I. Sample Selection Models

- Consider a population of women where only a sub-sample are engaged in market employment and report wages.
- Suppose we are interested in identifying the determinants of the wages of these working women to make statements regarding the determinants of wages for all women.
- The differences between the workers and non-workers determines whether an issue of selection bias arises.
- To illustrate this characterize each individual by her endowments of observable and unobservable characteristics.

- First assume that the working sub-sample is chosen randomly from the population.
- If the working sub-sample have similar endowments of characteristics as the non-working sample there is no reason to suspect selectivity bias will be induced by examining the working sample.
- That is, as the sample is randomly chosen the average characteristics, in terms of both observable and unobservables, of the working sample should be similar to the average characteristics of the population.

- Now consider where the decision to work is no longer random and as a consequence the working and non-working samples potentially have different characteristics.

- Sample selection bias arises when some component of the work decision is relevant to the wage determining process. That is, when some of the determinants of the work decision are also influencing the wage.

- However, if the relationship between the work decision and the wage is purely through the observables one can control for this by including the appropriate conditioning variables in the wage equation.

- Thus, sample selection bias will not arise purely on the basis of differences in observable characteristics.

- However, if we now assume the unobservable characteristics affecting the work decision are correlated with the unobservable characteristics affecting the wage we generate a relationship between the work decision and the process determining wages.
- Controlling for the observable characteristics when explaining wages is insufficient as some additional process is influencing the wage, namely, the process determining whether an individual works or not.
- If these unobservable characteristics are correlated with the observables then the failure to include an estimate of the unobservables will lead to incorrect inference regarding the impact of the observables on wages. Thus a bias will be induced due to the sample selection

- This discussion highlights that sample selectivity operates through unobservable elements, and their correlation with observed variables, although often one can be alerted to its possible presence through differences in observables across the two samples.

- However, this latter condition is no means necessary, or even indicative, of selection bias. Although this example is only illustrative it highlights the generality of the issues and their relevance to many economic examples.

- The possibility of sample selection bias arises whenever one examines a sub-sample and the unobservable factors determining inclusion in the sub-sample are correlated with the unobservables influencing the variable of primary interest.

II. The Model

- The conventional sample selection model has the form:

$$
\begin{aligned}
y_i^* &= x_i'\beta + \epsilon_i; i = 1..N && (1)\\
d_i^* &= z_i'\gamma + v_i; i = 1..N && (2)\\
d_i &= 1\ if\ d_i^* > 0;\ d_i = 0\ otherwise && (3)\\
y_i &= y_i^* * d_i; && (4)
\end{aligned}
$$

  where y_i^* is a latent endogenous variable with observed counterpart $y_i; d_i^*$ is a latent variable with associated indicator function d_i reflecting whether the primary dependent variable is observed and where the relationships, between d_i and d_i^*, and y_i and y_i^* respectively, are shown in (3) and (4). (1) is the equation of primary interest and (2) is the reduced form for the latent variable capturing sample selection; x_i and z_i are vectors of exogenous variables; β and γ are vectors of unknown parameters; and ϵ_i and v_i are zero mean error terms with $E[\epsilon_i|v_i] \neq 0$.

- We let N denote the entire sample size and use n to denote the sub-sample for which $d_i = 1$

.

- At this point we allow z_i to contain at least one variable which does not appear in x_i although this is sometimes seen to be a controversial assumption and, as such, we return to a discussion of it below.
- While this exclusion restriction is typically not necessary for parametric estimation it is generally crucial for semi-parametric procedures. For now we assume that x_i is contained in $z_i.$ The primary aim is to consistently estimate β.

-

- Well known that ordinary least squares (OLS) estimation of β over the sub-sample corresponding to $d_i = 1$ will generally lead to inconsistent estimates due to the correlation between x_i and ϵ_i operating through the relationship between ϵ_i and v_i.

- A number of remedies, however, exist. The first is maximum likelihood estimation and relies heavily on distributional assumptions regarding ϵ_i and v_i. A second approach is characterized by two-step procedures which approximate or eliminate the non-zero expectation of ϵ_i conditional on v_i

III. Maximum Likelihood Estimation

A. Parametric methods

- The first solution to sample selection bias was suggested by Heckman (1974) who proposed a maximum likelihood estimator. This requires distributional assumptions regarding the disturbances and Heckman made the following assumption:

- Assumption 1: ϵ_i and $v_i,\ i = 1..N,$ are independently and identically distributed $N(0, \sum)$ where $\sum = \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon v} \\ \sigma_{\epsilon v} & \sigma_v^2 \end{pmatrix}$ and (ϵ_i, v_i) are independent of z_i.

- Using Assumption 1 it is straightforward to estimate the parameters for the model in Section 3 by maximizing the following average log likelihood function:

$$L = \frac{1}{N}\sum_{i=1}^{N}\{d_i{*}\ln[\int_{-\infty}^{-(z_i'\gamma)} \phi_{\epsilon v}(y_i - x_i'\beta, v)dv] + (1-d_i){*}[\ln\int_{-(z_i'\gamma)}^{\infty}\int_{-\infty}^{\infty}\phi_{\epsilon v}(\epsilon, v)d\epsilon dv]\} \tag{5}$$

where $\phi_{\epsilon v}$ denotes the probability density function for the bivariate normal distribution.

- This is closely related to the Tobit estimator although it is less restrictive in that the parameters explaining the censoring are not constrained to equal those explaining the variation in the observed dependent variable.
- For this reason it is also known as Tobit type two (see, for example, Amemiya 1984).

- As estimation heavily relies on the normality assumption the estimates are inconsistent if normality fails. This is an unattractive feature although it is straightforward to test the normality assumption using tests such as those proposed by Gourieroux, Monfort, Renault and Trognon (1987) and Chesher and Irish (1987). It is also possible to employ the conditional moment framework of Newey (1985) and Tauchen (1985) as discussed in Pagan and Vella (1989).

- It is clear that estimation would be simplified if $\sigma_{\epsilon v} = 0$ as (5) would then reduce to the product of the two marginal likelihoods.
- That is, a product of the likelihood function explaining whether d_i was equal to 1 or $0,$ over the entire N sample, and the likelihood function explaining the variation in y_i for the n sub-sample satisfying $d_i = 1.$

- Alternatively, when $\sigma_{\epsilon v} \neq 0$ it is necessary to evaluate double integrals. Moreover, since there is no selection bias when $\sigma_{\epsilon v} = 0$ a test of this hypothesis is a test of whether the correlation coefficient $\rho_{\epsilon v}$ is equal to zero as this parameter captures the dependence between ϵ and v. Alternatively, one could estimate under the null hypothesis of no selection bias and test $\sigma_{\epsilon v} = 0$ via the lagrange multiplier or the conditional moment approaches.

- When the model is estimated by maximum likelihood the parameter estimates are fully efficient.
- This is an important characteristic as several alternative estimators do not require the same parametric assumptions for consistency. However, the relaxation of parametric assumptions is accompanied by an efficiency loss. Accordingly, the maximum likelihood estimates are a benchmark to examine the efficiency loss of these procedures under normality (see, for example, Nelson 1984).

- Furthermore, much of the development of this literature is devoted to the trade-off between efficiency and robustness. While maximum likelihood is the best when the model is correctly specified there exists a willingness to trade-off some efficiency for estimators which are more robust to distributional assumptions.
- The ease of implementation is also an important issue.

- One way to relax the normality, while remaining in the maximum likelihood framework, was suggested by Lee (1982, 1983).
- Lee proposes transforming the stochastic components of the model into random variables which can be characterized by the bivariate normal distribution.

- For example, suppose the errors ϵ and v are drawn respectively from the non-normal but known distributions $F(\epsilon)$ and $G(v)$.
- It is possible to transform ϵ and v into normal disturbances via the functions J_1 and J_2 , which involve the inverse standard normal distribution function, such that:

$$\epsilon^* = J_1(\epsilon) \equiv \Phi^{-1}[F(\epsilon)] \quad (6)$$

$$v^* = J_2(v) \equiv \Phi^{-1}[G(v)] \quad (7)$$

where the transformed errors ϵ^* and v^* now have standard normal distributions.

- The joint distribution of the transformed errors is now fully characterized through the bivariate normal distribution. Furthermore, it is possible to construct a likelihood function for the transformed errors as is done in equation (5) noting, however, that an additional set of parameters, characterizing $F(\epsilon)$ and $G(v)$, must be estimated.

B. Semi-nonparametric methods

- To avoid distributional assumptions it is possible to employ the general estimation strategy of Gallant and Nychka (1987) who approximate the underlying true joint density with:

$$b_{\epsilon v} = (\sum_{k=0}^{K} \sum_{j=0}^{J} \pi_{kj} \epsilon^k v^j) \phi_\epsilon \phi_v \tag{8}$$

where $b_{\epsilon v}$ denotes the true joint density; ϕ_ϵ and ϕ_v denote the normal densities for ϵ and v respectively; and π_{kj} denote unknown parameters.

- The basic idea is to multiply the product of these two marginal normal densities by some suitably chosen polynomial such that it is capable of approximating the true joint density.

