course, in this case, the estimates of the main and interaction effects will incorporate the effects of the case/control sample design.) If the model contains interactions of

v and x in addition to v-effects and v-y effects, then the coefficients on these interactions would contain additional sampling factors that must be removed to obtain consistent estimates of the corresponding population interactions.) The simplification above was first noted by Anderson (1972).

8. EXTENSIONS AND CONCLUSIONS

Both the WESML and CML estimators are computationally practical in a variety of endogenous sampling situations, and have been widely used. In general, neither estimator dominates the other. Monte Carlo experience is that the WESML estimator is more efficient when the weights for different alternatives are nearly the same, and that CML is more efficient when the weights differ substantially across alternatives. The FICLE estimator has not been widely used.

When the population qualification factors r(s) are unknown, and consistently estimated by f(s) obtained from auxiliary data, then the estimators described above are consistent. However, in computing the asymptotic covariance matrices of the estimators, it is necessary to take account of presence of estimated quantities in estimation criterion. This will in general contribute additional terms to the asymptotic covariance matrix; see Newey and McFadden (1995). A more efficient procedure is to estimate the r(s) jointly using the sample and auxiliary data. Hsieh, Manski, and McFadden (1985) develop the procedures for doing this.

Extensions of the theory of endogenous sampling can be made to more complex applications, and to more complex sources of auxiliary information, such as duration data (with length-biased sampling) and endogenously recruited panel data,; see Lancaster and Imbens (1990) and McFadden (1996).

9. SELECTION

There are a variety of econometric problems where dependent variables are discrete, censored at lower or upper limits, or truncated or selected so they are not always observed. It is often convenient to model the behavior of such variables as the result of a two-stage process,

$$\begin{bmatrix} \textit{Exogenous} \\ \textit{Variables} \end{bmatrix} \longrightarrow \begin{bmatrix} \textit{Latent} \\ \textit{Dependent Variables} \end{bmatrix} \longrightarrow \begin{bmatrix} \textit{Observed} \\ \textit{Dependent Variables} \end{bmatrix},$$

where there are intermediate unobserved (latent) variables that are in the first stage determined by exogenous variables through a conventional linear model, and observed dependent variables that in the second stage are determined by some non-linear mapping. The structure of the first mapping, the dimensionality of the latent variables, and the structure of the non-linear mapping can all be varied to fit particular applications. Historically, latent variable models come from psychometrics, where both the mappings from exogenous variables to latent variables, and from latent variables to observed dependent variables are linear, and the critical feature is that the dimensionality of the latent variables is much lower than the dimensionality of the observed dependent variables. A classical psychometric application is to ability testing, where the observed dependent variables are responses to test items, and the latent variables are *factors* such as verbal, quantitative, and motor abilities. In their most general form, these are called Multiple-Indicator, Multiple Cause (MIMC)

models, and analysis of the mapping from latent to observed dependent variables is called *factor analysis*. An example of an economic application of MIMC models is the Friedman permanent income hypothesis, where the observed dependent variables are measured yearly incomes and there is a single latent variable, permanent income. These lecture notes will discuss the second major application of latent variable models, to situations where the mapping from latent to observed dependent variables is nonlinear, and the observed dependent variables are not necessarily continuous.

A fairly general notation for a model with m latent variables for each observation unit is $y_j^* = x_j\beta + \varepsilon_j$, where j = 1,...m. This can be written more compactly in matrix notation as $y^* = X\beta + \varepsilon$, where $y^* \in \mathbb{R}^m$ is a m×1 vector of latent variables for one observation, X is a m×k array of explanatory variables whose rows are the x_j vectors, β is a k×1 vector of parameters, and ε is a m×1 vector of disturbances with a multivariate density $f(\varepsilon|\theta)$ that contains additional parameters θ. This notation can accommodate β parameters that differ across equations by introducing variables in each x_j in interaction with dummies for the different equations. The observed dependent variables are given by a mapping $y = h(y^*)$ that is in general nonlinear and many-to-one. Some examples illustrate the possibilities, and indicate the scope of possible applications:

(1) $$y^* \in \mathbb{R}^1 \text{ and } y = h(y^*) = \begin{cases} +1 & if\ y^* \geq 0 \\ -1 & if\ y^* < 0 \end{cases}$$

generates a binomial response model. An application might be to firms' decisions to go bankrupt or stay in business, where y^* is latent *expected* profit; see also application (5) below.

