Chapter 5

Econometric Approaches to the Evaluation Problem: Structural Models

5.1 Introduction

The methods described in Chapter 4 assume that the selection into the program occur on the basis of observable (to the analyst) individuals' characteristics. In fact, this is the most troubling feature of these methods. Depending on the quality of the data at the analyst's disposal, it may or not may be attractive to assume that the analyst knows as much as people, being studied. When analyst's knowledge is perfect, matching methods would yield robust results. However, if it is implausible to assume that the analyst observes all the individuals' characteristics that affect their decision to participate, the dependence between Y_i and the training indicator variable D_i is not eliminated even after controlling for X_i. Under such circumstances, the method of matching is not robust anymore and the problem has to be approached with alternative econometric methods. Also, since X_i may not be equal to Z_i so that variables that affect participation may be different from those affect outcomes (in fact it is usually assumed that X_i is included in Z_i), we use both X_i and Z_i terms on this chapter.

The traditional econometric approach to the selection model adopts a more conservative approach and allows for selection on unobservables. Several methods have been proposed to cope with this problem. These methods, initially developed by Roy (1951), and extended by Lewis (1974), Gronau (1974), Heckman (1974, 1978, 1979) are going to be discussed in this chapter.

5.2 Historical Review

Roy (1951) with the development of the *General Equilibrium model* set up the conditions for the growth of structural (model-based) econometric approaches in evaluation studies. In his seminal work, he considered two occupational statuses, say 0 and 1, in which income-maximizing agents can work. Agents are allowed to have heterogeneous skills and, more important, are free to enter the sector that gives them the highest income. However, they are restricted to only one occupation for a period. Under the assumption of normality for the logarithms of the incomes, Roy concluded that the distribution of earnings depends on certain "real" factors, i.e. the character of the distributions of various kinds of human skill and the state of technique existing in different occupations. The desires of the individuals in the community for various sorts of goods were found also of great importance, but they were only able to exert their influence within the framework determined by skill and technique. It must be emphasized, though, that the conclusions reached are dependent upon the normality assumption of the log incomes. They would not necessarily be true if other types of distribution were considered.

Extending the work of Roy, Heckman (1974) observes a population of women that either participate in the labor force (sector 1 participants) or do not participate (sector 0 participants). Their decision to work depends on the wage they face in the market. Only when their asking wage exceeds the offered wage the women work. Under the normality assumption for the logarithms of wages, the probability that a woman works, her asking wage and her offered wage are mainly sought to be estimated from a common set of parameters.

The problem in obtaining these estimates come from the missing information (missing data) occurred from the inability to report wages for the "sector 0" women. Approaching this problem by estimating a wage function on a subsample of working women to estimate the missing wages and then performing OLS regression over all wages would give an estimate of the mean offered wage. However, as Heckman (1974) explained, this procedure would lead to biased parameter estimates for the wages because of the, possibly, selective nature of work decision.

Kalachek and Raines (1970) estimate an "expected hours of work" equation for all women by separately estimating hours of work relation of working women, and an equation determining the probability that a woman works. Leibowitz (1973) describes a procedure where observations of non-working women are assumed to lie on the same hours of work function as observations of working women, with a particular value of zero for their hours of work. However, all these applications led to implausible estimates due to the perceived assumption that conditional on X_i and Z_i, outcomes Y_i are mean independent of D_i, or mathematically:

$$E\left(Y_i \middle| X_i, Z_i\right) = E\left(Y_i \middle| D_i = 1, X_i, Z_i\right) = E\left(Y_i \middle| D_i = 0, X_i, Z_i\right) \tag{5.1}$$

The plausibility of this assumption has also been questioned sharply by Heckman and Sedlacek (1985). Although in an experimental setting, randomization can produce such data, observational studies do not satisfy this assumption due to the selective nature of participation decision.

Instead of these erroneous approaches, Heckman (1974) applied a prototypic procedure for the estimation of the mean wage of married women without the implausible restriction (5.1). Specifically, he develops two simultaneous behavioral models; the first one determines the offered wage to the woman while the second her asking wage. If a woman works, her hours of work adjust to equate these wages if she has freedom to set her working hours. If a woman does not work, no offered wage matches her asking wage (in fact her asking wage is greater than the offered). The simultaneous estimation of this models, lead to unbiased estimators of the probability that a woman works, of her actual hours of work given that she works, of the potential market wage rates facing non-working women and of the implicit shadow price (the highest amount of the offered wage for which a woman is not interested in working) for non-working women.

Upon these findings, Gronau (1974) reports that the wage rate, and thus participation in the labor force, depends not only on the wage offered (a function of individuals' market characteristics), but also on the job-search strategy a person follows. The higher the person's wage demands, the higher the wage he can expect, though the probability of finding an adequate job is lower. Ignoring this relationship results in selectivity bias.

Under that perspective, the author also comments on the selectivity biases that occur in wage comparisons. He examines the selection bias in the case where the population under study is homogeneous. Specifically, he recognizes that in such a population the bias α measures the extend to which the mean market wage, $\overline{E}$, of employed persons in the group overstates the mean μ_W of their common wage-offer distribution. Adopting this, he proceeds with the extension of selection bias in wage comparisons between two homogeneous groups.

Lewis (1974) examines selection bias under the assumption of a heterogeneous population, an issue not considered by Gronau (1974). He also recognizes the existence of selection bias in fields other than wage comparisons and labor-force participation. As illustrative examples, he mentions the returns from schooling (compares the wages of persons with different amounts of schooling) and the return to geographic migration (compares the destination wages of migrants with the corresponding origin wages of non-migrants).

Later, Heckman (1978) develops the "dummy endogenous variable model" and supports that it produces credible estimates of program impacts. As a member of the simultaneous equation models family, this model is described to include both discrete (endogenous) and continuous random variables based on normally distributed latent random variables. In Heckman's (1978) seminal work, conditions for its existence and identification criteria are provided, and consistent estimators are proposed.

Heckman (1979) adopts a similar approach and discusses the bias that results from using nonrandom selected samples to estimate behavioral relationships. Specifically, he develops a computationally tractable technique based on simple regression methods to estimate behavioral functions that account for selection bias in the case of a censored sample. Asymptotic properties of the estimator are also derived. This model and its descendants are mainly discussed in this chapter.

Singh et al (1979) propose a model where the variable being studied (duration of the postpartum amenorrhea) is directly related to the probabilities of selection in the study. The selection bias that results, leads in underestimation of the level of the variable. The nature of selection bias is then investigated and an analytical procedure is outlined to adjust the sample estimate of the mean value in which they are interested. Greene (1981)

comments on Heckman's (1979) method and proposes an alternative modeling approach to correct for selection bias.

5.3 Useful Notation

In the context of a social program, assume a sample of individuals with heterogeneous skills (target population) that are free to decide between *participation* (sector 1) or *not participation* (sector 0). Following Roy (1951), each individual is restricted to belong to only one sector at a time. There are no costs of changing sectors and investment is ignored.

5.3.1 The Evaluation Problem

Suppose that an analyst observes the post-program incomes of a sample of N individuals ($i = 1, ...,$ N). Denoting as T_i the $N \times J$ skill vector for persons i with skills j, post-program outcomes can be modeled as a function of T_i and a $N \times 1$ vector of unobserved (to the analyst) individual characteristics u_i. In the general econometric approach, the functions of outcomes for each sector are postulated in the following way:

$$Y_{0i} = g_0(T_{0i}, u_{0i}) \quad \text{and} \quad Y_{1i} = g_1(T_{1i}, u_{1i})$$

Considering *additive separability* assumption, outcomes Y_{0i} and Y_{1i} can be represented, respectively, as:

$$Y_{1i} = g_1(T_i) + u_{1i} \quad \text{and} \quad Y_{0i} = g_0(T_i) + u_{0i}$$

Each person has a (Y_{0i}, Y_{1i}) pair but only one outcome can be observed at a time for unit i according to its sector decision. Then, the unobserved outcome is regarded as a *latent variable*.

From the econometricians' standpoint, utility theory indicates that an individual i with endowments T_i participates in the program (sector 1), if participation maximizes his post-program earnings, that is if:

$$Y_{1i} \succ Y_{0i} \Leftrightarrow g_1(T_{1i}) + u_{1i} \succ g_0(T_{0i}) + u_{0i}$$

where $g_1(T_{1i})$ and $g_0(T_{0i})$ denote utility functions (see McFadden, 1975). The evaluation problem does not allow a straight comparison of the outcomes. Simple regression methods for estimation of $\overline{Y}_1$ and $\overline{Y}_0$ across individuals results in biased estimates due to the selective nature of participation decision. Thus, other approaches have to be considered.

5.3.2 Description of the Structural Approach

The conventional econometric approach addresses selectivity in terms of unobservables to the analyst individuals' characteristics u_i by postulating econometric (structural) models. Heckman (1974) corrects for selection bias in the estimation of program impacts through the estimation of the participation probability for each individual, given a vector of attributes. To attain this result in terms of a simple computational procedure, Heckman (1978, 1979) partitions the endowments vector T_i into two, not necessarily disjoint, sets (X_i, Z_i). Then, he estimates the models:

$$\begin{aligned} Y_{1i} &= g_1(X_{1i}) + u_{1i} \\ Y_{0i} &= g_0(X_{0i}) + u_{0i} \\ D_i &= Z_i\gamma + v_i \end{aligned} \qquad (5.3)$$

where the last equation accounts for the probability of participation. In a linear regression framework, equation (5.3) is specialized in the familiar form:

$$\begin{aligned} Y_{1i} &= \alpha_i + \beta_1 X_{1i} + u_{1i} \\ Y_{0i} &= \alpha_0 + \beta_0 X_{0i} + u_{0i} \\ D_i &= Z_i\gamma + v_i \end{aligned}$$

To estimate adequately the above models an *exclusion restriction* is invoked. More specifically, Z_i vector must contain at least one variable that does not appear in X_i. In a parametric setting this restriction is not strictly required, but in semi-parametric procedures is essential to be satisfied.

Regarding selection on unobservables, expression

$$E\left(u_i \middle| D_i, X_i, Z_i\right) \neq 0$$

indicates that participation depends on the unobserved variable u_i, resulting in selection bias. For this reason D_i is regarded as an *endogenous* variable. Alternatively, selection on unobservables can be represented by the relationship

$$E\left(u_i \middle| v_i\right) = \rho \times v_i \neq 0$$

that indicates the correlation between the unobservables u_i and v_i.

5.3.3 Definition of Mean Impacts

In terms of a linear model, the gain from participation in a program for each person *i* can be expressed as:

$$\Delta_i = Y_{1i} - Y_{0i} = \left(\alpha_1 + \beta_1 X_{1i} + u_{1i}\right) - \left(\alpha_0 + \beta_0 X_{0i} + u_{0i}\right)$$

The gain has two components, namely the gain for the average person with characteristics X_i

$$\alpha_1 + \beta_1 \mathrm{X}_1 - \alpha_0 - \beta_0 \mathrm{X}_0$$

and the idiosyncratic gain

$$u_{1i} - u_{0i}$$

The idiosyncratic components may be observed by the person deciding participation in the program, but not by the econometrician who evaluates the program. In this notation, the mean impacts of a social program can be formulated.

The most commonly estimated mean parameters in modern applied econometrics are the Average Treatment Effect (ATE) and the Effect of Treatment on the Treated (TT). At this point it is worth discussing them in a non-experimental framework.