- The estimate of $b_{\epsilon v}$ must represent a density and this imposes some restrictions on the chosen expansion and the values of π_{kj}. Gallant and Nychka show that the estimates of β and γ are consistent providing the number of approximating terms tends to infinity as the sample size increases.
- While Gallant and Nychka provide consistency results for their procedure they do not provide distributional theory.
- However, when K and J are treated as known inference can be conducted as though the model was estimated parametrically.

IV. Two-Step Estimation

- While the semi-nonparametric procedures can be computationally challenging the maximum likelihood procedures of Heckman and Lee are relatively straightforward. However, their use in empirical work is relatively uncommon.
- The more frequently employed methods for sample selection models are two-step estimators. In considering the two-step procedures it is useful to categorize them into three “generations”.

- The first fully exploits the parametric assumptions.
- The second relaxes the distributional assumptions in at least one stage of estimation.
- The third type are semi-parametric in that they relax the distributional assumptions.

A. Parametric Two-Step Estimation

- To examine the two-step version of the fully parameterized model we retain Assumption 1.
- The primary equation of interest over the n sub-sample corresponding to $d_i = 1$ can be written:

$$y_i = x_i'\beta + \epsilon_i; i = 1..n \tag{9}$$

  recalling OLS estimation leads to biased estimates of β since $E[\epsilon_i|z_i, d_i = 1] \neq 0$ (that is, the conditional mean of y is misspecified).

- The general strategy proposed by Heckman (1976,1979) is to overcome this misspecification through the inclusion of a correction term which accounts for $E[\epsilon_i|z_i, d_i = 1]$.
- To employ this approach take the conditional expectation of (9) to get:

$$E[y_i|z_i, d_i = 1] = x_i'\beta + E[\epsilon_i|z_i, d_i = 1]; i = 1..n.$$

- Using Assumption 1 and the formula for the conditional expectation of a truncated random variable we note that $E[\epsilon_i|z_i, d_i = 1] = \frac{\sigma_{\epsilon v}}{\sigma_v^2}\{\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}\}$ where $\phi(.)$ and $\Phi(.)$ denote the probability density and cumulative distribution functions of the standard normal distribution.
- The term in curly brackets is known as the inverse Mills ratio.

- To obtain an estimate of the inverse Mills ratio we require the unknown parameters γ and σ_v. By exploiting the latent structure of the underlying variable capturing the selectivity process, and the distributional assumptions in Assumption 1, we can estimate γ/σ_v by Probit.

- Thus the two-step procedure suggested by Heckman (1976,1979) is to first estimate γ over the entire N observations by maximum likelihood Probit and then construct an estimate of the inverse Mills ratio.

- One can then consistently estimate the parameters by OLS over the n observations reporting values for y_i by including an estimate of the inverse Mills ratio, denoted $\hat{\lambda}_i$, as an additional regressor in (9).

- More precisely, we estimate:

$$y_i = x_i'\beta + \mu\hat{\lambda}_i + \eta_i \tag{10}$$

  by OLS to obtain consistent estimates of β and $\mu,$ where η_i is the term we use throughout the paper to denote a generic zero mean error uncorrelated with the regressors and noting $\mu =\frac{\sigma_{\epsilon v}}{\sigma_v^2}.$

- This procedure is also known as a "control function" estimator (see, for example, Heckman and Robb 1985a,b).

- The t-test on the null hypothesis $\mu = 0$ is a test of $\sigma_{\epsilon v} = 0$ and represents a test of sample selectivity bias. Melino (1982) shows this represents the optimal test of selectivity bias, under the maintained distributional assumptions, as it is based on the same moment as the lagrange multiplier test.
- That is, both the lagrange multiplier test and the t-test for the coefficient on $\hat{\lambda}_i$ are based on the correlation between the errors in the primary equation and the errors from the selection equation noting that the inverse Mills ratio is the error from the Probit equation explaining selection.
- We return to this interpretation of the inverse Mills ratio below.

- The Heckman two-step estimator is straightforward to implement and the second step is only complicated by the standard errors having to be adjusted to account for the first step estimation (see, for example, Heckman 1976,1979, Greene 1981 and Maddala 1983).

- However, one concern is related to identification.

- While the inverse Mills ratio is nonlinear in the single index $(z_i'\gamma)$ the function mapping this index into the inverse Mills ratio is linear for certain ranges of the index.

- Accordingly the inclusion of additional variables in z_i in the first step can be important for identification of the second step estimates. However, there are frequently few candidates for simultaneous exclusion from x_i and inclusion in z_i. In fact, many theoretical models impose that no such variable exist.

- For example, empirical models based on the Roy (1951) model often employ the estimated covariances to infer the nature of the sorting. The underlying economic model often imposes the same variables to appear in both steps of estimation.

- Thus many applications constrain $x_i = z_i$ and identify β through the non-linearity in the inverse Mills ratio. However, as the inverse Mills ratio is often linear the degree of identification is often “weak” and this results in inflated second step standard errors and unreliable estimates of $\beta.$ This has proven to be a major concern (see, for example, Little 1985) and remains a serious point of contention.

- Given this is a relatively important issue for empirical work it has been the object of several monte-carlo investigations (see, for a recent example, Leung and Yu 1996).
- While most studies find that the two-step approach can be unreliable in the absence of exclusion restrictions Leung and Yu (1996) conclude that these results are due to the experimental designs.
- They find that the Heckman two-step estimator is effective providing at least one of the $x_i's$ display sufficient variation to induce tail behavior in the inverse Mills ratio.

- An examination of the inverse Mills ratio reveals that while it is linear over the body of permissible values the single index can take, it becomes non-linear at the extreme values of the index.
- Accordingly, if the $x_i's$ display a relatively large range of values, even in the absence of exclusion restrictions, then it is likely that the data will possess values of the single index which induce the non-linearity and assists in model identification. However, despite this finding these two-step procedures should be treated cautiously when the models are not identified through exclusion restrictions.

- It is worth reformulating the Heckman two-step estimator from a different and more restrictive perspective as this provides some insight into models we examine below.

- By imposing the restrictive assumption that the parameters are the same for each subsample it is possible to view the sample selection model as a model with a censored endogenous regressor. That is rewrite the model as:

$$\begin{aligned} y_i &= x_i'\beta + \theta d_i + \epsilon_i; i = 1..N && (11) \\ d_i^* &= z_i'\gamma + v_i; i = 1..N && (12) \\ d_i &= 1\ if\ d_i^* > 0;\ d_i = 0\ otherwise && (13) \end{aligned}$$

  where rather than sample selection we have an endogenous dummy variable.

- Estimating β and θ over the whole, or any chosen, sub-sample results in inconsistent estimates due to the correlation between d_i and ϵ_i operating through the non-zero covariance $\sigma_{\epsilon v}$.
- This is known as an "endogenous treatment model" and is closely related to the sample selection model (see, for example, Heckman 1978).

- It is well known, (see, for example, Hausman 1978 and Heckman 1978), that the inconsistency in (11) can be overcome by; i) projecting d_i onto z_i to obtain $\hat{d}_i$ and then replacing d_i with $\hat{d}_i$; or ii) obtaining the residuals from this projection, $\hat{v}_i$ and including both $\hat{v}_i$ and d_i in (11).

- A similar approach, which exploits the distributional assumptions and the dichotomous nature of the $d_{i,}$ involves estimating γ by Probit and then computing the corresponding Probit residual.

- That is, by using our distributional assumptions we can rewrite (11) as:

$$y_i = x_i'\beta + \theta d_i + \mu v_i + \eta_i.$$

- The Probit residual is known as a generalized residual (see Gourieroux et al. 1987) and has the form $d_i * \frac{\sigma_{\epsilon v}}{\sigma_v^2}\left[\frac{\phi(z_i'\hat{\gamma})}{\Phi(z_i'\hat{\gamma})}\right] + (1 - d_i) * \left[\frac{-\phi(z_i'\hat{\gamma})}{1-\Phi(z_i'\hat{\gamma})}\right]$. This can be identified as the inverse Mills ratio for the entire sample.

- This term possesses two important characteristics of a residual. First, it has mean zero over the whole sample.

- Second, it is uncorrelated with the variables which appear as explanatory variables in the first step Probit model.

- As the inclusion of the generalized residual accounts for the correlation between ϵ_i and $d_i,$ it is possible to estimate β over either sub-sample corresponding to $d = 0$ or $d = 1$ after including the generalized residual.
- This model is identified in the absence of exclusion restrictions due to the non-linearity of the residual. Also note that the generalized residual is uncorrelated with the $z_i's,$ over the whole sample, by construction.

- Thus the consequences of a high degree of collinearity between the generalized residual and the $z_i's$, which is a concern in the sample selection model, does not arise.

- An advantage of this interpretation is that it generalizes to alternative forms of censoring. Moreover, if we assume $E[\epsilon_i|v_i]$ is a linear function we can also relax the distributional assumptions for v_i. We return to this below.