(2) $$y^* \in \mathbb{R}^1 \text{ and } y = h(y^*) = \begin{cases} y^* & if\ y^* \geq 0 \\ 0 & if\ y^* < 0 \end{cases}$$

generates a *censored data* (*Tobit*) model. An application might be to expenditure on clothing in a one-week observation period, where zeros are common.

(3) $$y^* \in \mathbb{R}^1 \text{ and } y = h(y^*) = \begin{cases} y^* & if\ y^* \geq c \\ NA & if\ y^* < c \end{cases},$$

where NA means no observation is available and c is a constant, generates a *truncated data* model. An application might be to competitive (among buyers) auction prices for units of a good, where a transaction is observed only if a bid exceeds a reservation price c. In case $y^* < c$, one may in one variant of this model observe x, and in another variant observe nothing about x.

(4) $y^* \in \mathbb{R}^1$ and $y = h(y^*)$ is given by $y = i$ if $\lambda_i \leq y^* < \lambda_{i+1}$ for i = 0,...,J, with $\lambda_0 = -\infty$ and $\lambda_{J+1} = +\infty$,

where λ_1 to λ_J are parameters. This mapping generates an ordered response or *count* model. An application might be to household choice of number of children, or to wealth or income within brackets established by the questionnaire.

(5) $$y^* \in \mathbb{R}^2 \text{ and } y = h(y^*) = \begin{cases} (+1, y_2^*) & if\ y_1^* \geq 0 \\ (-1, NA) & if\ y_1^* < 0 \end{cases}$$

has the following interpretation: if $y_1^* \geq 0$, then $y_1 = +1$ is an indicator for this, and $y_2 = y_2^*$ is observed. If $y_1^* < 0$, then $y_1 = -1$ is an indicator for this, and y_2 is not observed. Variants may have x_2 observed or not when y_2 is unobserved. An application is to bankruptcy decisions of the firm, where y_1^* is expected profit and y_2^* is realized profit. This is termed a *bivariate selection* model.

(6) $y^* \in \mathbb{R}^m$ and $y = h(y^*)$ is a mapping from $\mathbb{R}^m$ into $\{1,...,m\}$, where $y = i$ if $y_i^* \geq y_j^*$ for $j \neq i$.

This generates a *multinomial* response model in which the observed response corresponds to the maximum of the latent variables. An application might be to choice of occupation.

(7) $y^* \in \mathbb{R}^m$ and $y = h(y^*)$ is a mapping from $\mathbb{R}^m$ into $\{-1,+1\}^m$, with $y_j = +1$ if $y_j^* \geq 0$, and $y_j = -1$ otherwise.

This generates a *multivariate binomial* response model. An application might be to panel data on employment status.

(8) $y^* \in \mathbb{R}^m$ and $y = h(y^*)$ is a mapping from $\mathbb{R}^m$ into $\{0,1,2,...\}^m$, with $y_j = k_j$ for an integer k_j if $\lambda_{ij} \leq y_{i,j+1} < \lambda_{i,j+1}$.

This is a multivariate ordered response or *count* model. An application is to numbers of units purchased of each of m goods.

Let $\mathbf{A}(y)$ denote the set of y^* that map into observation y; this can be written as $\mathbf{A}(y) = h^{-1}(y)$, where h^{-1} denotes the inverse of the (possibly) many-to-one mapping h. Then, the probability of an observation can be written

$$g(y|X,\beta,\theta) = \int_{A(y)} f(y^*-X\beta|\theta)dy^*.$$

The integral should be interpreted as extending over the dimensions where the condition $y^* \in h^{-1}(y)$ gives a range of values. In the Tobit example (2) above, $y = 0$ implies $h^{-1}(0) = (-\infty,0]$, and the integral is over this interval. However, $y > 0$ implies $h^{-1}(y) = y$, and $g(y|X,\beta,\theta) = f(y-X\beta|\theta)$ without integration. In the bivariate selection model (5), the observation $(+1,y_2)$ requires integration in one dimension, $g((+1,y_2)|X,\beta,\theta) = \int_0^{+\infty} f(y_1^*-x_1\beta,y_2-x_2\beta|\theta)dy_1^*$, while the observation (-1,NA) requires integration in both dimensions, $g((-1,NA)|X,\beta,\theta) = \int_{-\infty}^{0} \int_{-\infty}^{+\infty} f(y_1^*-x_1\beta,y_2^*-x_2\beta|\theta)dy_1^*dy_2^*$.