The Average Treatment Effect parameter (ATE)

This parameter is defined as the expected gain from participating in the program for a randomly chosen individual. The ATE, conditional on $X_i = x_i$, is given by:

$$\begin{aligned} ATE(X) = E\left(\Delta_i \middle| X_i\right) &= E\left(Y_{1i} - Y_{0i} \middle| X_i\right) \\ &= \left(\alpha_1 + \beta_1 X_{1i} - \alpha_0 + \beta_0 X_{0i}\right) + E\left(u_{1i} - u_{0i} \middle| X_i\right) \end{aligned}$$

Note that due to selectivity $E\left(u_{1i} - u_{0i} \middle| X_i\right) \neq 0$ and thus a simple mean comparison does not yield plausible estimates of the mean effect. In practice, most evaluation studies do not estimate $E\left(\Delta_i \middle| \mathrm{X}_i\right)$. Since it does not account for participation status D_i, ATE(X) does not answer economically interesting questions in program evaluations and other estimators are preferred.

The Treatment on the Treated parameter (TT)

This is the average gain from treatment for those that actually select participation. In other words this parameter expresses the gain from moving a participation person with attributes X_i from the non-participation to participation state. This is what is called *wage gap* by Heckman (1990b). Treatment on the Treated effect, conditional on X_i is calculated by:

$$\begin{aligned} TT(X) = E\left(\Delta_i \middle| D_i = 1, X_i, Z_i\right) &= E\left(Y_{1i} - Y_{0i} \middle| D_i = 1, X_i, Z_i\right) \\ &= \left(\alpha_1 + \beta_1 X_{1i} - \alpha_0 + \beta_0 X_{0i}\right) + E\left(u_{1i} - u_{0i} \middle| D_i = 1, X_i, Z_i\right) \end{aligned}$$

Again, selectivity causes $E\left(u_{1i} - u_{0i} \middle| D_i = 1, X_i, Z_i\right) \neq 0$ and specific methods are required to evaluate the above expression. In practice, this is the most popular parameter in evaluation studies because it reflects the difference of participants' outcomes had they not participated in the program $\left(Y_{0i} \middle| D_i = 1, X_i, Z_i\right)$ from participants' outcomes under participation $\left(Y_{1i} \middle| D_i = 1, X_i, Z_i\right)$.

5.4 The Conventional Selection Bias Model (Tobit Type – II Censoring)

By definition, evaluation of the above mean parameters requires unbiased estimation of the corresponding linear models for participants and non-participants. Several procedures, either parametric or semi-parametric, have been proposed. At this point, we discuss them, beginning with the *conventional selection bias* model of Heckman (1979), which is a two-equation model of the form:

$$Y_i^{(1)} = X_i\beta + u_i; \qquad i = 1,...,N \tag{5.10}$$

$$D_i^{(1)} = Z_i \times \gamma + v_i; \qquad i = 1,...,N \tag{5.11}$$

$$D_i = 1 \ \ if \ \ D_i^{(1)} \succ 0; \ \ D_i = 0 \ \ otherwise \tag{5.12}$$

$$Y_i = Y_i^{(1)} \times D_i \tag{5.13}$$

The crucial feature of the above formulation is the censoring rule that is imposed as:

Censoring Rule 1 (Tobit Type – II censoring rule)
Only the sign of $D_i^{(1)}$ is observed and this sign determines whether an individual participates or not.

$D_i^{(1)}$ is a latent variable that accounts for the participation decision. Specifically, when participation is identified by a prescribed rule, i.e. one participates if the offered wage in he faces in the market exceeds a threshold (e.g. his asking wage), then equation (5.11)

determines the probability of participation, denoted by $P(D_i = 1|X_i, Z_i)$. Relatively to equation (5.12) that defines $D_i = 1 \;\; if \;\; D_i^{(l)} \succ 0$, the choice of zero as a threshold reflects to an inessential normalization. γ is again a $J \times 1$ vector of parameters of the effect of attributes in the participation probability while β is a $J \times 1$ vector of parameters of the effect of attributes on the post-program wages.

The importance of equation (5.13) in this formulation is minor. Its inclusion indicates that only Y_{1i} (participants' outcome) are observed (sample selectivity) while its exclusion indicates that both Y_{1i} and Y_{0i} can be recorded. Model (5.10) is known literarily as the *primary model* and *the model of interest*, while (5.11) as the *participation model,* the *selection model* and the *discrete choice model*.

This formulation is applicable in several social fields, e.g. evaluation of training programs for workers. The primary interest is to estimate parameter β of equation (5.10) for participants and non-participants (if (5.13) does not exist). As we denoted before, OLS procedure leads in biased estimates due to selectivity on unobservables that causes $E(u_{Di}|D_i, X_i, Z_i) \neq 0$. Therefore, other procedures have to be considered.

5.5 Parametric Methods

5.5.1 Maximum Likelihood Estimation

Heckman (1974) proposed an appealing procedure to account for sample selection bias. He stated the following assumption:

Assumption 1

u_i and v_i, i = 1, …, N, are independent of Z_i, and independently and identically distributed (iid) over the entire population (participants and non-participants) with the bivariate Normal distribution N(0, Σ), where:

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{vu} & \sigma_v^2 \end{bmatrix}$$

The phrase "over the entire population", inserted in Assumption 1 is crucial. Basically, it discriminates the selection models from the mixture-distribution models where the distribution of u_i, i = 1, …, N, is defined only for a subpopulation of persons (e.g. participants). In a discussion for the latter family of models, Maddala (1983) states that despite its observational equivalence with sample selection bias models, mixture-distribution models cannot be applied in evaluation studies.

Under Assumption 1 the parameters of the model can be estimated by Maximum Likelihood method. The log-likelihood to be maximized is:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left\{ D_i \times ln\left[\int_{-Z_i'\gamma}^{\infty} \phi_{uv}\left(Y_i - X_i\beta, v_i\right) dv_i \right] \right.$$
$$\left. + \left(1 - D_i\right) \times \left[\ln \int_{-Z_i'\gamma}^{\infty} \int_{-\infty}^{\infty} \phi_{uv}\left(u_i, v_i\right) du_i dv_i \right] \right.$$

where φ_{uv} denotes the probability density function for the bivariate normal distribution of (u_i, v_i). Maximum Likelihood method is easy to be implemented while it yields consistent and fully efficient parameter estimates given Assumption 1. Relaxation of this assumption is accompanied by an efficiency loss.

5.5.2 2-Step Estimation

Given the necessary distributional assumptions, a likelihood function, which accounts explicitly for the selection mechanism is theoretically easy to be derived and maximized. Nevertheless practically, the likelihood equations may be non-linear and thus the form of the function will be relatively complicated while the required computer programming will be difficult and the computational costs high. These computational difficulties led Heckman (1979) to propose a simple 2-step estimator. The author suggested estimation of γ in the first step by modeling D_i as a dichotomous variable as in logit or probit analysis. Under Assumption 1, then, $\hat{\gamma}$ contributes to estimate β and Y_i without bias.

More mathematically, recall model (5.10) – (5.13) and remember that estimation of equation (5.10) is of primary interest although equation (5.11) have also to be estimated using conventional methods. Heckman (1979) suggested overcoming the bias problem through the inclusion of a correction term (*control function*) that accounts for $E\left(u_i \middle| X_i, Z_i, D_i\right)$.

To see this observe that if all individuals participated *(D_i = 1)*, the primary equation (5.10) would be:

$$E\left(Y_{1i} \middle| X_{1i}, Z_i\right) = \mathrm{X}_{1i} \times \beta_1 \tag{5.15}$$

However, from the population N, only a subsample N_1 participates under a specific selection rule that produces selectivity. Therefore (5.15) is rewritten as (see Panaretos, 1974):

$$\begin{aligned} E\left(Y_{1i} \middle| D_i = 1, X_{1i}, Z_i\right) &= \mathrm{X}_{1i} \times \beta_1 + E\left(u_i \middle| D_i = 1, X_i, Z_i\right) \\ &= \mathrm{X}_{1i} \times \beta_1 + E\left(u_i \middle| v_i \succ -Z_i\gamma, X_i\right) \end{aligned} \tag{5.16}$$

The regression function (5.16) depends X_i and Z_i. Regression estimators of its parameters fit on the selected sample omit the final error term as a regressor, so that the bias that results from using non-randomly selected samples to estimate behavioral relationships is seen to arise from the econometric problem of *omitted variables* of Griliches (1957).

To describe how this problem is solved remember that Assumption 1 validates the property:

$$\begin{aligned} E\left(u_i \middle| v_i\right) &= \sigma_v^{-1} \times \sigma_{uv} \times v_i \\ &= \sigma_u \times \rho \times v_i \end{aligned}$$

indicating the selectivity bias problem that according to Olsen (1980) occurs from the dependence (correlation ρ) between u_i and v_i. By using the *theory of biserial correlation* of Kotz, Johnson and Campell (1986, Vol. 4), one corrects for selectivity in the sample by using the Mill's ratio:

$$R_i = \frac{1 - \Phi\left(-Z_i\gamma / \sigma_v\right)}{\phi\left(-Z_i\gamma / \sigma_v\right)}$$

"R_i" is a monotone decreasing function of the probability that an observation is selected into the sample, $\Phi\left(Z_i\gamma / \sigma_v\right)$. This formula has appeared in approximating binomial and geometric probabilities in terms of standard Normal distribution (see Johnson and Kotz, 1972). In particular, φ and Φ are, respectively, the density and the distribution function for a standard normal variable and

$$lim_{\Phi\left(-Z_i\gamma/\sigma_v\right)\to 1} R_i = 0 \quad and \quad lim_{\Phi\left(-Z_i\gamma/\sigma_v\right)\to 0} R_i = \infty$$

(A more detailed discussion on the Mill's ratio is found in Appendix 1).

The primary model of outcome Y_{1i} for participants $D_i = 1$ can be written as:

$$E\left(Y_{1i} \middle| D_i = 1, X_i, Z_i\right) = X_i\beta + \sigma_u \times \rho \times E\left[\frac{1 - \Phi\left(-Z_i\gamma / \sigma_v\right)}{\phi\left(-Z_i\gamma / \sigma_v\right)} \middle| D_i = 1, X_i\right]$$

$$= X_i\beta + \sigma_u \times \rho \times \left[\frac{\phi\left(-Z_i\gamma / \sigma_v\right)}{1 - \Phi\left(-Z_i\gamma / \sigma_v\right)}\right]$$

$$= X_i\beta + \sigma_u \times \rho \times \left[\frac{\phi\left(-Z_i\gamma / \sigma_v\right)}{\Phi\left(Z_i\gamma / \sigma_v\right)}\right]$$

$$= X_i\beta + \sigma_u \times \rho \times \lambda_i \qquad (5.17)$$

where λ_i is the inverse Mill's ratio whose coefficient $\rho \times \sigma_u$ accounts for the selectivity bias. In practice, λ_i is not known. Heckman (1979) suggested a procedure that estimates λ_i and in extend β, $\rho \times \sigma_u$ and Y_i adequately.

Figure 5.1: Heckman's 2-step procedure

1. Estimate the parameters of the probability that $D \geq 0$

$$P(D_i^{(1)} \mid Z_i, D_i \geq 0) = Z_i\gamma / \sigma_v = Z_i\gamma^*$$

using probit analysis for the full sample[1].

2. From the above estimator one can estimate γ^* and hence λ_i. All of these estimators are consistent.

3. The estimated value of λ_i may be used as a regressor in equation (5.17) fit on the selected subsample. These regression estimators are consistent for β and $\rho \times \sigma_u$.