- The parametric procedures are based on the exploiting the relationship between ϵ_i and v_i operating through the distributional assumptions.
- Bivariate normality dictates that the relationship between the disturbances is linear.
- Accordingly, one may test, or even correct, for departures from normality by including terms which capture systematic deviations from linearity. Lee (1984) suggests approximating the true density by the product of normal density and a series of Hermite polynomials.
- Although the test that Lee motivates is based on the lagrange multiplier framework he also presents a variable addition type test in which (10) is augmented with the additional terms.

- Pagan and Vella (1989) adopt a similar approach and follow Gallant and Nychka (1987) and approximate the bivariate density of ϵ_i and v_i as:

$$b_{\epsilon v} = (\sum_{k=0}^{K}\sum_{j=0}^{J}\pi_{kj}\epsilon^k v^j)\phi_{\epsilon v}$$

  recalling $\phi_{\epsilon v}$ is the bivariate normal density; the $\pi's$ denote unknown parameters; and $\pi_{00} = 1.$

- If we set $K = 0$, let $\phi_{\epsilon|v}$ denote the conditional normal density of ϵ given v, and $p = b_v/\phi_v$, then:

$$\begin{aligned} E[\epsilon_i|v_i] &= \sum_{j=0}^{J} p^{-1}\pi_{0j}(\epsilon\phi_{\epsilon|v}d\epsilon)v^j \\ &= \sum_{j=0}^{J} p^{-1}\pi_{0j}\rho v^{j+1}. \end{aligned}$$

- Thus under the null hypothesis of joint normality:

$$E[\epsilon_i|d_i = 1] = \pi_{00}E[v_i|d_i = 1] + \sum_{j=1}^{J}\pi_{0j}E[v_i^{j+1}|d_i = 1]$$

since $p = 1$ under the null hypothesis.

- A test of normality is to add on the higher order terms and test whether they are jointly zero. To compute these terms we can use the recursive formula provided in Bera, Jarque and Lee (1984). They are proportional to the inverse Mills ratio and take the form $E[v_i^{j+1}|z_i, d_i = 1] = (z_i'\gamma)^j\{\phi(z_i'\gamma)/[\Phi(z_i'\gamma)]\}$.
- Thus one computes these higher order terms and inserts them in (10) and tests whether they are jointly significant. Given the nature of the Hermitian expansion the additional terms employed by Pagan and Vella are similar to those suggested by Lee.

- It was quickly recognized that the heavy reliance of the two-step procedures on normality could be partially relaxed by replacing Assumption 1 with the relatively weaker Assumption 2:
- Assumption 2: The distribution of v_i is known and ϵ_i is a linear function of v_i.
- This presents no advantage over the Heckman two-step procedure if we assume that v_i is normal as Assumption 2 implies joint normality.
- It does however allow us to replace the normality of v_i with alternative distributional assumptions thereby allowing consistent first-step estimation by methods other than Probit

- One procedure is suggested by Olsen (1980) who assumes that v_i is uniformly distributed.
- One can now replace the inverse Mills ratio with a simple transformation of the least squares residuals derived from the linear probability model (that is, the residuals from regressing d_i on z_i).

- Olsen shows that when the disturbances in the selection equation are uniformly distributed this two-step estimator is consistent. More formally, Olsen shows that:

$$E[\epsilon_i|z_i, d_i = 1] = \rho\sigma_v(3)^{\frac{1}{2}}(z_i'\gamma - 1).$$

- This procedure generally produces results similar to Heckman two-step procedure.
- This follows from the high degree of collinearity between the respective corrections.

- The Olsen estimator requires the exclusion from the primary equation of at least one variable which appears in the reduced form as the model can no longer be identified through the non-linear mapping of the index $z_i'\gamma$ to the correction term.
- A test of selectivity bias is captured through a test of statistical significance of the coefficient of the correction term as this parameter captures the linear relationship between the two disturbances.

- A more general approach, to relax joint normality while remaining within the parametric framework, is proposed by Lee (1982,1983).
- A useful case is where the marginal distribution of ϵ_i is normal and the marginal of v_i is known but non-normal.
- Thus the distribution of ϵ and the transformed disturbance v^* is bivariate normal and their dependence is captured by their correlation coefficient.
- More importantly, the relationship between the disturbances is linear.

- To implement the two-step version of the Lee maximum likelihood estimator we note that $d_i = 1$ when $v_i < z_i'\gamma$.
- This implies, from (7) that $J_2(v_i) < J_2(z_i'\gamma)$.
- It follows that $\Pr[d_i = 1] = \Phi[J_2(z_i'\gamma)] = G(z_i'\gamma)$.
- Thus we can now write the conditional expectation of (9) as:

$$E[y_i|z_i, d_i = 1] = x_i'\beta + \rho\sigma_{\epsilon v}\phi[J_2(z_i'\gamma)]/G[z_i'\gamma]. \tag{14}$$

- Thus first we estimate the γ from the discrete choice model by maximum likelihood where we employ $G(v_i)$ as the distribution function for v_i.
- We then substitute the estimate of γ into (14) and estimate by least squares.
- Lee (1982) generalizes this approach such that J_2 is a specified strictly increasing transformation such that $v_i < z_i'\gamma \iff J_2(v_i) < J_2(z_i'\gamma)$.

- Let $\mu_J = E[J_2(v)]$ denote the expected value of v and let σ_J^2 denote the variance of $[J_2(v)]$. Furthermore assume that ϵ_i can be written as:

$$\epsilon_i = \tau[J_2(v) - \mu_J] + \eta_i \tag{15}$$

  where η_i and $J_2(v_i)$ are independent and where $\tau = 0$ if the disturbances are uncorrelated.

- If we write the conditional mean of the truncated disturbance as:

$$\mu(J_2(z_i'\gamma)) = E[J_2(v_i)|J_2(v_i) < J_2(z_i'\gamma)]$$

then:

$$E[y_i|z_i, d_i= 1] = x_i'\beta + \rho\sigma_\epsilon/\sigma_J \left(\frac{\mu(J_2(z_i'\gamma))}{G(z_i'\gamma)} - \mu_J\right). \tag{16}$$

- While this methodology provides some flexibility it crucially depends on the assumption in (15).
- This approach, in the conventional sample selection model, is typically associated with the use of Logit.
- It is particularly attractive when there are multiple unordered outcomes as maximum likelihood estimation of the first step can be computationally difficult.

V. Semi-Parametric Two-Step Estimation

- An early criticism of the parametric sample selection estimators was their reliance on distributional assumptions
- While this can be relaxed, through the use of different distributional assumptions, it is appealing to consider alternatives which have a limited reliance on parametric assumptions.

- To consider the available procedures replace Assumption 2 with a weaker statement about the disturbances.

Assumption 3: $E[\epsilon_i|z_i, d_i = 1] = g(z_i'\gamma)$ where g is an unknown function.

- Assumption 3 is known as an index restriction. While the parametric two-step approaches implicitly define the function $g(.)$ through the distributional assumptions, or assume it explicitly, the semi-parametric procedures seek to avoid the imposition of such information.

- Estimation under Assumption 3 rather than Assumptions 1 and 2 raises two difficulties.
- First, it is no longer possible to invoke distributional assumptions regarding v_i to estimate γ.
- Second, we cannot use distributional relationships to estimate $E[\epsilon_i|z_i, d_i = 1]$. The first problem is overcome through non-parametric or semi-parametric estimation of the binary choice model.

- For example, it is possible to estimate γ by the procedures of Cosslett (1983), Gallant and Nychka (1987), Powell, Stock and Stoker (1987), Klein and Spady (1993) and Ichimura (1993), without imposing distributional assumptions on v_i.
- With these estimates it is straightforward to compute an estimate of the single index $z_i'\gamma$ and the second difficulty can then be overcome in a number of ways.

VI. Semi-Parametric Estimation

- First, why should we be interested in semi-parametric methods?
- Although non and semi-parametric procedures are very popular with econometrician theorists the ideas have caught on far less with applied people.
- This is partially due to a number of factors.

- i) Despite the fact they are simple they are typically not well understood.
- ii) Frequently implementation is not seen to be straightforward as various choices have to be made and standard software is typically not applicable.
- iii) There is an incorrect sense that the methods are not useful in that they do not work well or, in some cases, give the same kind of results as parametric methods.

- We will now see that many of the available methods are directly applicable to many economic problems and show there is a very large number of models which could be estimated by simple extensions on the available methods.s.

- The outline is the following. We will start with simple density estimation. Then we will examine how to estimate conditional moments via the use of density estimates. Once we have the ability to estimate conditional moments we will examine several models in which the estimated conditional moments appear as important explanatory variables.

VII. Density Estimation

A. Empirical Distribution

- Start with the simple problem of estimating the cumulative distribution function (CDF) $F_Z(z) = \Pr(Z \le z)$ of a random variable Z.
- Let $1(A)$denote the indicator of some event A which is equal to 1 if A occurs and zero otherwise.
- Since $F_Z(z)$ is a probability it can be estimated from an iid sample $Z_1 ... Z_n$ as

$$\widehat{F}_Z(z) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i \le z)$$

- This is the empirical distribution.
- This puts weight of 1/n on each observation.

- While it has good properties it has the feature of being discrete.
- Thus it is not too good for density estimation.
- The lack of smoothness is also a problem when we look at estimators which are based on the use of the non-parametric procedures.