Consider the log likelihood of an observation, $l(\beta,\theta) = \log g(y|X,\beta,\theta)$. The *score* with respect to the parameters $\gamma = (\beta,\theta)$ is

$$\nabla_\gamma l(\beta,\theta) = \frac{\int_{A(y)} \left\{\nabla_\gamma \log f(y^*-X\beta|\theta)\right\} f(y^*-X\beta|\theta)dy^*}{\int_{A(y)} A(y) f(y^*-X\beta|\theta)dy^*}$$

$$= \mathbf{E}\ \left\{\nabla_\gamma \log f(y^* - X\beta \,|\, \theta) | y^* \in h^{-1}(y)\right\}\ ;$$

that is to say, *the score of the observation* y *can be expressed as the conditional expectation of the score of the latent variable model, conditioned on the event that the latent vector yields* y. If these integrals can be evaluated analytically or numerically, then it is usually feasible to do maximum likelihood estimation of the parameters. Even when the integrals are intractable, it may be possible to approximate them by simulation methods.

The basic latent variable model setup above can be extended in several ways. For time-series or panel data, X may contain variables determined by lagged latent variables. If disturbances are serially correlated, one confronts all the problems of identification, stationarity, and consistent estimation that occur in conventional linear systems, plus additional problems of dealing with initial conditions. The leading author who has worked on these problems is Heckman. The latent variable model can also be extended to have a more full-blown simultaneous-equations form, with complex paths linking observed and latent variables, with a *multiple-indicator, multiple-cause* structure. Leading authors on MIMC models are Goldberger and Joreskog.

10. THE BIVARIATE SELECTION PROBLEM

An important economic application of latent variable models is to the problem of *selection*: Who or what we can observe about economic agents is influenced by their behavior, so that our data are not representative of the whole population. Our analysis needs to correct for the effects of selection if we are to make consistent inferences about the population. A classic example of selection occurs in the study of wages and hours worked of married women. These variables are observed only for women who are working, but the same economic factors that determine these variables also influence the decision to work. For example, an unobserved disturbance that gives Mrs. Smith a higher-than-average potential wage and Mrs. Jones a lower than-average potential wage is more likely to induce Mrs. Smith into the labor force than Mrs. Jones. Then, a regression of wage on family characteristics using data for workers will typically overestimate the potential wage of non-workers. The econometric analysis of this problem provides a good tutorial for a broad spectrum of selection problems that arise because of economic behavior or because of survey design (e.g., deliberate stratification).

Consider a bivariate latent variable model with normal disturbances,

$$\begin{aligned} y^* &= x\beta + \varepsilon\ , \\ w^* &= z\alpha + \sigma\nu\ , \end{aligned} \tag{21}$$

where x and z are vectors of exogenous variables, not necessarily all distinct, α and β are parameter vectors, again not necessarily all distinct, and σ is a positive parameter. The interpretation of y^* is latent desired hours of work, and of w^* is latent log potential wage. The disturbances ε and ν have a standard bivariate normal distribution

$$\begin{bmatrix} \varepsilon \\ \nu \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)\ , \tag{22}$$

with zero means, unit variances, and correlation ρ.

There is a nonlinear *observation rule* determined by the application that maps the latent variables into observations. A typical rule might be "Observe y = 1 and w = w^* if $y^* > 0$; observe y = -1 and do not observe w when $y^* \le 0$". This could correspond, for example, to an application where the event of working (y = 1) or not working (y = 0) is observed, but actual hours worked are not, and the wage is observed only if the individual works ($y^* > 0$). It is sometimes convenient to 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. The probability of working is $P(y=1|x) = P(\varepsilon > -x\beta) = \Phi(x\beta)$, and of not working is $P(y=-1|x) = P(\varepsilon \le -x\beta) = \Phi(-x\beta)$, where Φ is the standard univariate cumulative normal. This can be written compactly as

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

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

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

and

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

The joint density can be written as the product of the marginal density of one component times the conditional density of the other,