4. One can consistently estimate σ_u by the following procedure. Denote as $C = \rho \times \sigma_u$, as $\hat{u}_i$ the residual for the i^{th} observation obtained from step 3, and the estimator of C by $\hat{C}$. Then, an estimator of σ_u is:

$$\hat{\sigma}_u = \frac{\sum_{i=1}^{N_1} \hat{u}_i^2}{N_1} - \frac{\hat{C}}{N_1} \times \sum_{i=1}^{N_1} \left(\hat{\lambda}_i \times \hat{Z}_i - \hat{\lambda}_i^2 \right)$$

where $\hat{\lambda}_i$ and $\hat{Z}_i$ are the estimated values of λ_i and Z_i obtained from step 2. This estimator of σ_u is consistent.

Intuitively from equation (5.17), a test on the null hypothesis $C = 0$ is equivalent to $\sigma_{uv} = 0$ and represents a test for the existence of sample selectivity bias. When $C = 0$, no selection bias exists since equation (5.17) reduces to (5.15) where all individuals participate. Melino (1982) shows that the square of the t – statistic, t^2, approximates the Langrange multiplier test and supports that it is the optimal test of selectivity bias. Moreover, by the standard large sample criteria, this test is equivalent to the likelihood

[1] Since $D_i^{(1)}$ is a continuous variable OLS method over the full sample is applied to estimate γ^*. Then estimation of the probability that $D_i = 0$ or $D_i = 1$ is conducted by the Probit model.

ratio and Wald tests. Due to its computational simplicity, this test is also recommended by Heckman, Tobias and Vytlacyil (2000) as a diagnostic tool for selection bias.

Performing this estimation procedure, Vella (1998) estimates adequately the wages of a sample of 2.300 females taken from the 1987 wave of the National Longitudinal Survey in America. However, another question is raised here. The above model specification assumes that participation occurs if one criterion is satisfied, e.g. if and only if the wage offer exceeds the asking wage of the individuals. This assumption seems to be restrictive in most practical situations. It is possible that more than one criterion often determine participation. As Heckman (1979) indicated "*sample selectivity may arise in practice for two reasons. First, there may be self-selection by the individuals or data units being investigated. Second, sample selection decisions by analysts or data processors operate in much the same fashion as self-selection*". Unlike the experimental and matching procedures, modeling selection bias allows the analyst to study the same problem from different perspectives (different selection rules) and under different distributional assumptions. This aspect has been criticized sharply by Holland (1989) who considers as impractical the existence of the several different estimators for the different assumptions. This opinion is reviewed at the final paragraph of this chapter.

Heckman, Lalonde and Smith (1999) indicate the importance of estimating the *indirect benefits* from participation to the program and propose further research on this field. The number of non-experimental studies in this aspect is very limited. Typical work is the one of Davidson and Woodbury (1993), who consider the effects of a bonus program on the search behavior of participants and non-participants to evaluate the indirect effects of the program.

5.6 Dummy Endogenous Variable Model

An alternative representation of the selection bias model (5.10) – (5.13) is the one where the primary regression equation includes a coefficient for the dummy variable indicating participation. Adopting Assumption 1, this model can be written as:

$$Y_i = X_i\beta + \theta \times D_i + u_i; \qquad \text{i} = 1, ..., \text{N} \tag{5.20}$$

$$D_i^{(l)} = Z_i\gamma + v_i; \qquad i = 1, ..., N \tag{5.21}$$

$$D_i = 1 \;\; if \;\; D_i^{(l)} \succ 0; \;\; D_i = 0 \;\; otherwise \tag{5.22}$$

and $E\left(u_i \middle| D_i, X_i, Z_i\right) \neq 0$, $E\left(u_i \middle| v_i\right) = \rho \times v_i \neq 0$.

For example, consider the case of the effect of laws in the status of migrants. Let Y_{1i} indicate the outcome of migrants (payments) while D_i is an indicator that reflects the state's population sentiment toward migrants. If sentiment for migrants is sufficiently favorable $\left(D_i^{(l)} \succ 0\right)$, the state may enact antidiscrimination legislation and the presence of such legislation is denoted with $D_i = 1$. In the outcome equation (5.20) both the presence of a law and migrants' characteristics (X_i) affect their measured outcomes. Therefore both of them must be included in the regression equation.

The dummy variable D_i can be characterized as endogenous since it is correlated with the disturbance u_i, $Corr(D_i, u_i) \neq 0$, operated through the non zero covariance σ_{uv}, indicating the selective nature of the sample. For this reason this model is known as "dummy endogenous variable" model. This model is different from the (5.10) – (5.13) representation in two ways. First, Y_i outcomes are observed for both participants and non-participants and thus the model allows for a parallel estimation and comparison of their outcomes. Second, this model assumes a common parameter β for participants and non-participants while the previous one do not. This feature seems to be its major limitation related to the previous formula.

Again, a simple OLS procedure would lead to inconsistent estimates due to the fact that $E(u_i|v_i) \neq 0$. Hausman (1978) suggested that this inconsistency could be overcome by:

a) Projecting $D_i^{(l)}$ onto Z_i to obtain $\hat{D}_i^{(l)}$ and then replacing D_i with $\hat{D}_i$ in (5.20) and perform OLS regression to estimate the parameters of interest.
b) Projecting $D_i^{(l)}$ onto Z_i to obtain the residuals $\hat{v}_i$ and then including both $\hat{v}_i$ and D_i in (5.20) and perform OLS.

Heckman (1978) adopts an approach similar to the one described above for the conventional selection bias model. Under Assumption 1, parameter γ of the participation equation can be consistently estimated by OLS over the full sample. To this extend, the Generalized or Probit residuals of Gourieroux et. al. (1987) can be attained in a manner similar to the computation of the Mill's ratio for both participants and non-participants:

$$v_i = D_i \times C \times \left[\frac{\phi\left(-Z_i\gamma^*\right)}{\Phi\left(Z_i\gamma^*\right)} \right] + \left(1 - D_i\right) \times \left[\frac{-\phi\left(-Z_i\gamma^*\right)}{1 - \Phi\left(-Z_i\gamma^*\right)} \right]$$

Then the regression model to be identified is:

$$Y_i = X_i\beta + \theta \times D_i + C \times v_i + \varepsilon_i$$

and parameters β, θ and C can be consistently estimated for both $D_i = 1$ and $D_i = 0$ by using least squares procedure. More interesting, this model is identified without any exclusion restrictions due to the nonlinearity of the generalized residual term. Also since the generalized residuals are independent with Z_i by construction, there would not be collinearity between $\hat{v}_i$ and Z_i, that is $Corr(\hat{v}_i, Z_i) = 0$. This zero correlation is due to the derivation of the generalized residual as the score for the intercept from the Probit model evaluated at each data point. Gourieroux et. al. (1987, page 15) show how this zero correlation is attained.

5.7 Properties of the estimators

Heckman (1979) mentions that the 2-step estimator although consistent and simpler than Maximum Likelihood, lacks efficiency. He refers to this inefficiency as a consequence of the heteroscedacity apparent from the formula:

$$E\left(V_{1i}^2 \middle| X_{1i}, \lambda_i, v_i \geq -\mathrm{Z}_i \times \gamma\right) = \sigma_u^2 \left[\left(1 - \rho_{\varepsilon v}\right) + \rho_{\varepsilon v}\left(1 + Z_i\lambda_i - \lambda_i^2\right)\right]$$

where $\rho_{\varepsilon v} = Corr(\varepsilon_i, v_i)$. The variance of ε_i never exceeds σ_u^2, the population variance, because the term in braces is never greater than unity. Heteroscedacity is produced since the variance in participation status decrease with increased selection. For example, assuming $\rho_{\varepsilon v} \neq 0$, when people shift from non-participation to participation, the variance in the participants the log outcomes increases while the variance in the non-participants log outcomes decreases. It is suggested that a Generalized Least Squares estimator for the second step account for this heteroscedacity.

Wales and Woodland (1980) present some evidence of the efficiency gain for the Maximum Likelihood against 2-step estimator. Nelson (1984) shows that by defining the parameter of model (5.10) – (5.13) as $\delta = [\beta', C]'$ and the corresponding variables as $A = [X_i, \lambda_i]$, we can take an estimate of δ and its variance as:

$$\hat{\delta} = \left[A'A\right]^{-1} \times A'Y$$
$$V\left(\hat{\delta}\right) = \sigma_{11}^2 \left(A'A\right)^{-1} \times A'WA \times \left(A'A\right)^{-1} \qquad (5.24)$$

As the correlation between X_i and λ_i increases the condition of the matrix being inverted in (5.24) worsens and $V\left(\hat{\delta}\right)$ increases too. This is precisely the situation where the Least Squares bias becomes large and an alternative estimator is needed. Despite all these cautions, virtually all empirical work involving the more complicated of the models with sample selection have been solved using the 2-step estimator instead of the more efficient but computationally difficult ML procedure.

Heckman (1979) also reffers to the standard errors estimation of his conventional model. He claims that the computed standard errors for the 2-step estimator always underestimate the correct asymptotic standard errors and thus they can be used only as lower bounds in statistical inference. Greene (1981) shows the implausibility of that claim by proving that the conventional "incorrect" standard errors can either be larger or smaller than their "correct" counterparts.

Manning, Duan and Rogers (1987) approach the problem from a rather different perspective. They claim that although the inverse Mill's ratio is non linear in the single index $(Z_i\gamma^*)$, the function mapping this index into the inverse Mill's ratio is linear for

certain values of the index. Accordingly, simultaneous exclusion of variables from X_i and inclusion of them in Z_i, in the first estimation step must be implemented. However, this tradeoff of variables is not often possible. Thus many applications constrain $X_i = Z_i$ and β is identified through the non-linearity in the inverse Mill's ratio. Unless exclusion restrictions are invoked, the OLS procedure results to inflated second step standard errors and unreliable estimates of β.

Leung and Yu (1996) do not agree with the above conclusions. Based on a result by Gronau (1974), suggesting that Mill's ratio method breaks down if Z_i is composed of mutually exclusive and exhaustive sets of dummy variables and Z_i is contained in X_i, the authors conduct several Monte Carlo investigations, to find that Heckman 2-step estimator is effective under some circumstances even in the absence of exclusion restrictions. More specific, they suggest that although Mill's ratio is linear over the body of permissible values the single index can take, it becomes non-linear at the extreme values of the index. Hence, even with the absence of exclusion restriction when at least one of the X_i*'s* displays sufficient variation to induce tail behavior in the inverse Mill's ratio, it is likely that the data will posses values of the single index which induce non-linearity and this assists in model identification (plausible standard errors and reliable estimates of β).

5.8 Testing the Normality Assumption

From the above analysis it is obvious that the normality assumption is very crucial for both the ML and the Heckman's 2-step estimators. If normality fails, the estimates of the model (5.10) – (5.13) are inconsistent. This is an unattractive feature and thus it is essential to test the normality assumption in advance. Various tests have been proposed. One of them is that developed by Lee (1984), who, as later Galland and Nychka (1987), assumes that the density of u_i and v_i is not Normal and approximates it as a product of a normal density and a series of Hermite polynomials $H_{rs}(u_i, v_i)$:

$$f_{uv} = \phi_{uv}\left[1 + \sum\sum_{r+s\geq 3} \alpha_{rs} H_{rs}(u_i, v_i)\right]$$

Testing for $\alpha_{rs} = 0$ with the Lagrange multiplier test, one tests for normality because when H_0: $\alpha_{rs} = 0$ cannot be rejected then $f_{uv} = \varphi_{uv}$. However, this test leads to complicated calculations of the scores. For this reason is not applicable in most practical situations.