B. Density Estimation

- We can construct a density estimator by adding a bit of noise that smooths over the discreteness.
- Let
$$\widetilde{Z} = \overline{Z}_n + hU$$
where the cdf of $\overline{Z}_n$ is the empirical CDF, h is a positive scalar and U is continuously distributed with pdf $K(u)$.
- If we let
$$F_U(u) = \int_{-\infty}^{u} K(t)dt$$
be the cdf of U.

- Then one can show that the CDF of $\widetilde{Z}$ is

$$F_{\widetilde{Z}}(z) = \sum_{i=1}^{n} F_U\left(\frac{z - Z_i}{h}\right)/n$$

  and the corresponding estimator of the pdf is

- $$\widehat{f}_h(z) = \frac{dF_{\widetilde{Z}}(z)}{dz} = \sum_{i=1}^{n} K_h(z - Z_i);\ K_h(u) = h^{-1}K\left(\frac{u}{h}\right)$$

- This is the kernel density estimator.

- The function $K(u)$ is the kernel and h is the bandwidth
- Note that h increases the density becomes smoother.
- At the same time as h increases we have more noise and this introduces more bias.
- Consistency generally requires that $h \to 0$ as the sample size increases. However, at a rate which is not too fast (i.e. $nh \to \infty$)

- The choice of kernel is far less important than the choice of the bandwidth.
- However, the kernel should have certain properties
- Popular choices are the normal kernel or Epanechnikov

$$K(u) = 1(|u| \leq 1)(1 - u^2)(.75)$$

- Alot of the ideas that we will look employ the use of density estimation so it is useful to see how easily it can be programmed.

- The following is gauss program for density estimation using a normal kernel.

```
new;
n=100;
x=rndn(n,1);
fxn=zeros(n,1);
j=1;
sx=stdc(x);
h=1.06*sx*n^(-.2);
do until j>n;
d=(x[j]-x)/h;
fx1=pdfn(d);
fxn[j]=meanc(fx1/h);
j=j+1;
endo;
library pgraph;
xx=sortc(x~fxn,1);
xy(xx[.,1],xx[.,2]);
```

- An important feature of the kernel estimator is that is biased. The logic behind this naturally is that the density at a certain point is a weighted average of ”all” the data when really we only want the data at that point. (This is the logic why $h \to 0$ as the sample size increases)
- In general both the bias and the variance depend on the kernel used and the shape of the density.
- As bias and variance reduction methods are a very important part of implementing kernel estimation (as we will discuss below) it is useful to examine the expression for the bias for the kernel estimator.

- Bias Result: If we assume that $f_o(z)$ is twice continuously differentiable with bounded second derivative, $\int K(u)du = 1$, $\int K(u)udu = 0$, $\int u^2K(u)du < \infty$ then

$$E\left[\widehat{f}_h(z)\right] - f_o(z) = h^2 f_o'' \int K(u)u^2du/2 + o(h^2)$$

- Variance Result: If $f_o(z)$ is continuous and bounded, $\int u^2K(u)du < \infty$, $h \to 0$, and $nh \to \infty$ then

$$Var\left[\widehat{f}_h(z)\right] = f_o(z) \int u^2K(u)du/(nh) + o(1/(nh))$$

- Note that if we choose a very small h the bias in the density estimate is small but the variance is large.
- A very small h means we are using a small number of points and this will may lead to an estimate displaying many wiggles.
- On the other hand a large h oversmooths the density.

- In practice h should be set to achieve the best trade off between bias and variance.
- One possibility is to focus on some approximation to the integrated mean squared error

$$MISE = \int \left[\left(Bias\ \widehat{f}_h(z)^2 \right) + V(\widehat{f}_h(z) \right] dz$$

- Since Bias^2 is $O(h^4)$ and variance is $O(nh)^{-1}$ the value for which they are of the same order of magnitude is $h \propto n^{-1/5.}$
- The AMISE is now $O(n^{-4/5})$
- We can see that by $h \to 0$ the MSE vanishes slower than $\frac{1}{n}$ and thus the non-parametric density estimator converges at a rate slower than $\frac{1}{\sqrt{n}}$.
- This slower convergence occurs because of the avoidance of bias by $h \to 0$ means that the fraction of the observations that the estimator uses shrinks to zero.

C. Bias Reduction

- Given that the density estimator is biased it is common to use bias reduction methods.
- This is a potentially important issue given, particularly in multiple index models, since the type of bias reducing method employed will influence the final estimates when the density estimator is used as a means of computing conditional moments.
- Two types of bias reducing methods which can be employed are i) higher order kernels and ii) local smoothing.
- Higher order kernels are functions which integrate to one, but integrals of some positive powers of u are negative over some range. These have the feature of reducing bias.
- Although they work in theory there is some concern that they may not work so well in practice. This is particularly true for the multiple index models.

- Local smoothing is where we use a different bandwidth depending on whereabouts in the density of the random variable we are. For example, when we are in the dense part of the density the bandwidth shrinks while once we are in a section of the density where the observations are scarce we use relatively more observations which means a bigger bandwidth.
- The optimal bandwidth, in the MSE sense, in this case can be shown to be

$$h_i = cf(z_i)^{-.5}$$

- Thus this requires the use of a pilot density in that we estimate the density once to obtain $f(z_i)$ and then we use this as the bandwidth.

- The program below employs local smoothing to estimate the density of x.

- new;

```
n=100;
x=rndn(n,1);
fxn=zeros(n,1);
fxln=zeros(n,1);
j=1;
sx=stdc(x);
h=1.06*sx*n^(-.2);
do until j>n;
d=(x[j]-x)/h;
fx1=pdfn(d);
fxn[j]=meanc(fx1/h);
j=j+1;
endo;
library pgraph;
xx=sortc(x~fxn,1);
j=1;
h=.25*(fxn.^(-.5));
do until j>n;
d=(x[j]-x)./h;
fx1=pdfn(d);
fxln[j]=meanc(fx1./h);
j=j+1;
endo;
xx=sortc(x~fxn~fxln,1);
xy(xx[.,1],xx[.,2 3 ]);
```

Additional Issues

- Two additional issues are worth considering.
- Extension to multivariate density estimation. This can be of particular value although once the dimension of the random variables is higher than 2 there can be some difficulties in implementing the method of estimation.
- Nevertheless, the extension to two allows the estimation of a large class of estimators.
- The simplest way to proceed to two is to characterize the joint density as the product of the two univariate kernels. This works better in practice if the two random variables are orthogonalized.

- The second important issue is the choice of the bandwidth.
- We will frequently see that the bandwidth has to be in some range and frequently involves the use of some unknown constant.
- Accordingly there is always a degree of arbitrariness about the choice of $h.$
- Frequently there are some rules of thumb which can be employed.
- A common approach is to employ Cross Validation methods which is based on minimizing some estimate of the integrated MSE.

VIII. Estimation of Conditional Mean

- Frequently in economics we are interested in the use of conditional expectations.
- Usually we are interested in the value of some endogenous value on the basis of some conditioning values.
- In the case when the dimension of the conditioning values is large we can use non-parametric regression.
- This kind of approach typically suffers from the curse of dimensionality.

- In economics when there is a large conditioning set it is useful to reduce the dimensionality by making a single index restriction. That is,

$$\begin{aligned} & E[y_i|x_{1i}, ...x_{ki}] \\ = \; & E[y_i|x_i'\beta] \end{aligned}$$

- In most of the models that we will examine, the single index assumption will be imposed.
- Thus we can look at deriving the expectation of $E[y_i|x_i]$ where there is a single x as this will be useful for examining $E[y_i|x_i'\beta]$
- Suppose that the x_i are iid random variables and we are interested in estimating the conditional mean of y given $x.$
- Since $m(x_i)$ is the mean of the conditional density $f(y_i|x_i) = f(Y|X = x_i)$ we can use the density estimation procedure from above.

- Now, by definition

$$m = \int_{-\infty}^{\infty} y f(y,x)dy / \int_{-\infty}^{\infty} f(y,x)dy$$

- We can evaluate these property at the density estimates for $f(.,.)$.
- Assume X is a scalar and consider a bivariate kernel $k(u_1, u_2)$ where $\int t(k(t, u_2)dt = 0.$
- Let the data be $(Y_1, X_1), ...(Y_n, X_n)$ and $K(t, u_2) = \int k(t, u_2)dt.$

- By change of variables $t = \frac{(y_i - Y)}{h}$ we get
- $$\begin{aligned} \int y\widehat{f}_h(y,x)dy &= n^{-1}h^{-2}\sum yk\left(\frac{y-Y_i}{h}, \frac{x-X_i}{h}\right)dy \\ &= n^{-1}h^{-1}\sum\int(Y_i+ht)k(t,\frac{x-X_i}{h})dt \\ &= n^{-1}h^{-1}\sum Y_i\int k(t,\frac{x-X_i}{h})dt \\ &= n^{-1}h^{-1}\sum Y_iK(\frac{x-X_i}{h}) \end{aligned}$$
- $$\begin{aligned} \int \widehat{f}_h(y,x)dy &= n^{-1}h^{-2}\sum k\left(\frac{y-Y_i}{h}, \frac{x-X_i}{h}\right)dy \\ &= n^{-1}h^{-1}\sum K(\frac{x-X_i}{h}) \end{aligned}$$

- This can be written as

$$\widehat{m} = \frac{\int y\widehat{f}_h(y,x)dy}{\int \widehat{f}_h(y,x)dy} = \frac{n^{-1}h^{-1}\sum Y_i K(\frac{x-X_i}{h})}{n^{-1}h^{-1}\sum K(\frac{x-X_i}{h})}$$

- Thus this formula shows that estimating the conditional mean is straightforward and only requires the use of the marginal density.
- The interpretation is also quite clear. That is, it takes a weighted average of the Y's on the basis of how far the values of x_i are from the candidate value of X.