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

The density of (y^*,w^*) can then be written

$$(23) \qquad f(y^*,w^*) = \frac{1}{\sigma}\,\varphi\left(\frac{w^*-z\alpha}{\sigma}\right)\cdot \frac{1}{\sqrt{1-\rho^2}} \cdot\varphi\left(\frac{y^*-x\beta-\rho(w^*-z\alpha)/\sigma}{\sqrt{1-\rho^2}}\right)$$

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

Now consider the log likelihood of an observation, $l(\alpha,\beta,\sigma,\rho)$. In the case of a non-worker (y = -1 and w = NA), the density (23) is integrated over $y^* < 0$ and all w^*. Using the second form in (23), this gives probability $\Phi(-x\beta)$. In the case of a worker, the density (23) is integrated over $y^* \ge 0$. Using the first form in (23)

$$(24)\qquad e^{l(\alpha,\beta,\sigma,\rho)} = \begin{cases} \Phi(-x\beta) & if\ y = -1 \\ \frac{1}{\sigma}\varphi\left(\frac{w-z\alpha}{\sigma}\right)\cdot\Phi\left(\frac{x\beta+\rho\left(\frac{w-z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) & if\ y = 1 \end{cases} .$$

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,

$$(25)\qquad l^1(\beta) = \log \Phi(yx\beta),$$

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

$$(26)\qquad l^2(\alpha,\beta,\sigma,\rho) = -\log\sigma + \log\varphi\left(\frac{w-z\alpha}{\sigma}\right) + \log\Phi\left(\frac{x\beta+\rho\left(\frac{w-z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) - \log\Phi(x\beta).$$

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, at least in the first step of estimation. Maximization of l^1 is a conventional binomial probit problem, which can be done easily with many canned programs. 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. The first case, maximization of l^2 in all the parameters, provides estimates whose variances are estimated by the inverse of the information matrix for l^2. The maximization of l^2 with an estimate of β substituted in requires use of the formula for the variance of a GMM estimator containing an embedded estimator; see the lecture notes on this topic. Neither of these procedures is fully efficient, and the two methods cannot be ranked in terms of efficiency.

When ρ = 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, note that l^2 reduces to the log likelihood for a regression with normal disturbances, implying that the maximum likelihood estimates for α and σ will be the OLS estimates. However, when $\rho \neq 0$, selection matters and regressing of w on z will <u>not</u> 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 $d\varphi(\varepsilon)/d\varepsilon = -\varepsilon\cdot\varphi(\varepsilon)$, one has

(27) $\mathbf{E}\{w|z,y=1\} = z\alpha + \sigma\mathbf{E}\{v|y=1\} = z\alpha + \sigma \int_{-x\beta}^{+\infty} \mathbf{E}\{v|\varepsilon\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$

$= z\alpha + \sigma\rho \int_{-x\beta}^{+\infty} \varepsilon\varphi(\varepsilon)d\varepsilon/\Phi(x\beta) = z\alpha + \sigma\rho\varphi(x\beta)/\Phi(x\beta) \equiv z\alpha + \lambda M(x\beta),$

where $\lambda = \sigma\rho$ and $M(c) = \varphi(c)/\Phi(c)$ is called the inverse Mill's ratio. (As a computational note, it is much better when calculating M to use a direct approximation to this function, rather than taking the ratio of computational approximations to φ and Φ.) Further, using the relationship

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

and the integration-by-parts formula

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

one obtains

(28) $\mathbf{E}\{(w-z\alpha)^2|z,y=1\} = \sigma^2\mathbf{E}\{v^2|y=1\} = \sigma^2 \int_{-x\beta}^{+\infty} \mathbf{E}\{v^2|\varepsilon\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$

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

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

Then,

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

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

It is possible to go on and compute higher moments, using the recursion formula:

$$\mu(c,k,\lambda) \equiv \mathbf{E1}(\varepsilon>c)\cdot(\varepsilon-\lambda)^k = \int_{\varepsilon=c}^{\infty} (\varepsilon-\lambda)^k\varphi(\varepsilon)d\varepsilon = -(c-\lambda)^{k-1}\varphi(c) - \lambda\cdot\mu(c,k-1,\lambda) + (k-1)\cdot\mu(c,k-2,\lambda).$$