Chesher and Irish (1987) introduce a second normality test. They specify an extended censored model in which some, or even all, parameters vary independently across realizations and then describe a procedure that derives various diagnostic tests readily applicable to normal based models for censored data.

Under the assumption that:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{vu} & \sigma_v^2 \end{pmatrix}\right]$$

Gourieroux et. al. (1987) derives a normality test statistic that is a simple function of the Generalized residuals

$$\tilde{\hat{u}}_i^0 = \tilde{u}_1\left(\tilde{\theta}_n^0\right), \quad \tilde{\hat{v}}_i^0 = \tilde{v}_1\left(\tilde{\theta}_n^0\right)$$

where $\tilde{u}_i(\theta) = \underset{\theta}{\mathrm{E}}\left[u_i(\theta) \middle/ Y_i\right]$ and $\tilde{v}_i(\theta) = \underset{\theta}{\mathrm{E}}\left[v_i(\theta) \middle/ Y_i\right]$ are the generalized residual and the S-generalized residual, respectively.

Pagan and Vella (1989) discuss the RESET normality test. Similar to Lee (1984), they assume that the density of u_i and v_i is not Normal and approximate it, as Galland and Nychka (1987) do, with:

$$f_{uv} = \left(\sum_{j=0}^{J}\sum_{k=0}^{K} \pi_{jk} u_i^k v_i^j\right) \times \phi_{uv} \tag{5.25}$$

where φ_{uv} is the bivariate normal density of u_i and v_i and $\pi_{00} = 1$. From (5.25) they obtain the conditional expectation of u_i that expresses selection bias:

$$E\left(u_i \middle| v_i\right) = \int u_i \times f_{u|v} du_i = \int u_i \times \left(f_{uv} / f_v\right) du_i = \sum_j \sum_k b^{-1} \pi_{jk} \times \left(\int u^{k+1} \phi_{u|v} du_i\right) \times v_i^j$$

where $f_{u|v}$ is the conditional density of u_i given v_i, $\varphi_{u|v}$ is the conditional normal density and $b = f_v / \varphi_v$. Setting K = 0:

$$E(u_i|v_i) = \sum_j b^{-1}\pi_{oj}\left(\int u_i \phi_{u|v} du_i\right) \times v_i^j$$
$$= \sum_j b^{-1}\pi_{0j} \times \rho \times v^{j+1}$$

and therefore under the null hypothesis of normality:

$$E(u_i|D_i = 1) = E(v_i|D_i = 1) + \pi_{01} \times E(v_i^2|D_i = 1) + ... + \pi_{0J} \times E(u_i^J|D_i = 1)$$

as $b = 1$ when errors are distributed bivariate Normal because $f_v = \varphi_v$.
From the above, testing for $\pi_{0j} = 0, j = 1, ..., J$, is a test for normality. Amemiya (1973) gives expressions for $E(v_i^j|D_i = 1)$, for $j = 1, 2, 3, 4$. By using the recursive formula of Bera, Jarque and Lee (1984), finds that these are proportional to λ_i, $(1 - Z_i\gamma^* \times \lambda_i)$, $\left(2 + (Z_i\gamma^*)^2 \times \lambda_i\right)$ and $\left(3 - 3(Z_i\gamma^*) \times \lambda_i - (Z_i\gamma^*)^3 \times \lambda_i\right)$, respectively. Hence, a test for normality is to add on the variables $(Z_i\gamma^*)^j \times \lambda_i$, $j = 1, 2, 3, 4$ to the 2-step estimator (5.17) and test if they are jointly zero.

Finally, a rather empirical but at least applicable way to test for normality in self-selective samples is to replace Assumption 1 with a weaker distributional assumption that do not impose joint normality for the error terms. Then, one can compare the results and if these are significally different the normality assumption is rejected. Apart from testing purposes, this assumption can be used to formulate a different, independent approach for the estimation of the parameters of a selective sample model without the restriction of the joint normality. In the next paragraph, we will discuss this approach.

5.9 Relaxation of Normality

An alternative, weaker assumption than Assumption 1, can be considered:

Assumption 2

The distribution of v_i is known, different from Normal, and u_i is a linear function of v_i. Again the distribution of u_i and v_i is defined over the entire population.

In the above assumption the normality case has been excluded since it would lead to joint normality of u_i and v_i. When the distribution of v_i is not normal, one cannot proceed with OLS method to estimate λ_i in the first step. Various other methods have been proposed.

The method of Olsen

Olsen (1980) assumes that v_i is uniformly distributed and replaces the inverse Mill's ratio with a simple transformation of the least squares residuals derived from the linear probability model (model that regress D_i with Z_i). Specifically, by assuming that v_i is uniformly distributed in [0, 1], the regression equation of interest becomes:

$$Y_i = X_i\beta + \delta(Z_i\gamma - 1) + \varepsilon_i$$

$$\delta = \rho \times \sigma_u \times \sqrt{3}$$

Since γ is typically not known, one has to estimate it by a linear probability model. Then, the primary model to be estimated is:

$$Y_i = X_i\beta + \delta(Z_i\hat{\gamma} - 1) + n_i$$

$$n_i = -\delta \times Z_i(\hat{\gamma} - \gamma) + \varepsilon_i$$

Olsen (1980) compares this estimator with the classical Heckman's 2-step estimator. He concludes that the two procedures lead to similar results under the same assumptions.

The apparent difference is in the conditions required to identify the effect of sample selection. In Heckman's procedure the same set of regressors may be used in the regression and probit models without encountering perfect collinearity when using the Mill's ratio correction (see Leung and Yu, 1996). On the other hand, Olsen's correction requires the presence of a regressor in the linear probability model, which does not appear in the regression model. In order to identify the effect of sample selection one could include higher order powers and cross products in the linear probability model but exclude them from the regression equation. Although Olsen's method seems practically more general than Heckman's (1979) method, it stills limits the distribution of v_i to Uniform (0, 1). Possibly, an alternative estimator that allowed a more general specification for the v_i distribution would yield better results in many practical cases.

The methods of Lee

A more general approach to relax joint normality is proposed by Lee (1982). He supposes that the marginal distributions of u_i and v_i are specified but their joint distribution is not. The joint bivariate distribution of interest should allow unrestricted correlation between the disturbances u_i and v_i. Assuming that the completely specified marginal distribution of u_i and v_i are respectively *F(u)* and *F(v)*, each of u_i and v_i can be transformed into a standard normal random variable N(0, 1):

$$u_{i*} = J_1(u_i) = \Phi^{-1}\big(F(u_i)\big)$$

and

$$v_{i*} = J_1(v_i) = \Phi^{-1}\big(F(v_i)\big)$$

Then their joint distribution is specified as:

$$H\big(u_i, v_i; \rho\big) = B\big[J_1(u_i), J_2(v_i); \rho\ \big] \qquad (5.18)$$

which is the bivariate normal N(0, 0 , 1, 1, ρ) since u_i and v_i are transformed into standard Normal variables.

Under these assumptions, Lee shows that maximizing the log-likelihood below over β and γ^*:

$$\ln L(\beta, \gamma^*, \rho) = \sum_{i=1}^{N} \{ D_i \ln g((Y_i - X_i\beta)/\sigma_u) + D_i \ln \Phi\left(J_1(-Z_i\gamma^*) - \rho \times J_2((Y_i - X_i\beta)/\sigma_u)/\sqrt{1-\rho^2}\right) - D_i \ln \sigma_u + (1 - D_i) \times ln(1 - F(-Z_i\gamma^*)) \}$$

where g is the density function of u_i, one obtains unbiased and efficient estimates of the parameters of interest.

A more interesting assumption is the one where the distribution of u is normal N(0, 1) and the marginal distribution of v_i is known but non-normal. From the model (5.10)-(5.13) we see that $D_i = 1$ if and only if $v_i \succ -Z_i\gamma$. Given any absolutely continuous distribution function $F(v)$, the transformation $J_1 = \Phi_0^{-1}F(v)$ is a strictly increasing function. Therefore, we have $D_i = 1$ if and only if $J_1(-Z_i\gamma^*) \prec J_1(v)$. The censored regression model with given normal marginal distribution $G(u)$ of u_i, arbitrary marginal distribution $F(v)$ of v_i and the bivariate distribution (5.18) is statistically equivalent to the model with

$$Y_i^{(1)} = X_i\beta + u_i$$
$$D_i'^{(1)} = J_1(Z_i\gamma) + v_i'$$

where $v_i' = J_1(v_i)$ is a standard Normal random variable and (u_i, v_i') are bivariate normally distributed. The regression equation can be written as:

$$Y_{1i} = X_i\beta + C \times \frac{\phi(J_1(-Z_i \times \gamma^*))}{F(Z_i\gamma^*)} + \varepsilon_{1i} \qquad (5.19)$$

An estimate of γ^* can be obtained by the Maximum Likelihood method where one employs *F(v)* as the distribution function for v_i. Then substituting $\hat{\gamma}$ into (5.19) we estimate β and C by Ordinary Least Squares (OLS).

Finally, Lee's (1982, 1983) work is focussed on relaxation of normality by assuming that the error terms (u_i, v_i) are jointly distributed according to the Student – t_d distribution. Since the value of the degrees of freedom, d, affects the tail behavior of the distribution of errors, by varying d, he produces a flexible class of models, which can depart significantly from the "normal" model. As Lydall (1968) said, *wage data tend to be fat tailed due to measurements errors in earnings and hours and because wages are often defined by dividing earnings by hours*. In such cases, a multivariate Student-t_d distribution with the appropriate degrees of freedom is possibly a better approximation of the error distribution than Normal.

To see mathematically how the selection bias problem is solved in this case, let us assume that $t_d(\mu, m)$ denotes the multivariate Student – t_d distribution with mean μ, scale matrix m and variance equal to $[d/(d-2)] \times m^{-1}$, where d are the degrees of freedom[2]. Then t_d is the standardized univariate Student – t_d density with mean 0 and scale parameter equal to 1 with T_d the associated cumulative distribution function. The marginal distribution of v_i is *F(v)*. In order to transform it into a t_d density, define $J_{T_d}(v) \equiv T_d^{-1}(F(v))$. Then, the regression equation, corrected for the selection bias is:

$$E\left(Y_i \middle| D_i = 1, X_i, Z_i\right) = X_i\beta + C \times \left[\left(\frac{d + \left(J_{T_d}\left(-Z_i\gamma^*\right)^2\right)}{d-1}\right) \times \left(\frac{t_d\left(J_{T_d}\left(-Z_i\gamma^*\right)\right)}{F\left(Z_i\gamma^*\right)}\right)\right]$$

Alternatively, if we set

$$g\left(v_i, d\right) \equiv \left(\frac{d + \left[J_{T_d}\left(v_i\right)\right]}{d-1}\right) \times t_d\left(J_{T_d}\left(v_i\right)\right)$$

[2] The mean exists for d > 1 and the variance exists for d > 2.

the regression equation takes the simpler form:

$$E\left(Y_i \middle| D_i = 1, X_i, Z_i\right) = \quad X_i\beta + C \times \frac{g\left(v_i, d\right)}{F\left(Z_i\gamma^*\right)}$$

5.10 Instrumental Variable Estimation (IV)

A rather new approach, developed for program evaluation applications, is the method of Instrumental Variables (IV). Although this method is considered from Heckman, LaLonde and Smith (1999) as a variant of the matching method because of the assumptions it poses for the identification of the various parameters, it is mentioned here because it explicitly makes use of structural equation models.