- This program shows how to adjust the density estimation programs to estimate conditional means

```
/* Estimating conditional mean using constant kernel and local smoothing*/
new;
n=100;
x=rndn(n,1);
y=-.1*x+.5*x^2;
fxn=zeros(n,1);
fxln=zeros(n,1);
j=1;
sx=stdc(x);
h=1.06*sx*n^(-.2);
do until j>n;
d=(x[j]-x)/h;
fx1=pdfn(d);
num=fx1.*y;
fxn[j]=sumc(num)./sumc(fx1);
j=j+1;
endo;
j=1;
h=.25*(fxn.^(-.5));
do until j>n;
d=(x[j]-x)./h;
fx1=pdfn(d);
num=fx1.*y;
fxln[j]=sumc(num)./sumc(fx1);
j=j+1;
endo;
xx=sortc(x~fxn~fxln~y,1);
xy(xx[.,1],xx[.,2 3 4]);
```

```
/*ESTIMATING CONDITIONAL MEAN USING ESTIMATED BIVARIATE DENSITY*/
new;
n=100;
x=rndn(n,2);
y=5*x[.,1]+.5*x[.,2].^2+1*rndn(n,1);
fxn=zeros(n,1);
fxmm=zeros(n,1);
j=1;
h=1.06*n^(-.4);
do until j>n;
d=(x[j,.]-x)/h;
fx1=pdfn(d);
fxn[j]=meanc(prodc(fx1’/h));
fxm=sumc(prodc(fx1’/h).*y);
fxmm[j]=fxm./sumc(prodc(fx1’/h));
j=j+1;
endo;
library pgraph;
xx=sortc(y~fxmm,1);
xy(xx[.,1],xx[.,2]);
```

Series Regression

- The computation of a conditional mean has the intrepretation of a kernel regression.
- This has obvious problems when the dimension of the conditioning set is big.
- Another approach to non-parametric regression is to emply least squares using flexible functional forms.
- In this case we would approximate $g_0(x)$ by a linear combination of some approximating functions $p_{jK}(x)$.

- The estimator of $g_0(x)$ is the predicted value from regressing Y_i on $p^K(X_i)$ for $p^K(x) = (p_{1K}(x)....p_{KK}(x))'$.
- Consistency is obtained by letting K grow with the sample size.
- If we let $Y = (Y_1, ...Y_n)'$ and $P = [P^K(X_1), ...P^K(X_n)]'$ then the estimator is

$$\widehat{g}(x) = p^K(x)'\widehat{\beta}, \text{ where } \widehat{\beta} = (P'P)^{-1}P'Y$$

- Popular choices for P are polynomials and splines.

- Polynomials have the disadvantage that they are based on global approximations and thus they are sometimes not ideal for local fitting.

- Splines are more suited for local fitting.

- The series approximations are well suited for multivariate cases by including cross products etc.

- There is a large literature on the convergence rates for $\widehat{g}(x)$.

- An advantages of series is that they are well suited to several problems in econometrics and they are easily employed for models imposing additivity.

- Note that the K serves the same purpose as h.

- Robinson's Estimator
- In many types of models there may be a non-parametric element which is of potential interest or is form of nuisance.
- In these models it would be useful to estimate the parameters of interest while eliminating the non-parametric component.
- Suppose the model has the form
- $$Y_i = X_i'\beta + g(Z_i) + u_i$$

  where $g(.)$ is an unknown function.

- One way to estimate the parameters and the $g(.)$ function is to solve the following problem

$$\arg\min_{\beta,g(.)} \sum \left[Y_i - X_i'\beta - g(Z_i)\right]^2$$

- The minimization can occur in 2 steps. First, solving for g over fixed β, substituting that minimum into the objective function, and then minimizing over $\beta.$
- The minimizer over $g(.)$ for fixed β is

$$E\left[Y_i - X_i'\beta|Z\right] = E\left[Y_i|Z\right] - E\left[X|Z\right]\beta$$

- Substituting this back gives

$$\beta_0 = \arg\min_{\beta} \sum E\left[\left\{[Y_i - E[Y_i|Z]] - [X - E[X|Z]]' \beta\right\}\right]$$

- This estimator can be easily implemented by replacing the terms $E[Y_i|Z]$ and $E[X|Z]$ with their non-parametric expectations. This can be done either using the kernel or series methods discussed above.

- Note that one can also recover an estimate of the $g(.)$ function by performing the following non-parametric regression

$$Y - X_i\beta_0 = g(Z_i) + u_i$$

- Note that the above has been extended to models where there are parameters inside the $g(.)$ function. This widely extends the number of applications that are available.

- Klein and Spady Estimator
- This paper examines the discrete choice model
- Model has the form we have discussed above. That is,
-

$$\begin{aligned} y_i^* &= x_i'\beta + u \qquad (17) \\ y_i &= 1 \text{ if } y_i^* > 0 \qquad (18) \\ &= 0 \text{ otherwise} \end{aligned}$$

- A key assumption is that the model is a single index model. That is the probability that y is equal to 1 is a function only of the index $x_i'\beta.$

- Recall that the log likelihood function for this model is

$$\log\ L = \sum_{i=1}^{n} y_i \log F(x_i'\beta) + \sum_{i=1}^{n}(1 - y_i) \log F[1 - (x_i'\beta)] \tag{19}$$

where $F(.)$ is some generic likelihood function. Note that we can also write this as

$$\log\ L = \sum_{i=1}^{n} y_i \log \{\Pr(y_i = 1|x_i'\beta)\} + \sum_{i=1}^{n}(1-y_i) \log \{1 - (\Pr(y_i = 1|x_i'\beta))\} \tag{20}$$

- KS make the observation that via the use of Bayes Law one can write

$$\Pr(y_i = 1|x_i'\beta) = \Pr(y_i = 1) * \frac{g_{x_i'\beta|y=1}}{g_{x_i'\beta}}$$

  where $g_{x_i'\beta|y=1}$ is the conditional density of the index given y is equal to 1 and $g_{x_i'\beta}$ is the unconditional density for the index.

- Writing the probability has the advantage that rather than conditoning on the index, we evaluate the index conditioned on something else (namely the $y's$).
- We have seen above that to estimate these densities and conditional densities is very straightforward.
- Thus we evaluate the above density and maximize the associated likelihood function.
- What is identified?
- Local Smoothing.
- Trimming of the criterion function.

Semiparametric Least Squares

- Another estimator we will consider the SLS estimator of single index model. This is the estimator proposed by Ichimura (1993).
- The model has the following form
- $$y_i = \tau\left(v\left(X_i, \beta\right)\right) + \varepsilon_i$$

  where $\tau(.)$ is an unknown function and v is a form of aggregation for the exogenous variables X and the unknown parameter vector $\beta.$ By making the single index assumption we can rewrite the above as
- $$y_i = \tau\left(X_i\beta\right) + \varepsilon_i$$

- The idea of the estimator is to use non-linear least squares to define the $\beta' s.$ That is, we want to estimate

$$\underset{\beta\tau(.)}{\arg\min} \sum \left[y_i - \tau\left(X_i\beta\right)\right]^2$$

- If we knew the form of the $\tau\,(.)$ it would be straightforward to estimate the $\beta.$ However, we can replace this with a non-parametric estimate of $E[y_i|X_i\beta].$ Thus we proceed by minimizing over $\tau(.)$ for a fixed $\beta.$

- Thus the procedure works in the following way. For a given β we have the index for each i. Using this index we can now estimate, non-parametrically, $E[y_i|X_i\beta]$, by kernel methods as discussed before. We then continue to search for values of β until we minimize the sum of the squared errors from above.

- Some identification restrictions need to be imposed. First, there can be no constant in X and one of the $\beta's$ must be set to 1.
- Even with any normalization it is still possible to evaluate the partial derivative $\partial y / \partial X$.
- This procedure if very easy to implement and its small sample properties are good. Despite this however, it is not frequently used in empirical work.

- Now that we see how to estimate the index semiparametrically we can return to the sample selection model
- Using the index restriction write the conditional expectation of the primary equation as:

  $$E[y_i|z_i, d_i = 1] = x_i'\beta + g(z_i'\gamma);\ i = 1..n.$$

  noting that it is not possible to distinguish an intercept term in x_i from an intercept in $g(.)$.
- Accordingly, the intercept term is not identified in the following procedures.
- We discuss below, however, some ways to infer the value of the intercept.
- Given consistent estimates of the single index the issue is how the $g(.)$ function is approximated.