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

(30) $w = z\alpha + \sigma\rho M(x\beta) + \zeta,$

where ζ is a disturbance that satisfies $\mathbf{E}\{\zeta|y=1\} = 0$. This ignores the heteroskedasticity of ζ, but it is nevertheless consistent. This regression estimates only the product $\lambda \equiv \sigma\rho$, but consistent estimates of σ and ρ could be obtained in a second step: The formula for the variance of ζ,

(31) $\mathbf{V}\{\zeta|x,z,y=1\} = \sigma^2\{1 - \rho^2 M(x\beta)[x\beta + M(x\beta)]\},$

suggests obtaining an estimate of σ^2 by regressing the square of the estimated residual, ζ_e^2, on one and the variable $M(x\beta_e)[x\beta_e + M(x\beta_e)]$, where β_e is the estimated parameter vector. Then, the estimated coefficients *a* and *b* in the regression

$$\zeta_e^2 = a + b\{M(x\beta_e)[x\beta_e + M(x\beta_e)]\} + \xi \tag{32}$$

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

The GMM estimator above is asymptotically inefficient because it fails to correct for heteroskedasticity, but more fundamentally because there are common parameters between the regression and the variance of the disturbances, and because the disturbance ζ is not normally distributed, so there is information in moments beyond the first two. The first of these inefficiencies could be eliminated by an estimated GLS-type transformation: From the first-step NLLS regression and the estimator of σ described above, calculate the weight

$$\tau^2 = 1 - \rho_e^2 M(x\beta_e)[x\beta_e + M(x\beta_e)],$$

and then rerun a weighted NLLS regression,

$$w/\tau = (z/\tau_e)\alpha + \sigma\rho(M(x\beta_e)/\tau_e) + (\zeta/\tau_e). \tag{33}$$

The variance of this regression is now σ^2, so that all the parameters of the original problem are estimated by the regression parameters plus the estimated variance of the regression.

The NLLS estimator above involves about the same amount of calculation as full maximum likelihood estimation, so that the latter method is usually preferable because it is asymptotically efficient, and the standard errors obtained from the information matrix are easier to calculate than the two-step GLS standard errors. However, there is an alternative two-step estimation procedure, due to Heckman, that requires only standard computer software, and is widely used:

[1] Estimate the binomial probit model,

$$P(y|x,\beta) = \Phi(yx\beta) \ , \tag{34}$$

by maximum likelihood.

[2] Estimate the linear regression model,

$$w = z\alpha + \lambda M(x\beta_e) + \zeta, \tag{35}$$

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 two additional steps,

[3] Estimate σ^2 using the procedure described in (12), with estimates λ_e from the second step and β_e from the first step; and

[4] Estimate the weighted linear regression model

(36) $$w/\tau = (z/\tau)\alpha + \lambda M(x\beta_e)/\tau + (\zeta/\tau),$$

where

$$\tau^2 = \{1 - \rho_e^2 M(x\beta_e)[x\beta_e + M(x\beta_e)]\},$$

and the parameters in this weight come from the first and second steps, plus

$$\rho_e^{\,2} = \lambda_e^{\,2}/\sigma_e^{\,2}$$

with λ_e^2 from step two and σ_e^2 from step three.

The standard errors of the first-step estimates β_e are obtained from the binomial probit maximum likelihood. However, the second-step estimates α_e and λ_e have standard errors that are not given correctly by the regression (35), both because the errors are heteroskedastic and because a right-hand-side variable contains embedded parameters from an earlier step; see the lecture notes on GMM estimation with embedded estimates for the formulas for the correct standard errors.

One limitation of the bivariate model is most easily seen by examining the regression (35). Consistent estimation of the parameters α in this model requires that the term $M(x\beta|)$ 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 (35) would be consistent under the much more relaxed requirements for consistency of OLS estimators. To investigate this issue further, consider the bivariate selection model (21) with the following more general distributional assumptions: (i) ε has a density $f(\varepsilon)$ and associated CDF $F(\varepsilon)$; and (ii) ν has $\mathbf{E}(\nu|\varepsilon) = \rho\varepsilon$ and a second moment $\mathbf{E}(\nu^2|\varepsilon) = 1 - \rho^2$ that is independent of ε. Define the truncated moments

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

and

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

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

$$\mathbf{E}(w|z,y=1) = z\alpha + \sigma\rho\mathbf{E}(\varepsilon|\varepsilon>-x\beta) = z\alpha + \sigma\rho J(x\beta),$$

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

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:

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

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

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

Now suppose one runs the regression (30) with an inverse Mill's ratio term to correct for self-selection, when in fact the disturbances are not normal and (36) is the correct specification. What bias results? The answer is that the closer M(xβ) is to J(xβ), the less the bias. Specifically, when (36) is the correct model, regressing w on z and M(xβ) 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.