Instrumental variables are those variables excluded from some equations and included in others, and therefore are correlated with some outcomes only through their effect on other variables. To make it clear, let us suppose that in a social program, persons sort into the program on the basis of an unobserved factor, e.g. ability. "Ability" may raise earnings and more able people participate, but participation may not raise the earnings of any given person. To evaluate such a program, an instrument $Q_i \in Z_i$ is often sought that determines participation but that does not directly affect earnings and does not depend on "ability". Angrist, Imbens and Rubin (1996) and Rosebaum (1996) provide a simple example of a relative social program.

5.10.1 Definition of IV Estimator

Take the dummy endogenous variable model of Heckman (1978). The Instrumental Variable method is focused on the estimation of the effect of participation status D_i on outcomes. By means of model (5.20) – (5.22), the parameter of interest is θ. Following Durbin (1954), Imbens and Angrist (1994) set the theoretical background for the existence of instruments and the plausibility of the IV estimator. Specifically, they pose the following conditions:

Condition 5.1 (Existence of Instruments):

Let Q_i be a random variable with $\mathfrak{I}$ be the support of Q. Define for each $q_i \in \mathfrak{I}$ a random variable $D(Q_i)$ with $D(q_i) = 1$ if an individual with instrument (characteristics) $Q_i = q_i$ participates and zero otherwise. Now define a random variable W such that (i) for all $w_i \in \mathfrak{I}$ the triple $(Y_0, Y_1, D(W_i))$ is jointly independent of Q_i and (ii)$P(W)=E(D_i \mid Q_i=w_i)$ is a non-trivial function of W.

Condition 5.2 (Monotonicity):

For all q_i, $w_i \in \mathfrak{I}$, either $D(q_i) \geq D(w_i)$ or $D(q_i) \leq D(w_i)$ for every individual. This condition is inserted in order to obtain adequate estimates of the IV estimator.

Condition 5.3

g(Q) is a function from the support of Q to $\mathfrak{R}$, such that (i) either for all q_i, $w_i \in \mathfrak{I}$, $P(q_i) \leq P(w_i)$ implies $g(q_i) \leq g(w_i)$, or, for all q_i, $w_i \in \mathfrak{I}$, $P(q_i) \leq P(w_i)$ implies $g(q_i) \geq g(w_i)$; (ii) $Cov(d_i, g(q_i)) \neq 0$.

Theorem 5.1 (Definition of the IV estimator)

Suppose that Conditions 1, 2 and 3 are satisfied. Let Q be a discrete random variable with support $\{Q_{0i}, Q_{1i}, \ldots, Q_{Ki}\}$, ordered in such a way that if $l < m$ then $P(Q_l) \leq P(Q_m)$. Then, if $Cov(D_i, g(Q_i)) \neq 0$, the IV estimator for the effect of D_i on Y_i using $g(Q_i)$ as an instrument estimates

$$\hat{\theta}_{IV} = \frac{Cov(Y_i, Q_i)}{Cov(D_i, Q_i)} = \frac{\sum_{i=1}^{N} Y_i Q_i \Big/ \sum_{i=1}^{N} Q_i - \sum_{i=1}^{N} Y_i (1-Q_i) \Big/ \sum_{i=1}^{N} (1-Q_i)}{\sum_{\iota=1}^{N} D_i Q_i \Big/ \sum_{i=1}^{N} Q_i - \sum_{\iota=1}^{N} D_i (1-Q_i) \Big/ \sum_{i=1}^{N} (1-Q_i)} \qquad (5.26)$$

As shown, the IV estimator is defined as the ratio of sample covariances.

The important contribution of Imbens and Angrist (1994) on the IV procedure is the definition of a mean parameter based on the IV framework. More specifically, under the above conditions and on the following:

Assumption 5.1 (Stable Unit Treatment Value Assumption or SUTVA):

This assumption is defined by Rubin (1980) in terms of an IV framework. SUTVA implies that the potential outcomes are unrelated to the treatment status of other individuals. More formally, it assumes that:

a) *If $Q_i = Q_i'$, then the causal effect of instruments on D_i for person i is $D_i(Q_i) = D_i(Q')$*

b) *If $Q_i = Q_i'$ and $D_i = D_i'$, then the causal effect of instruments on Y_i for person i is $Y_i(Q_i, D_i) = Y_i(Q_i', D_i)$*

Assumption 5.2 (Random Assignment)

The treatment assignment Q_i is random: $P(Q_i = c) = P(Q_i = c')$

Assumption 5.3 (Exclusion Restriction)

$Y(Q_i, D_i) = Y(Q_i', D_i)$ for all q_i, q_i' and for all d_i

Assumption 5.4 (Nonzero Average Causal Effect of Q on D)

The average causal effect of Q_i on D_i, $E(D_i(1) - D_i(0)) \neq 0$

Then, the Local Average Treatment Effect (LATE) is defined as:

$$E\left(Y_i \middle| Q_i = q_i\right) - E\left(Y_i \middle| Q_i = w_i\right) = \left[P\left(q_i\right) - P\left(w_i\right)\right] \times E\left[Y_i\left(1\right) - Y_i\left(0\right) \middle| D_i\left(q_i\right) - D_i\left(w_i\right) = 1\right]$$

LATE is defined as the expected outcome gain for those induced to receive treatment through a change in the instrument from $Q_i = q_i$ to $Q_i = w_i$. The variable Q_i affects the treatment decision, since is contained in Q_i in the selection equation, but does not affect directly the outcome Y_i.

More schematically, Angrist, Imbens and Rubin (1996) show that LATE estimates the average causal effect of a specific subpopulation of persons. To see this, let us suppose two different choices. We define an unobserved variable (IV) that affect choices Q_i. When $Q_i = 1$, the individuals are induced to participate while $Q_i = 0$ denotes inexistence of any induction. An individual may, under either value of Q_i, participate in a program and thus $(D(Q_i) = 1)$ or may not and $(D(Q_i) = 0)$. The following table is enlightening:

Table 5.2: Causal effects of Q_i on Y_i, $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$, for the population of units classified by $D_i(0)$ and $D_i(1)$

		$D_i(0)$	
		0	*1*
$D_I(1)$ / $D_i(1)$	*0*	$Y_I(1, 0) – Y_i(0, 0) = 0$ Never – takers	$Y_i(1, 0) – Y_i(0, 1) = -[Y_I(1) – Y_i(0)]$ Defiers
	1	$Y_i(1, 1) – Y_i(0, 0) = Y_i(1) – Y_i(0)$ Compliers	$Y_i(1, 1) – Y_i(0, 1) = 0$ Always-takers

According to the authors, LATE parameter is designed to estimate the average effect in outcomes of compliers. These individuals behave more normally that the others. They who encouraged to participate, do participate and they who did not encouraged, do not participate. Thus, LATE parameter can be formulated as:

$$LATE = E\left[Y_i(1) - Y_i(0) \middle| D_i(1) - D_i(0) = 1\right] = \frac{E\left[Y_i\left(D_i(1),1\right) - Y_i\left(D_i(0),0\right)\right]}{E\left[D_i(1) - D_i(0)\right]}$$

Monotonicity condition and Nonzero Average Causal Effect of Q_i on D_i assumption restrict the definition of LATE to "compliers" only (*key assumption*). In this formulation, LATE takes into account the outcomes of the normally behave persons. In other words, it restricts evaluation in a specific, homogeneous group of people, the "compliers", embedded within the *Rubin's Causal Model* (see Holland, 1986). The advantages of embedding the IV approach in the *Rubin's Causal Model* are that it clarifies the nature of critical assumptions needed for a causal interpretation, and moreover allows the analyst to consider sensitivity of the results to deviations from the "key assumption" in a straightforward manner.

It is obvious that all other individuals, apart from compliers, do not behave normally. A simple examination of the above table shows that "Always Takers" and "Never Takers" have $D_i(1) – D_i(0) = 0$ and Assumption 5.4 is violated while "Defiers" have $D_i(1)–D_i(0) = -1$ and monotonicity condition is violated because they seem to participate without induction and do not participate with inducion. For these reasons, Angrist, Imbens and Rubin (1996) support the plausibility of their LATE parameter.

Regardless of the above arguments, it is not clear how LATE parameter accounts for selectivity bias in a non-experimental framework since it is simply expressed as a simple mean average. In fact, this seems to be a major limitation of this estimator that makes it applicable only under experimental designs where selectivity is balanced rather than eliminated through randomization. To that extend, Heckman (1996) reconsiders LATE from a different perspective to control for selection bias.

5.10.2 Criticisms

The above parameter has been subject of criticisms by several authors. Its assumptions and the fact that is restricted in the estimation of causal effects for only a subpopulation of persons are the main arguments against LATE. More specifically, Robins and Greenland (1996) comment that in many cases ATE parameter can be of greater public interest and provide a specific example from a clinical trial where almost all individuals are willing to take treatment independent of whether or not they induced to. In this case ATE estimates the average treatment effect for persons who actually receive treatment compared to those who do not, while LATE does not estimate adequately the average effect of treatment since it refers only to a specifically behave subpopulation. However, the authors recognize the superiority of LATE parameter in several cases because it takes into account the choice variable D_i in the analysis. Angrist, Imbens and Rubin (1996, rejoinder) agree with Robins and Greenland that in some cases ATE provides more interesting results than LATE. They also add that IV framework allows them to compute informative bounds for the ATE parameter.

Heckman (1996) has also studied the restriction of LATE to compliers only. He finds LATE conditions unattractive once they clearly stated and that are based on unspecified and implicit behavioral assumptions. Robins and Greenland (1996) and Heckman (1996) stress that LATE is an average causal effect for a subpopulation that cannot be identified in the sense that we cannot label all individual units in the population as compliers and non-compliers. The latter author also disagrees in the comment of Angrist, Imbens and Rubin that econometricians do not clearly state their behavioral assumptions and adds that the literature includes many models to test the assumptions of LATE that are ignored by them.