- The first suggestion to estimate the model semi-parametrically is found in Heckman and Robb (1985a).
- They suggest a two-step estimator in which the first step is the non-parametric estimation of $\Pr[d_i = 1|z_i]$, which is also known as the propensity score (see, for example, Rosenbaum and Rubin 1983).
- The second step is to approximate the $g(z_i'\gamma)$, which is equal to $E[\epsilon_i|z_i, d_i = 1]$, through a Fourier expansion in terms of $\Pr[d_i = 1|z_i]$.

- Cosslett (1991) proposes a two-step procedure in which he first estimates $\hat{\gamma}$ via the non-parametric maximum likelihood estimator outlined in Cosslett (1983).

- The first step approximates the marginal distribution function of the selection error, $\hat{F}(.)$, as a step function constant on a finite number J of intervals $\{\hat{I}_j \equiv [\hat{c}_{j-1}, \hat{c}_j), j = 1..J$ and $c_0 = -\infty, c_J = \infty\}$.

- In the second step Cosslett estimates the primary equation while approximating the selection correction, $g(.)$ by J indicator variables $\{1(z_i'\hat{\gamma} \in \hat{I}_j)\}$.

- Consistency requires that J increase with the sample size.

- Newey (1988) suggests estimating the single index by some semi-parametric procedure.
- He then approximates $g(z_i'\gamma)$ by $\hat{g}(z_i'\hat{\gamma}) = \sum_{k=1}^{K} \alpha^k (z_i'\hat{\gamma})^{k-1}$ where $\hat{\gamma}$ is some first step estimate and K, denoting the number of terms in the approximating series, is allowed to grow with the sample size.
- The second step is then estimated by OLS while setting K equal to some fixed number. An advantage of the Newey approach is that the estimates are $\sqrt{n}$ consistent and it is straightforward to compute the second step covariance matrix.

- The above estimator employs the orthogonality conditions $E[\epsilon_i - g(z_i'\gamma)|d_i = 1, z_i] = 0$ to define the estimator of β.
- Newey argues efficiency gains can be obtained if the additional orthogonality conditions implied by the independence of $\{\epsilon_i - g(z_i'\gamma)\}$ and z_i are exploited.
- Newey notes that $\{\epsilon_i - g(z_i'\gamma)\}$ is uncorrelated with any function of z_i.

- To employ the additional orthogonality conditions implied by this independence define $\varsigma_j(\epsilon_i) \;\; (j = 1..J)$ as some function of ϵ_i, and $\xi_j(z_i'\gamma) = E(\varsigma(\epsilon_i)|z_i, z_i'\gamma)$.
- Newey then defines a generalized method of moments estimator based on the orthogonality conditions $E[k(\varsigma_j(\epsilon_i) - \xi_j(z_i'\gamma))]$ where k is some function of z_i and $z_i'\gamma$.

- .An alternative approach to the elimination of selection bias is based on an estimation strategy suggested by Robinson (1988) in which the endogeneity is purged from the model through a differencing process.
- Powell (1987) exploits the index restriction in estimation by identifying observations by their value of this single index.

- The underlying intuition is that if two observations i and j have similar values for the single index generating the selection bias, then it is likely that subtracting the j^{th} observation from the i^{th} observation will eliminate the selection bias.
- Powell (1987) suggests an instrumental variable estimator based on pairwise comparisons of all observations in the sample where the contribution of each comparison is weighted by the difference in the values of the single index.

- The estimator of β,denoted β_p, has the form:

$$\begin{aligned}\beta_p &= \left\{[^n_2]^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n} m_{ijn}(w_{ij}x_{ij})'\right\}^{-1} \\ &\quad \left\{[^n_2]^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n} m_{ijn}(w_{ij}y_{ij})'\right\}\end{aligned}$$

  where w_{ij}, x_{ij} y_{ij} denote $(w_i - w_j)$, $(x_i - x_j)$ and $(y_i - y_j)$ respectively and m_{ijn} captures a weight depending on the distance between the values of the single indices for the i^{th} and j^{th} observations; and the $w'_{ij}s$ denote some chosen instruments.

- The weight is constructed such that observations that are nearby, in terms of the single index, have a greater contribution than those far apart.
- As the weights are unobserved the first step is semi-parametric estimation of the single indices $z_i'\gamma$ and $z_j'\gamma.$

- A similar approach, based on differencing out the selectivity bias, first uses the index restriction to rewrite the primary equation as:

$$y_i = x_i'\beta + g(z_i'\gamma) + \eta_i; i = 1..n. \tag{21}$$

- With an estimate of $z_i'\gamma$ we condition (21) on $z_i'\gamma$ to get:

$$E[y_i|z_i'\gamma] = E[x_i|z_i'\gamma]'\beta + g(z_i'\gamma); i = 1..n.$$

- We subtract this conditional expectation from (21) to get:

$$y_i - E[y_i|z_i'\gamma] = \{x_i - E[x_i|z_i'\gamma]\}'\beta + \eta_i; i = 1..n \tag{22}$$

which can be estimated by OLS since the component reflecting the selection bias has been eliminated.

- A closely related estimator to that proposed by Powell (1987) is Ahn and Powell (1993).
- The innovation in the Ahn and Powell procedure is their use of non-parametric kernel methods to compute the propensity scores $Pr[d_i = 1|z_i]$ and $Pr[d_j = 1|z_j]$.
- They then use these probabilities in place of the estimated single indices $z_i'\gamma$ and $z_i'\gamma$ in the computation of the weights $m_{ijn}.$ This is an important variation on Powell (1987) as it relaxes the single index assumption.
- However, it is accompanied by a substantial increase in computational requirements as it is necessary to estimate the first step non-parametrically. While the second step is $\sqrt{n}$ consistent one would expect some efficiency loss due to the manner in which the first step is estimated.

- The final semi-parametric procedure for the conventional sample selection model that we consider was proposed by Ichimura and Lee (1991) and is based on their estimator for models with multiple indices.

- Although this procedure is a single equation estimator it is well motivated in the sequential equation framework. Recall that the model for the sub-sample has the following form:

  $$E[y_i|d_i = 1, z_i] = x_i'\beta + g(z_i'\gamma); i = 1..n. \tag{23}$$

  which implies:

  $$E[(y_i - x_i'\beta)|z_i'\gamma] = g(z_i'\gamma). \tag{24}$$

- Equations (23) and (24) characterize the relationship between β, γ and $g(.)$.
- The Ichimura and Lee procedure is based on the following iterative non-linear least squares approach.
- With estimates of β and γ we employ (24) to estimate $g(.)$ non-parametrically. T
- hen using (23) and our estimate of $g(.)$ we can estimate β and γ.

A. Conditional Expectations and Bounds

- While we have focussed on the estimation of the β, one may be interested in conditional expectation of $E[y_i|, z_i, d_i]$.

- This originally arose in Lee (1978) and Willis and Rosen (1979) where the model comprised two mutually exclusive and exhaustive sectors. Wages were observed for each individual for the sector in which the individual was located.

- One objective in those studies was to compute the wage for the sector in which the individual was not observed. We now examine the case where y_i is only observed for a sub-sample.

- The generalization to the case of y_i observed for everyone, although individuals are in different sectors, is straightforward.

- Suppose the two "sectors" refer to market and non-market employment and the variable y reflects the offered market wage.

- Furthermore, assume the errors are bivariate normally distributed and we estimate the model by the Heckman two-step procedure. Denote the parameter estimates for the sub-sample of those engaged in market employment, as β_M.

- Consider the expectation $CE_{1i} = x_i'\beta_M$.
- This represents the expected market wage for an individual randomly selected from the sample.
- That is, the conditioning set does not contain any information regarding the sector in which the individual is actually located. This is the approach adopted in Lee (1978) and Willis and Rosen (1979).
- However, the expectation CE_{1i} can be "improved" via the inclusion of information relaying the chosen sector. For example, $CE_{2i} = E[y_i|z_i, d_i = 1] = x_i'\beta_M + \mu\lambda_i^M$ represents the expected wages for those already located in the respective sectors noting that the λ^M denotes the inverse Mills ratios for those in the market sector and the μ is the estimated parameter capturing the covariance between the errors across equations.

- This latter expectation varies from CE_{1i} in that it includes the respective returns to the unobservables associated with market sector (see, for example, Vella 1988).

- Accordingly, one may consider the following counterfactual wages as conditional expectations of interest. Namely, $CE_{3i} = E[y_i|z_i, d_i = 0] = x_i'\beta_M + \mu\lambda_i^N$ which represents the expected wages for those in the non-market sector if they obtained market employment noting that λ^N is the inverse Mills ratio for those in the non-market sector.

- Once again the term $\mu\lambda_i^N$ captures the market return to the unobservables.

- Lee (1995) extends and generalizes this approach by providing a general strategy for estimating the conditional expectations of the outcomes not chosen.
- Furthermore, Lee provides the formulae for the conditional expectations of outcomes models with polychotomous outcomes in models with sample selection bias.