Calculation for some standard distributions shows that when disturbances deviate from normal, M may not be a good approximation to J, implying that bias due to misspecification can be substantial. For example, consider as alternatives to the normal density for ε the logistic density,

$$f(\varepsilon) = e^{-a\varepsilon}/(1+e^{-a\varepsilon})^2, \; a = \frac{\sqrt{3}}{\pi},$$

and the bilateral exponential density,

$$f(\varepsilon) = (1/2\sqrt{2})\cdot e^{-|\varepsilon|/\sqrt{2}}.$$

For these densities, the function J can be calculated analytically. For the logistic density, one obtains $J(\varepsilon) = -\varepsilon + (1/a)\cdot\log(1+e^{a\varepsilon})\cdot(1+e^{-a\varepsilon})$, and for the bilateral exponential density, one obtains $J(\varepsilon) = e^{-c|\varepsilon|}\cdot(1 + c|\varepsilon|)/2cF(\varepsilon)$, where $F(\varepsilon) = \mathbf{1}(\varepsilon<0)\cdot e^{c\varepsilon} + \mathbf{1}(\varepsilon\geq 0)\cdot(1 - e^{-c\varepsilon})$ and $c = 2^{-1/2}$. The J(·) functions have the same qualitative shape for the normal, bilateral exponential, and logistic densities, but they are substantially shifted, so that there is at least significant bias to the estimated intercept in the regression if J is misspecified.

A natural question in semiparametric estimation is whether there is a robust method for estimating α that does not require that the distributions of ε and ν be fully parametric. It should be clear intuitively that approximating the unknown true J(·) function by a series of functions of ε, such as a low order polynomial in ε, should be sufficient to approximately span the space containing J(·), and that this in turn would be sufficient to eliminate for practical purposes any bias in estimation of α. The question would remain at to how many terms to use in an approximation.

REFERENCES

Cosslett, S. (1981) "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*; 49(5), 1289-1316.

Cox, D. and Hinckley, D (1974) *Theoretical Statistics*, London: Chapman and Hall.

Goldfeld, S.; Quandt, R. (1973) "A Markov Model for Switching Regressions," *Journal-of-Econometrics*; 1(1), 3-15.

Goldfeld, S.; Quandt, R. (1975) "Estimation in a Disequilibrium Model and the Value of Information," *Journal-of-Econometrics*; 3(4), 325-48.
Hausman, J.; Wise, D. (1977) " Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica*; 45(4),919-38.
Heckman, J. (1974) "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*; 42(4), 679-94.
Hsieh, D.; Manski, C.; McFadden, D. (1985) " Estimation of Response Probabilities from Augmented Retrospective Observations," *Journal-of-the-American-Statistical-Association*; 80(391),651-62.
Lancaster, T.; Imbens, G. (1990) "Choice-Based Sampling of Dynamic Populations," in Hartog, J.; Ridder, G.; Theeuwes, J., eds. *Panel data and labor market studies*. Contributions to Economic Analysis, vol. 192, Amsterdam: North-Holland, 21-43.
Lee, L. F.; Porter, R. (1984) "Switching Regression Models with Imperfect Sample Separation Information-With an Application on Cartel Stability," *Econometrica*; 52(2), 391-418.
Maddala, G. S.; Nelson, F. (1974) "Maximum Likelihood Methods for Models of Markets in Disequilibrium," *Econometrica*; 42(6), 1013-30.
Manski, C.; Lerman, S. (1977) "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*; 45(8), 1977-88.e
Manski, C. and D. McFadden (1981) "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in C. Manski and D. McFadden (eds) Structural *Analysis of Discrete Data*, Cambridge: MIT Press, 2-49.
McFadden, D. (1996) "On the Analysis of 'Intercept & Follow' Surveys," University of California Berkeley working paper.
Newey, W. and D. McFadden (1995) "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden, eds. *Handbook of Econometrics*, Vol. 4, Amsterdam: North Holland, 2113-2247.