Heckman (1996) considers weaker conditions and more general behavioral assumptions to identify the Effect of treatment on the treated within an instrumental variable framework. Neither the monotonicity nor the full independence condition (implied by Exclusion Restriction assumption) needed for this identification. He imposes the following Mean Independence conditions:

$$E(u_0|X_i,Q_i)=0$$
$$E(\Delta_i|X_i,Q_i,D_i=1)=E(\Delta_i|X_i,D_i=1) \qquad (5.26)$$

and restates the condition that both Q_i and X_i determine D_i as an assumption

$$P(D_i=1|X_i,Q_i)\neq P(D_i=1|X_i)$$

where $P(D_i=1|X_i,Q_i)$ is a nontrivial function of Q_i. Then, by formulating the observed outcome for an individual as:

$$Y_i=D_iY_{1i}+(1-D_i)Y_{0i}=Y_{0i}+D_i(Y_{1i}-Y_{0i})$$

one obtains the mean observed outcome

$$\begin{aligned}E(Y_i|X_i,Q_i)&=E(Y_{0i}|X_i,Q_i)+E(\Delta_i|X_i,Q_i,D_i=1)\times P(D_i=1|X_i,Q_i)\\&=E(Y_{0i}|X_i)+E(\Delta_i|X_i,D_i=1)\times P(D_i=1|X_i,Q_i)\end{aligned}$$

by equation (5.26). By assuming two different instruments Q_i' and Q_i'', Heckman obtains a LATE parameter, not restricted to compliers

$$\begin{aligned}&E(\Delta_i|X_i,D_i=1)=E(Y_i|X_i,Q_i',D_i=1)-E(Y_i|X_i,Q_i'',D_i=1)\\&=E(Y_{0i}|X_i)+E(\Delta_i|X_i,D_i=1)\times P(D_i=1|X_i,Q_i')-E(Y_{0i}|X_i)+E(\Delta_i|X_i,D_i=1)\times P(D_i=1|X_i,Q_i'')\end{aligned}$$

Thus,

$$E\left(\Delta_i \middle| X_i, D_i = 1\right) = \frac{E\left(Y_i \middle| X_i, Q_i', D_i = 1\right) - E\left(Y_i \middle| X_i, Q_i'', D_i = 1\right)}{P\left(D_i = 1 \middle| X_i, Q_i'\right) - P\left(D_i = 1 \middle| X_i, Q_i''\right)} \tag{5.27}$$

This parameter evaluates the impact of the program for participants through a change in the instrument while accounting for selection bias, as shown by Heckman, Tobias and Vytlacil (2000). Angrist, Imbens and Rubin (1996, rejoinder), however, consider as implausible Heckman's expression of LATE. They claim that Heckman's key assumptions (mean independence) compare average outcomes for groups of individuals that the analyst does not know how they behave when faced up with an instrument Q_i. Thus, these groups are constituted by completely different people and LATE parameter may give implausible estimates. In addition, it is argued that the weaker assumptions imposed by Heckman (1996) (specifically the mean independence assumption instead of the full independence of Angrist, Imbens and Rubin) lack in scientific (economic) content. Since mean independence is weaker than full independence, it is preferred to the latter in many situations. However, in the case of instrumental variables, when mean independence hold, Q_i would be a valid instrument for the effect of D_i on Y_i but not on a transformation of Y_i such as $log(Y_i)$. Because of such cases, the stronger assumption has to be preferred. Finally, it must be noted that Heckman's pessimistic view of IV methods can be contrasted with the development of his views on a class of experimental evaluation designs with randomized eligibility. In these designs, units are randomly assigned an instrument Q_i with $Q_i = 1$ implying that unit i is eligible for a particular treatment and $Q_i = 0$ implying that unit i is not eligible to receive treatment. Formally, this is a special case of LATE model with $D_i(0) = 0$ (no defiers or always-takers), and hence monotonicity is automatically satisfied.

Moffit (1996) regards IV method as the most versatile and flexible technique, applicable in an enormous number of disparate applications. However, he criticizes the "compliers restriction" of LATE and comments that this parameter is based on untestable behavioral assumptions. Rosenbaum (1996) also shows the implausibility of this restriction in many practical situations by mentioning a clinical experiment for persons that suffer from Chronic Obstructive Lung Disease (COLD). At this example, it is

obvious that not all persons that were induced to participate, they did so. Since even in this small-size experiment the conditions of AIR were not satisfied, at a larger one, like a training program, they would not satisfied too. In addition, it would be very difficult to recognize individually the persons that are actually compliers. Possibly, such an attempt would result in large errors.

However, Rosenbaum (1996) agree with AIR in the fact that IV method possesses a significant role in many applications. In his example, when the mean effect of treatment on the treated (TT) is estimated according to the people that participated, the analyst results in overstated estimates. That is actually true because there may be people with specific characteristics e.g. healthier that surely participate whether induced or not and as healthiers they obtain greater values of outcomes. If the mean effect on outcomes were estimated by the means of the inducement to participate instead of the participation status itself, the estimate would be more sensible.

Rosenbaum's (1996) final comment is concerned with the lack of robustness of means as evaluation measures. He suggests using more robust estimators such as the Hodges-Lehman estimator (see Appendix 1 for a description) that performs median or quantile regression. Angrist, Imbens and Rubin (1996, rejoinder) consider this suggestion as an attractive one. They state that median and mean regressions are applied equally well to IV problems. However, little attention has been given in such an approach in econometrics and further attention is deserved.

5.10.3 The Local Instrumental Variable Effect

Heckman (1997) developed an additional mean parameter based on the IV approach. He called this parameter Local Instrumental Variable (LIV). Heckman and Smith (1998) and Heckman and Vytlacil (1999, 2000a) refer LIV as an estimator of the average treatment effect for individuals with a given value of v_i:

$$\begin{aligned} LIV(X) &= E\left(\Delta_i \middle| X_i, v_i\right) \\ &= X_i\left(\beta_1 - \beta_0\right) + E\left(u_{1i} - u_{0i} \middle| X_i, v_i\right) \\ &= X_i\left(\beta_1 - \beta_0\right) - E\left(u_{1i} - u_{0i} \middle| v_i\right) \end{aligned} \tag{5.28}$$

Evaluation of the LIV parameter at low values of v_i averages the outcome gain for those with unobservables making them least likely to participate, while evaluation of the LIV parameter at high values of v_i is the gain for those individuals with unobservables which make them most likely to participate. Since X_i is independent of v_i, the LIV parameter can be written as:

$$LIV(X) = E\left(\Delta_i \middle| X_i, Z_i, P(Q_i)\right) = \frac{\partial E\left(Y_i \middle| X_i, Z_i, P(Q_i)\right)}{\partial P(Q_i)} \qquad (5.29)$$

This formula gives the average effect for people who are just indifferent between participation or not at the given value of the instrument (i.e. for people who are indifferent at $P(Q_i) = P_i$). LIV$(X_i, P(Q_i))$ for values of P_i close to zero is the average effect for the individuals with unobservable characteristics that make them the most inclined to participate, and LIV$(X_i, P(Q_i))$ for values of P_i close to one is the average treatment effect for the individuals with unobservable characteristics that make them the least inclined to participate.

5.11 Calculation of Mean Parameters Using Heckman's 2-step Procedure

Estimation of mean gains and program impacts is a central feature in evaluation studies. Both statisticians and economists are focussed on this aim, although they approach it from different perspectives. While the former rely on randomized experiments or matching methods, economists rely on structural procedures based on Heckman's 2-step method to estimate parametrically unbiased mean parameters such as TT or ATE. Economists' approach is going to be reviewed in this paragraph.

Intuitively, ATE, TT, LATE and LIV can be estimated only when the outcomes for both the participants and non-participants are observed. Thus, in the formulation of the selection model (5.10) – (5.13), we discriminate earnings of participants from those of non-participants and consider the following structural models:

$$Y_{1i} = X_{1i}\beta_1 + u_{1i}$$
$$Y_{0i} = X_{0i}\beta_0 + u_{0i}$$
$$D_i = Z_i\gamma + v_i$$

The Normality assumption (Assumption 1) is then equivalently presented as:

$$\begin{bmatrix} u_{1i} \\ u_{0i} \\ v_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{u_1}^2 & \sigma_{u_1u_0} & \sigma_{u_1v} \\ \sigma_{u_0u_1} & \sigma_{u_0}^2 & \sigma_{u_0v} \\ \sigma_{vu_1} & \sigma_{vu_0} & \sigma_v^2 \end{bmatrix}\right)$$

where $\sigma_{u_1u_0} = \sigma_{u_0u_1}, \sigma_{u_1v} = \sigma_{vu_1}$ *and* $\sigma_{u_0v} = \sigma_{vu_0}$.

The Average Treatment Effect

It has already been mentioned that the Average Treatment Effect (ATE) does not answer economically interesting questions. Its simple form does not require formulating Heckman's estimator in order to be represented. The Average Treatment Effect conditional on $X_i = x_i$ is simply expressed as:

$$ATE(X) = E(\Delta_i | X_i) = X_i(\beta_1 - \beta_0) \qquad (5.30)$$

Heckman's 2-step method is primary oriented to the estimation of the rest mean impacts. At this paragraph we discuss how to obtain adequate estimates of the mean Effect of Treatment on the Treated, the Local Average Treatment Effect and the Local Instrumental Variable parameter. Both cases of normal and non-normal distribution for the error terms (u_i, v_i) are discussed.

The Effect of Treatment on the Treated (TT)

Under the normality assumption and the formulation of Heckman's 2-step estimator, the expression of the mean effect of Treatment on the Treated is:

$$\begin{aligned} TT(X) &= E\left(\Delta_i \middle| X_i, Z_i, D_i = 1\right) \\ &= X_i\left(\beta_1 - \beta_0\right) + \left(\rho_1 \times \sigma_{u_1} - \rho_0 \times \sigma_{u_0}\right) \times \frac{\phi\left(-Z_i\gamma^*\right)}{\Phi\left(Z_i\gamma^*\right)} \end{aligned} \tag{5.31}$$

where $\rho_D = Corr(u_{iD}, v_i), D = 0,1$.

Adopting Lee's (1982) assumptions of $u_i \sim N(0,1)$ and a known, non-Normal distribution for v_i, TT estimator is expressed as:

$$\begin{aligned} TT(X) &= E\left(Y_i \middle| X_i, Z_i, D_i = 1\right) \\ &= X_i\left(\beta_1 - \beta_0\right) + \left(\rho_1 \times \sigma_{u_1} - \rho_0 \times \sigma_{u_0}\right) \times \frac{\phi\left[J_2\left(-Z_i\gamma^*\right)\right]}{F\left(Z_i\gamma^*\right)} \end{aligned} \tag{5.32}$$

In the same way, by assuming that (u_i, v_i) are jointly Student – t_d distributed, TT is expressed as in equation (5.19).

The Local Average Treatment Effect (LATE)

In an econometric setting, this parameter can be evaluated by Heckman's (1979) 2-step procedure that accounts for selectivity bias. The mean effects under normality and non-normality assumptions, respectively, are:

$$\begin{aligned} LATE(X) &= E\left(\Delta_i \middle| X_i, P(Q_i'), P(Q_i'')\right) = E\left(\Delta_i \middle| X_i, -Q_i''\gamma \prec u_i \prec -Q_i'\gamma\right) \\ &= X_i\left(\beta_1 - \beta_0\right) + \left(\rho_1 \times \sigma_{u_1} - \rho_0 \times \sigma_{u_0}\right) \times \frac{\phi\left(-Q_i'\gamma^*\right) - \phi\left(-Q_i''\gamma^*\right)}{\Phi\left(Q_i'\gamma^*\right) - \Phi\left(Q_i''\gamma^*\right)} \end{aligned} \tag{5.32}$$

and

$$LATE(X) = E\left(\Delta_i \middle| X_i, P(Q_i'), P(Q_i')\right) = E\left(\Delta_i \middle| X_i, -Q_i'\gamma \prec u_i \prec -Q_i'\gamma\right)$$

$$= X_i(\beta_1 - \beta_0) + \left(\rho_1 \times \sigma_{u_1} - \rho_0 \times \sigma_{u_0}\right) \times \frac{\phi\left[J_2\left(-Q_i'\gamma^*\right)\right] - \phi\left[J_2\left(-Q_i'\gamma^*\right)\right]}{F\left(Q_i'\gamma^*\right) - F\left(Q_i'\gamma^*\right)} \quad (5.33)$$

The Local Instrumental Variable parameter (LIV)

The corresponding LIV parameters are calculated as:

$$\begin{aligned} LIV(X) &= E\left(\Delta_i \middle| X_i, V_i = v_i\right) \\ &= X_i(\beta_1 - \beta_0) + E\left(u_{1i} - u_{0i} \middle| V_i = v_i\right) \\ &= X_i(\beta_1 - \beta_0) + E\left(u_{1i} \middle| V_i = v_i\right) - E\left(u_{0i} \middle| V_i = v_i\right) \\ &= X_i(\beta_1 - \beta_0) + \left(\rho_1 \times \sigma_{u_1} - \rho_0 \times \sigma_{u_0}\right) \times v_i \end{aligned} \quad (5.34)$$

and

$$\begin{aligned} LIV(X) &= E\left(\Delta_i \middle| X_i, V_i = v_i\right) \\ &= X_i(\beta_1 - \beta_0) + E\left(u_{1i} - u_{0i} \middle| V_i = v_i\right) \\ &= X_i(\beta_1 - \beta_0) + E\left(u_{1i} \middle| V_i = v_i\right) - E\left(u_{0i} \middle| V_i = v_i\right) \\ &= X_i(\beta_1 - \beta_0) + \left(\rho_1 \times \sigma_{u_1} - \rho_0 \times \sigma_{u_0}\right) \times J_2(v_i) \end{aligned} \quad (5.35)$$

Heckman and Vytlacil (1999, 2000a) explain the relationships between these parameters. Here it is interesting to note that the most general mean effect seems to be LIV and all other parameters are average values of LIV but for values of v_i lying in different intervals.