- Manski (1989) focuses on the estimation of bounds for the conditional expectation (that is, $E[y_i|z_i]$ over the support of z_i) when y_i is only observed for either $d_i = 1$ or $d_i = 0$ but not both.
- Manski considers the case where z_i and d_i are observed over the whole sample and $E[y_i|z_i, d_i = 1]$ is observed. F
- irst, it is straightforward to see that:

$$E[y_i|z_i] = E[y_i|z_i, d_i = 1] \Pr[d_i = 1|z_i] + E[y_i|z_i, d_i = 0] \Pr[d_i = 0|z_i]. \tag{25}$$

- Manski assumes that the support of y_i conditional on $d_i = 0$ and z_i is known and lies in the interval $[K_{Lz}, K_{Uz}]$ which implies $K_{Lz} \leq E(y_i|z_i, d_i = 0) \leq K_{Uz}$. This, in turn with (25), implies:

$$\begin{aligned} E(y_i|z_i, d_i &= 1)\Pr(d_i = 1|z_i) + K_{Lz}\Pr(d_i = 0|z_i) \leq E(y_i|z_i) \quad (26) \\ &\leq E(y_i|z_i, d_i = 1)\Pr(d_i = 1|z_i) + K_{Uz}\Pr(d_i = 0|z_i). \end{aligned}$$

- The components of the bound are readily available in most contexts and Manski discusses the methodology for implementing the bound.

- The first important feature of (26) is that rather than focussing on a point estimate it provides a bound.
- A second feature is that it can be implemented non-parametrically as the components of (26) can be estimated from sample data without the imposition of parametric assumptions.
- A possible criticism is that the estimated bounds may be too wide to be informative. While this represents information in itself Manski (1994) shows how the bounds can be tightened through the use of additional information such as functional form and exclusion restrictions

.

IX. Sample Selection Models with Alternative Censoring Rules

- The model has the form:

$$\begin{aligned} y_i^* &= x_i'\beta + \epsilon_i; i = 1..N & (27)\\ d_i^* &= z_i'\gamma + v_i; i = 1..N & (28)\\ d_i &= h(\ d_i^*) & (29)\\ y_i &= j(d_i, y_i^*) & (30)\end{aligned}$$

  where we assume that ϵ_i and v_i are bivariate normally distributed and at this point we restrict β to be constant for all values of d_i.

- The selection mechanism now has the generic form $h(.)$ and the process determining the observability of y_i has the form $j(.)$.

A. Ordered Censoring Rules

- The first case we examine is where $h(.)$ generates a series of ordered outcomes through the following rule:

  $$\begin{aligned} d_i &= 1 \text{ if } -\infty < d_i^* \leq 0; d_i = 2 \text{ if } 0 < d_i^* \leq \mu_1; .... \\ ....d_i &= 3 \text{ if } \mu_1 < d_i^* \leq \mu_2; ...d_i = J \text{ if } \mu_{J-1} < d_i^*; \end{aligned}$$

  where the $\mu's$ denote separation points satisfying $\mu_0 < \mu_2.. < \mu_J$ where μ_0 and μ_J equal -∞ and +∞ respectively, and noting we may only observe y_i for a specified value(s) of d_i.

- That is, the $j(.)$ function specifies that $y_i = y_i^* * I(d_i = j)$. It is now necessary to incorporate the ordering of the outcomes when accounting for the selection bias.

- This model is considered in Vella (1993) and following that general methodology we estimate the first step by ordered Probit to obtain estimates of the $\mu's$ and $\gamma's$.

- We then compute the generalized residuals for each outcome, $d_i = j$, which take the form:

$$\frac{\phi(\hat{\mu}_{j-1} - z_i'\hat{\gamma}) - \phi(\hat{\mu}_j - z_i'\hat{\gamma})}{\Phi(\hat{\mu}_j - z_i'\hat{\gamma}) - \Phi(\hat{\mu}_{j-1} - z_i'\hat{\gamma})}$$

  which we include as the selection correction rather than the Inverse Mills ratio. We can then estimate over the various sub-samples corresponding to different values of d_i.

- A second model which exploits the ordering in the selection equation is the continuous selection model of Garen (1984).
- Garen considers the case where d_i^* is continuously observed (namely, $d_i = d_i^*$) and there is a subset of the N observations corresponding to each permissible value of d_i^*.

- Rather than estimate a different set of parameters for each value of d_i^* Garen suggests estimation of:

$$y_i = x_i'\beta + d_i\alpha + (d_i * v_i)\theta_1 + v_i\theta_2 + \eta_i \tag{31}$$

where we have replaced the y_i^* and d_i^* with their observed counterparts

- To implement this procedure we obtain an estimate of $v_i,$ denoted $\hat{v}_i,$ by estimating the reduced form by ordinary least squares. We then replace v_i with $\hat{v}_i$ and estimate (31) by ordinary least squares.
- Note that the ordering of the outcome variable in the selection equation is important as it ensures the residual has the appropriate interpretation.

B. Tobit Type Censoring Rule

- A further type of commonly encountered censoring is one in which the dependent variable in the selection equation is partially observed.
- This model is also known as Tobit type three. For example, in the labor supply context we often not only observe if the individual works but also observe the number of hours they work. This information can be exploited in estimation.

- To capture this process specify $h(.)$ as:

$$\begin{aligned} d_i &= d_i^* \; if \; d^* > 0 \\ d_i &= 0 \; otherwise \end{aligned}$$

  and specify the $j(.)$ function as:

$$y_i = y_i^* * I(d_i > 0).$$

- Thus the censoring variable is observed whenever it is greater than some threshold and equal to zero otherwise.
- Furthermore, the dependent variable on the primary equation is only observed when the censoring variable is positive.
- The appropriate way to estimate the censoring equation is by Tobit.

- Following Vella (1993) we compute the generalized residuals which take the form:

$$(1 - I_i) * \{\frac{-\phi(z_i\hat{\gamma})}{(1 - \Phi(z_i\hat{\gamma}))}\} + I_i * \{d_i - z_i\hat{\gamma}).$$

- Note that when the second step estimation is only over the sample for which $I_i = 1$ the residuals have a very simple form.
- It is clear that this procedure is closely related to the original Heckman (1976) two-step procedure. Accordingly we refer to this as a control function procedure.
- The strong reliance on normality of this control function estimator could be relaxed in a number of ways.

- First, one could relax normality in the second step by taking a series expansion around $\hat{v}_i$ which, for the observations corresponding to $I_i = 1,$ is equal to $d_i - z_i\hat{\gamma}.$

- Second, to relax normality in both steps one could estimate γ semi-parametrically by using the procedures in Powell (1984, 1986).

- Using this semi-parametric estimate of γ the residuals for the uncensored observations could be estimated and then the primary equation could be estimated by OLS while including the estimated residual, and possibly its higher order terms, as additional regressors.

- Lee and Vella (1997) suggest an estimator based on (21).
- They note that the selection bias operates through the reduced form.
- That is, the selection bias is generated by the presence of v_i in the primary equation.
- Accordingly, they suggest purging the model of the component contaminated with v_i.

- To do this they propose an estimator based on:
$$y_i - E[y_i|v_i] = \{x_i - E[x_i|v_i]\}'\beta + \eta_i; i = 1..n \tag{32}$$
To implement this procedure they require a $\sqrt{n}$ consistent semi-parametric estimator of the censored model to obtain an estimate of γ and they suggest the use of the estimators proposed by Powell (1984, 1986).

- They define the residuals for the sub-sample corresponding to $d_i > 0$ as $\hat{v}_i = d_i - z_i'\hat{\gamma}$ and propose estimation by the methodology in (32).

C. Unordered Censoring Rules

- A feature of these two extensions of the selectivity model is that while estimation is somewhat complicated by the presence of multiple outcomes it is greatly simplified by the imposition of ordering on the outcomes.
- When it is not possible to impose such ordering it is necessary to treat the outcomes in the first step as unordered.
- One possibility would be to estimate the first step by multinomial Probit and then compute the corresponding generalized residual to include as an additional regressor.
- Such an approach, however, will be difficult to implement whenever there are more than three outcomes.

- Two alternative approaches are those outlined by Lee (1983a), and Hay (1980) and Dubin and McFadden (1984).
- To analyze these approaches consider the following model:

$$\begin{aligned} y_{si} &= x'_{si}\beta_s + \epsilon_{si} \\ I^*_{si} &= z'_{si}\gamma_s + v_{si} \end{aligned} \tag{33}$$

  where the number of outcomes is given by $s = 1..M$.
- This model is characterized by a different parameter vector for each outcome.

- First, consider the approach of Lee (1983a). Assume the selection rule, determining the chosen outcome, is based on the following rule:

$$I_i = s \text{ iff } I^*_{si} > \max I^*_{ji}; j = 1..M; j \neq s.$$

- If we let $\kappa_{si} = \max I^*_{ji} - v_{si}$ it follows from the selection rule that:

$$I_i = s \text{ iff } \kappa_{si} < z'_{si}\gamma_s; j = 1..M; j \neq s.$$

Thus the model can now be characterized by a series of observed y_{si} which are only observed if $\kappa_{si} < z'_{si}\gamma_s$.