Heckman, Tobias and Vytlacil (2000) assess the performance of Heckman's 2-step estimator and compute the mean parameters discussed here under correct and incorrect model specification. They obtain sampling distributions of the estimators of different treatment parameters using both generated normal and non-normal data from Monte

Carlo simulations. They show that different kinds of data, give different estimates of the various treatment effects. Then they provide a simple model selection procedure given equal number of parameters across the various models. The procedure dictates to obtain estimates of the selection-corrected conditional mean functions for a variety of competing models and then select the one that minimizes the Sum Squared Residuals (SSR). This approach to model selection chooses the model whose conditional mean function provides the best fit to the observed data (see also Amemiya, 1980). Formally, one chooses the model, which minimizes the criterion:

$$\sum_{i=1}^{N}\left[\left(Y_i - D_i m\left(X_i, Z_i \middle| D_i = 1\right) - \left(1 - D_i\right) m\left(X_i, Z_i \middle| D_i = 0\right)\right)\right]^2$$

where $m(X_i, Z_i \,|D_i = 1)$ is the estimated selection conditional mean function in the participation state and $m(X_i, Z_i \,|D_i = 0)$ is the corresponding estimate of the non-participation state. The results of the simulations showed that the performance of the proposed model selection improves with the sample size n and with the degree of selectivity in the model. Earlier, Heckman (1984) had already developed a X^2 Goodness of Fit test for model specification.

It is worth noting here that the present econometric literature does not test the equality of the above mean differences. The analysis includes only tests for the parameters included in each mean estimate of Y_i. Intuitively, under Assumption 1, a t-test for the equality of the selection-corrected means in each case is a plausible test statistic. However, either when Assumption 1 is invalid or it is doubtful whether selection bias has been eliminated, a t-test cannot be applied. Instead, a numerical, non-parametric method (e.g. *bootstrap*) would yield a more plausible test for the equality of the means.

5.12 Calculation of ATE Using the Method of Bounding

This method is considered by Manski (1989) to estimate an informative bound on $E\left(Y_i \middle| X_i\right)$. The basic assumption invoked is that, conditional on X_i and on $D_i = 0$, the

distribution of Y_i is concentrated in a given interval $\{K_{0X}, K_{1X}\}$, where $K_{0X} \leq K_{1X}$. In other words,

$$P\{Y_i \in [K_{0X}, K_{1X}] | X_i, Z_i, D_i = 0\} = 1 \Rightarrow K_{0X} \leq E(Y_i | X_i, Z_i, D_i = 0) \leq K_{1X} \quad (5.36)$$

Taking the switching regression model (2.3), the mean outcome Y_i of an observed person is:

$$E(Y_i | X_i) = E(Y_i | D_i = 1, Z_i, X_i) \times P(D_i = 1) + E(Y_i | D_i = 0, Z_i, X_i) \times P(D_i = 0) \quad (5.37)$$

Applying inequality (5.36) to (5.37) results in:

$$E(Y_i | D_i = 1, X_i, Z_i) \times P(D_i = 1 | Z_i) + K_{0X} \times P(D_i = 0 | Z_i) \leq E(Y_i | X_i)$$
$$\leq E(Y_i | D_i = 1, X_i, Z_i) \times P(D_i = 1 | Z_i) + K_{1X} \times P(D_i = 0 | Z_i)$$

Thus the lower bound is the value that takes $E(Y_i | X_i)$ if in the non-selected subpopulation, Y_i always equals K_{0X}. The upper bound is the value of $E(Y_i | X_i)$ if all the non-selected Y_i 's equal K_{1X}.

The method of bounding has not been considered extensively in the literature. Timing may have played a significant role. In the early 1970's when this method was developed non-parametric regression analysis was just beginning to be formalized by statisticians. Preoccupation of researchers with the estimation of wage equations is possibly another reason since $logY_i$ has no obvious upper bound, although minimum wage legislation may enforce a lower bound. Finally, the important limitations of this method in the estimation of the bound have definitely played an important role. Manski (1989) refers to these limitations and also considers a version of the curse of dimensionality for the bounding estimator.

5.13 Semi-Parametric Methods

5.13.1 Maximum Likelihood Estimation

The perceived inconsistency of the Heckman's 2-step estimator indicated Maximum Likelihood method as the usual alternative in parameter estimation of econometric applications. The sensitiveness of this method in small deviations from Assumption 1 dictates relaxation of the (u_i, v_i) joint normality assumption.

The parametric methods on this subject have been described in the previous paragraph. They avoid the imposition of joint normality by requiring a known marginal distribution for the choice equation's disturbances. At this point we present the semi-parametric approach to the problem. The strategy focuses to avoidance of the distributional assumption for the disturbances by approximating their density function.

Gallant and Nychka's estimator

Based on *ERA model* of Phillips (1983), Galland and Nychka (1987) employ an estimation strategy to approximate the true joint density for the disturbances, $f(u_i,v_i)$. Using the selection bias model (5.10) – (5.13), the authors do not assume that the primary equation is linear. They proposed to take $f(u_i, v_i)$ to be of the parametric form:

$$\widetilde{f}(u_i, v_i) = \left[\sum_{j=1}^{M} \sum_{k=1}^{M} \pi_{jk} \times u_i^j \times v_i^k \right]^2 \times \exp\left\{ -(u_i/\delta_1)^2 - (v_i/\delta_2)^2 \right\}$$

where π_{ij} denotes unknown parameters of a Hermite polynomial and δ_1 and δ_2 are parameters of interest. Integration with respect to this form of density gives us:

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \widetilde{f}(u_i, v_i) = \sum_{j,k,p,q=1}^{M} \pi_{jk} \times \pi_{pq} \times \int_{a_1}^{b_1} u_i^{j+p} \times \exp\left(-u_i^2/\delta_1^2\right) \times \int_{a_2}^{b_2} v_i^{k+q} \times \exp\left(-v_i^2/\delta_2^2\right)$$

By using this formula one can approximate the log-likelihood of equation (5.14).

Vella (1998) indicates that the multiplication of a suitably chosen polynomial with the normal densities approximates well the true disturbance density. However, if another density, apart from normal, is more plausible a priori, it may be substituted for the normal. The approximation is improved for large values of K and J. Although Galand and Nychka (1987) provide consistency results for their procedure, they do not provide an analytical distributional theory.

Since the estimate of $f(u_i, v_i)$ must represent a density, some restrictions have to be imposed on the chosen Hermite expansion and on values of π_{ij}. For this reason the Hermite series is in the form of a squared polynomial and the coefficients π_{ij} are restricted so that the series integrate to one and has zero mean.

The literature does not include many applications of the Maximum Likelihood method in a semi-parametric framework. The implementation difficulties are possibly the main reason for this scarcity. However, Melenberg and Van Soest (1993) use this method to examine the determinants of the wages of married women, while accounting for market work decision, using data from Netherlands. Since jointly normality for the disturbances was not a satisfactory approximation for such data because of the tail behavior, the authors do not assume a specific distribution for (u_i, v_i) and approach the problem by the semi-parametric method of Galand and Nychka (1987).

5.13.2 Heckman's 2-step Estimation

As shown, econometric and sample selectivity models have found interesting applications in empirical studies. Apart from ML method, Heckman's 2-step procedure has been applied extensively on evaluation of social programs. In the simplest case this estimator leads to adequate results by assuming that the errors are jointly normally distributed. Models with parametric distributions however, may be subject to distributional misspecifications, which might result in inconsistent estimates. Despite the fact that these assumptions can be relaxed by invoking weaker assumptions for the parametric form of the joint distribution of errors, recent research efforts focuses in an alternative approach.

Semi-parametric methods have been proposed for the estimation of sample selection and self-selection models with discrete choice selection rules, when the analyst cannot

specify the distribution of errors. These methods are based on the replacement of Assumption 2 with a weaker statement for the disturbances:

Assumption 3 (Index Restriction)
$E(u_i|Z_i, D_i = 1) = g(v_i)$ where g is an unknown function.

Estimation of the primary model under Assumption 3 raises two difficulties. First, it is no longer possible to invoke distributional assumptions regarding v_i to estimate γ. Therefore, Ordinary Least Squares method is not longer an option. Second, one cannot use a control function based on distributional relationships of u_i and v_i to estimate $E(u_i|D_i, X_i, Z_i)$, as in Assumption 2. These two problems are confronted separately with the semi-parametric estimation strategies described in the following subparagraph.

5.13.2.1 Estimation of the Selection Model

Following McFadden (1973, 1975) and Manski (1975), when the binary choice probability model is derived from a random utility maximization model, the choice probability for one alternative has the form $F[V(Z_i\gamma)]$, where *V()* is a utility function, γ is a $J \times 1$ vector of parameter representing the systematic component of the utility, Z_i is a $N \times J$ vector of exogenous variables and *F[·]* indicates the cumulative distribution function of the random component of the utility. Based on this specification, Cosslett (1983) describes a method of estimating γ by maximizing the likelihood over γ and a space that contains all distribution functions without assuming any functional form for the distribution *F*. He proves that $\hat{\gamma}$ and $\hat{F}$ are consistently estimated. To this extend Newey (1990) provides a semi-parametric efficiency bound for the binary choice (selection) model.

Matzkin (1992) indicates the possibility to identify binary choice models or binary *threshold choice models* (see Georgescu-Roegen, 1958; Lioukas, 1984) without imposing any parametric structure either on the systematic function of observable exogenous variables or on the distribution of the random term. This identification result is employed to develop a fully non-parametric ML estimator for both the functions of observable

exogenous variables and the distribution of the random term. The estimator is known to be strongly consistent and a 2-step procedure for its calculation is developed.

Ahn and Powel (1993) consider a non-parametric regression estimator for the selection equation under the assumption that the error term v_i is continuously distributed with support on the entire real line and is independent of Z_i. Similar to Heckman (1979), the authors estimate the primary equation in a second step, described in the following paragraph.

Itchimura (1993) constructs a semi-parametric Least Squares (SLS) estimator and a weighted SLS (WSLS) estimator of coefficients up to a multiplicative constant for the selection equation. Both estimators exhibit $1/\sqrt{N}$ -consistency and asymptotic normality. A consistent estimator of the covariance matrix is also presented. Since SLS estimation does not require specifying a parametric error distribution, the method allows analysts to focus on specifying systematic effects of an econometric model and frees them from the distributional worries for a broader class of models.