- When the distribution function of κ_{si} is known we are able to proceed in the same manner as for the binary choice model. That is, we estimate:

$$E[y_i|z_i, d_i = 1] = x_i'\beta + \rho\sigma_{\epsilon v}\phi[J_2(z_i'\gamma)]/G[z_i'\gamma] \qquad (34)$$

  where we either assume the marginal distribution of the untransformed ϵ_i is normal or that the relationship between ϵ and transformed v_s is normal.

- We require a first step estimate of γ which accounts for the polychotomous nature of I_i.

- A popular way to proceed is to assume that v_{si} has a type 1 extreme value distribution which then allows estimation of γ_s by multinomial Logit.

D. Censoring Rules Based on Multiple Indices

- A feature of many of these models, even in the case of unconventional forms of censoring, is that the selection bias is treated purely as a function of a single index.
- Now consider the following re-characterization of our original model where the selectivity bias is a function of multiple indices:

$$
\begin{aligned}
y_i^* &= x_i'\beta + \varepsilon_i && (35)\\
d_{1i} &= I(z_i'\gamma_1 > -v_{1i}) && (36)\\
d_{2i} &= I(z_i'\gamma_2 > -v_{2i}) && (37)\\
y_i &= y_i^* * d_{1i} * d_{2i} && (38)
\end{aligned}
$$

where $I(.)$ is an indicator function and the additional notation is obvious

- The sample selection is now based on multiple indices and multiple criteria.
- The method of estimation relies crucially on;
- i) the relationship between v_{1i} and $v_{2i};$ and
- ii) the observability of the two indices d_{1i} and $d_{2i}.$

- The simplest case is where the disturbances are jointly normally distributed; v_{1i} and v_{2i} are uncorrelated and both d_{1i} and d_{2i} are observed.
- In that case it is relatively straightforward to use the procedures discussed above to compute the following correction terms to include as regressors in the primary equation:

$$\begin{aligned} E[\varepsilon_i | z_i, d_{1i} &= 1, d_{2i} = 1] \\ &= (\sigma_{\varepsilon v_1}/\sigma_{v1}^2)\frac{\phi(z_i'\gamma_1)}{\Phi(z_i'\gamma_1)} + (\sigma_{\varepsilon v_2}/\sigma_{v_2}^2)\frac{\phi(z_i'\gamma_2)}{\Phi(z_i'\gamma_2)}. \end{aligned}$$

- To implement this model one first independently estimates (36) and (37) by Probit to obtain γ_1 and γ_2.
- The corresponding two Inverse Mills ratios can then be computed and included as correction terms in the primary equation.

- While this model is easily estimated it restricts the error terms in the two censoring equations to be uncorrelated.
- This is an assumption that most empirical studies would be reluctant to impose.
- Moreover, one could imagine that the two different selection rules would be related in various ways above the nature of the correlation of their respective disturbances.

- Perhaps the most commonly encountered of this comes under the heading of partial observability examined by Poirier (1980).
- In Poirier's model neither d_{1i} or d_{2i} are observed but we observe their product $d_{3i} = d_{1i} * d_{2i}$.
- Furthermore, we assume we only observe the $y_i's$ for the sub-sample corresponding to $d_{3i} = 1$.
- Poirier examines the conditions for the estimation and identification of γ_1 and γ_2 by maximum likelihood while employing d_{3i} as the dependent variable.

- Furthermore, he also shows:

$$\begin{aligned} E[\varepsilon_i|z_i, d_{3i} &= 1] = \sigma_{\varepsilon v_1}\left[\frac{\phi(z_i'\gamma_1)\Phi(z_i'(\gamma_2 - \rho_{12}\gamma_1)/(1-\rho_{12}^2)^{\frac{1}{2}}}{\Phi^b(z_i\gamma_1, z_i\gamma_2; \rho_{12})}\right] \\ &+\sigma_{\varepsilon v_2}\left[\frac{\phi(z_i'\gamma_2)\Phi(z_i'(\gamma_1 - \rho_{12}\gamma_2)/(1-\rho_{12}^2)^{\frac{1}{2}}}{\Phi^b(z_i'\gamma_1, z_i'\gamma_2; \rho_{12})}\right] \end{aligned}$$

- where Φ^b denotes the bivariate normal distribution; ρ_{12} denotes the correlation coefficient for v_1 and v_2; and we normalize $\sigma_{v_1}^2 = \sigma_{v_2}^2 = 1$.
- To adjust for sample selection in this model one computes the above two additional terms to include as regressors in the conditional mean function for y_i.

E. Conditional Maximum Likelihood

- Thus far we have focussed on models where the dependent variable in the selection equation is censored and we have a continuous variable in the primary equation.
- In many instances we confront a continuous, or partially continuous, dependent variable in the selection equation and a censored or limited dependent variable in the primary equation.

- The model has the following structure:

$$y_i^* = x_i'\beta + \theta d_i + \epsilon_i; i = 1..N \tag{39}$$

$$d_i = z_i'\gamma + v_i; i = 1..N \tag{40}$$

$$y_i = l(y_i^*); \tag{41}$$

$$y_i = j(d_i) \tag{42}$$

where we assume that the error terms are jointly normally distributed with non-zero covariance and the $l(.)$ function maps the latent y_i^* into the observed y_i noting that at this stage the y_i is reported for the whole sample.

- This model is similar to (11)-(13) except that the censoring occurs in the primary equation and not the reduced form.

- The model in (39)-(41) is considered by Smith and Blundell (1986), where $l(.)$ generates Tobit type censoring, and by Rivers and Vuong (1989) for the case where $l(.)$ generates Probit type censoring.

- Single equation maximum likelihood estimation of (39), while accounting for the form of $l(.)$, will not produce consistent estimates of β due to the endogeneity of d_i.

- One method of estimation is conditional maximum likelihood by which one first employs the bivariate normality assumption to rewrite (39) as:

  $$y_i^* = x_i'\beta + \theta d_i + \mu v_i + e_i; i = 1..N$$

  where $\mu = \sigma_{\epsilon v}/\sigma_v^2$ and e_i is a zero mean and normally distributed error term. As v_i is normally distributed we are able to obtain a consistent estimate as $\hat{v}_i = d_i - z_i'\hat{\gamma}$ where $\hat{\gamma}$ denote the OLS estimates from (40).

- One then estimates:

$$y_i^* = x_i'\beta + \theta d_i + \mu\hat{v}_i + e_{1i}; i = 1..N \tag{43}$$

  by maximum likelihood noting that $e_{1i} = e_i + \mu(v_i - \hat{v}_i)$ has zero mean and, most importantly, is normally distributed.

- The normality is retained as $\hat{v}_i$ is a linear transformation of normally distributed random variables.

- This differs from the model in (11)-(13) as the censoring of the endogenous variable in the reduced form results in the generalized residuals being non-linear functions of normally distributed random variables and the variable $(v_i - \hat{v}_i)$ is subsequently non-normal.
- As the coefficient μ captures the correlation between the errors a $t-$test on the null $\mu = 0$ is a test of the weak exogeneity of d_i.

- The conditional maximum likelihood estimator can be extended to the sample selection case if the selection rule, captured by the $j(.)$ function in (42), is within a certain class of functions.
- For example, if y_i was only observed when $d_i > 0$ then the second step estimation would only involve the sub-sample satisfying this selection rule and one would still estimate the primary equation by maximum likelihood with $\hat{v}_i$ included.
- Moreover, in this sub-sample case it is necessary to include $\hat{v}_i$ even if $\theta = 0$ whereas this is unnecessary if we observe the whole sample. Finally, despite only observing y_i for specified values of d_i we are still able to estimate by maximum likelihood as the error term retains its normality despite the inclusion of $\hat{v}_i$.

- The above discussion illustrates that when the second step estimation of the primary equation is performed by maximum likelihood it is necessary to impose some restrictions on the mapping from the first step parameter estimates, and variables, to the residuals operating as correction factors.

- More explicitly, the inclusion of the correction factor cannot corrupt the normality of the primary equation's disturbance.

- This naturally does not apply to models estimated by full information maximum likelihood in which case the selection and primary equation's dependent variables can take any sensible form and estimation can proceed providing the likelihood function can be constructed.

- However, as noted above it is clear that maximum likelihood can be employed in the second step whenever the residuals are a linear function of the variables as this transformation preserves the assumed normality.

- One particular case of interest is considered by Vella (1992) who examines a model where the primary equation has a binary outcome variable and the selection equation has a dependent variable which is partially observed and has Tobit type censoring.

- For that model it is possible to perform the reduced form first step estimation over the entire sample by Tobit. One then estimates the Tobit residuals for the sub-sample corresponding to $d_i > 0$ which simply take the form $\hat{v}_i = d_i - z_i'\hat{\gamma}$ where the hats now denote the Tobit estimates.

- It is then possible to estimate the primary equation by Probit over the subset satisfying $d_i > 0$ while including $\hat{v}_i$ as an explanatory variable.

- However, while this is how one would proceed for a model with Probit type censoring it is possible to estimate a number of models, depending on the form of $l(.)$, provided they require normality.