Finally, Klein and Spady (1993) propose an estimator for the selection equation that does not make any assumptions concerning the functional form of the choice probability function. The estimator is shown to be consistent, asymptotically Normal and to achieve the semi-parametric efficiency bound of Newey (1990). Klein and Spady (1993) also consider a generalization of their estimator for the cases of *trinary* choice problems and ordered selection rules.

5.13.2.2 Estimation of the Primary Equation

The major difficulty on the estimation of the selection bias model (5.10) – (5.13) is the fact that one cannot use distributional relationships to estimate $E\left(u_i \middle| X_i, Z_i, D_i = 1\right)$. Without evaluating the error term, the parameters of the regression model cannot be estimated adequately due to selection bias problem. Many authors have studied ways to overcome this difficulty. All start by considering Assumption 3 and define the conditional expectation of the primary equation as:

$$E\left(Y_i \middle| X_i, Z_i, D_i = 1\right) = X_i\beta + g\left(v_i\right); \qquad \text{i} = 1,...,\text{N} \qquad (5.40)$$

In this relationship it is possible to distinguish an intercept term in X_i from an intercept in *g(.)*. For simplicity reasons here we discuss the issue of approximation of function *g(.)*. In a following paragraph, we explore some ways to infer about the value of the intercept.

Heckman and Robb's estimator

The first suggestion to estimate model (5.40) semi-parametrically is found in Heckman and Robb (1985a) who propose a 2-step estimator. In the first step, the parameter γ and the propensity score $P(D_i = 1|Z_i)$ are estimated non-parametrically. In the second step, the function $g(v_i) = E(u_i|X_i, Z_i, D_i = 1)$ is approximated through a Fourier expansion:

$$g(v_i) = \sum_{i=1}^{\infty} b_j \times \lambda_i \times (Z_i'\gamma)^j \qquad (5.41)$$

where b_j is the traditional vector of parameters and λ_i is the inverse Mill's ratio evaluated at $Z_i\hat{\gamma}$. Heckman and Robb (1985a) report consistent estimates for the parameters of the primary equation.

Powell and Robinson's estimators

Powell (1987) exploits the index restriction (Assumption 3) by identifying observations by their value of this single index $g(Z_i\gamma)$. That is, if two observations, *i* and *s*, have similar values for the single index generating selection bias, then it is likely that subtracting the s^{th} observation from the i^{th} will eliminate selection bias. Assuming this, he uses the index restriction to rewrite the primary equation as:

$$Y_i = X_i\beta + g(v_i) + \varepsilon_i; \qquad \text{i} = 1,..., \text{N} \qquad (5.42)$$

Then, the parameter β can be estimated in a number of ways. Powell (1987) eliminates the unobservable $g(Z_i\gamma)$ by differencing:

$$Y_i - Y_s = (X_i - X_s)'\beta + g(v_i) - g(v_s)$$

and then applies the weights w_{is} that would be close to zero whenever u_i was not close close to u_s. The weighted model:

$$w_{is} \times (Y_i - Y_s) = w_{is} \times (X_i - X_s)'\beta$$

effectively eliminates the index restriction $g(Z_i\gamma)$ and β can be consistently estimated by OLS.

Robinson (1988) uses an alternative approach. With an estimate of $Z_i\gamma$ and conditioning (5.42) on v_i, he obtains:

$$E(Y_i|v_i) = E(X_i|v_i) \times \beta + g(v_i); \qquad i = 1,..., N \qquad (5.43)$$

Then subtracting (5.43) from (5.42) and writing:

$$Y_i - E(Y_i|v_i) = \{X_i - E(X_i|v_i)\} \times \beta + \varepsilon_i; \qquad i = 1,..., N$$

he estimates both $E(Y_i|v_i)$ and $E(X_i|v_i)$ non-parametrically. Regressing the non-parametric residuals $Y_i - E(Y_i|v_i)$ against $X_i - E(X_i|v_i)$ enables him to estimate β with OLS. Note that one cannot have a matrix X_i equaling a constant since then $X_i - E(X_i|v_i) = 0$ and valuable information is lost since β cannot be estimated.

Powell's (1987) and Robinson's (1988) estimators are connected to each other. To see the connection suppose that $E(Y_i|v_i)$ is estimated by kernel method with a uniform density as the kernel. Then, $\hat{E}(Y_i|v_i) = N_h^{-1} \times \sum_{s \in I_h} Y_j$ where I_h are those values of $s = 1, ..., N$ that have X_s within $\pm h/2$ of X_i and N_h is their number. It is clear that:

$$Y_i - \hat{E}\left(Y_i|v_i\right) = N_h^{-1} \times \sum_{s \in I^*} \left(Y_i - Y_s\right) = \sum_{s=1}^{N} w_{is} \times \left(Y_i - Y_s\right) \tag{5.44}$$

where $w_{is} = 0$ if $s \notin I_h$ and equals N_h^{-1} if $s \in I_h$. In the same way $X_i - E\left(X_i|v_i\right)$ can be replaced by $X_i - \hat{E}\left(X_i|v_i\right) = N_h^{-1} \times \sum_{s \in I^*} \left(X_i - X_s\right) = \sum_{s=1}^{N} w_{is} \times \left(X_i - X_s\right)$ and obtain Powell's (1987) estimator.

Ahn and Powell's estimator

Ahn and Powell (1993) also approach the problem of selection bias from Powell's (1987) perspective. The estimation method for β is summarized in a two step-strategy. The first step estimates γ non-parametrically while in the second step the estimator of β is a weighted instrumental variable estimator of all pairwise differences $y_i - y_s$ of dependent variable on the corresponding pairwise differences $x_i - x_s$ of regressors. Differences $z_i - z_s$ are used in instruments, these being suitable functions of the conditioning variables $w_i = Z_i\gamma$ and $w_s = Z_s\gamma$. This estimator is given by:

$$\hat{\beta} \equiv \left[\hat{S}_{ZX}\right]^{-1} \times \hat{S}_{ZY}$$

where

$$\hat{S}_{ZX} \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{s=i+1}^{N} \hat{w}_{is} \left(Z_i - Z_s\right) \times \left(X_i - X_s\right)'$$

$$\hat{S}_{ZX} \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{s=i+1}^{N} \hat{w}_{is} \left(Z_i - Z_s\right) \times \left(Y_i - Y_s\right)'$$

$$\hat{w}_{is} \equiv \frac{1}{h_2} \times k\left(\frac{\hat{g}_i - \hat{g}_s}{h_2}\right) \times D_i \times D_s$$

k(.) is, again, a kernel function, *h* is a bandwidth and $Z_i \equiv Z\left(w_i\right)$ is an instrumental variable defined as $Z : \Re^q \to \Re^p$.

The paper gives conditions under which the proposed estimator is consistent and asymptotically normal. Newey and Powell (1993) consider its efficiency properties by calculating its semi-parametric efficiency bound under weak conditions. They result the estimator failed to achieve the efficiency bound of Newey (1990).

Lee's estimator

Lee (1994) suggests estimation of equation

$$E\left(Y_i \middle| X_i, Z_i, D_i = 1\right) = X_i\beta + g\left(v_i\right); \qquad \text{i} = 1,..., \text{N}$$

by instrumental variables. Lee's model also incorporates endogenous regressors in the primary equation. Given the structure of the estimator he describes his procedure as semi-parametric 2-stage Least Squares. To ensure that the estimator has desirable properties, Lee employs a trimming function $\tau(X_i)$. The estimator is then defined as:

$$\hat{\beta}_{S2LS} = \left[\hat{Z}'\hat{X}\left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'\hat{Z}\right]^{-1} \times \left[\hat{Z}'\hat{X}\left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'\hat{Y}\right]$$

where $\hat{Z}$, $\hat{X}$ and $\hat{Y}$ are matrices with typical elements $\left\{\tau\left(X_i\right)\left(Z_i - E\left(Z_i \middle| Z_i\hat{\gamma}\right)\right)\right\}$, $\left\{\tau\left(X_i\right)\left(X_i - E\left(X_i \middle| Z_i\hat{\gamma}\right)\right)\right\}$ and $\left\{\tau\left(X_i\right)\left(Y_i - E\left(Y_i \middle| Z_i\hat{\gamma}\right)\right)\right\}$, respectively. The estimator is shown to be asymptotically efficient.

Donald's estimator

Finally, Donald (1995) considers the prototypical issue of estimation of a sample selection model in which heteroscedacity exists. Although heteroscedacity results in inconsistent estimates, it is difficult to be tackled without distributional assumptions. For this reason Donald develops a semi-parametric estimator where in the first stage he assumes that the errors u_i and v_i are bivariate normally distributed with covariance matrix

Σ with diagonal elements σ_u^2 and σ_v^2, and off-diagonal σ_{uv}. From bivariate normality, the primary model can be written as:

$$Y_i = X_i'\beta + \rho \times \sigma_u \times \lambda_i + \xi_i \tag{5.45}$$

where, as before, λ_i is the inverse Mill's ratio evaluated at $Z_i\hat{\gamma}$ and ξ_i is a heteroscedastic error term with mean zero and variance:

$$V\left(\xi_i \middle| X_i\right) = \sigma_u - \rho \times \sigma_u \times \left[Z_i\hat{\gamma} \times \lambda_i + \lambda_i^2\right]$$

Since λ_i is bounded away from 0 and ∞ (as is required for the transformed second-stage regression to be valid), one can transform model (5.45) into the semi-parametric regression form:

$$\frac{Y_i}{\lambda_i} = \frac{X_i}{\lambda_i}\beta + \rho \times \sigma_u + \frac{\xi_i}{\lambda_i} \tag{5.46}$$

where the last term is a heteroscedastic error and $C = \rho \times \sigma_u$ is an unknown term that can be eliminated with the "differencing out method", discussed by Robinson (1988) and Vella (1998), as

$$\frac{Y_i}{\hat{\lambda}_i} - E\left[\left(\frac{Y_i}{\hat{\lambda}_i}\right)\middle| Z_i\right] = \left\{\frac{X_i}{\hat{\lambda}_i} - E\left[\left(\frac{X_i}{\hat{\lambda}_i}\right)\middle| Z_i\right]\right\} \times \beta + \eta_i$$

In this equation, parameter β can be estimated by OLS over the differenced sample given that

$$\frac{X_i}{\hat{\lambda}_i} - E\left[\left(\frac{X_i}{\hat{\lambda}_i}\right)\middle| Z_i\right] \neq 0$$

The resulting estimator is proved to be consistent and asymptotically normal.

Despite the huge literature in the semi-parametric methods and the extreme popularity of the 2-steps procedures, semi-parametric methods have rarely been employed. Intuitively, this is partially due to the relative difficulty in the implementation and the estimation of the associated covariance matrices required for inference. Newey, Powell and Walker (1990) employ a number of 2-steps semi-parametric procedures to real data from married women's hour of work that they compare with 2-step parametric models. They find very little difference between the point estimates and conclude that the parametric procedures perform well if the conditional mean of the model is correctly specified. Thus, in some cases, the regression function appears to be more important than specifying semi-parametrically the error distribution. Heckman (1990b) also comments that methods simpler than the semi-parametric procedures may be robust after all. However, in cases where it is implausible to assume a specific density function for the errors, semi-parametric methods are surely essential.

5.13.2.3 Identification of the Primary's Equation Intercept

As it was noted before, the intercept in the primary equation is difficult to be identified through the above semi-parametric procedures. Most of the above estimators find impossible to distinguish an intercept term in X_i from an intercept in $g(.)$ since they absorb it into the definition of $E(u_i|Z_i, D_i = 1) = g(Z_i'\gamma)$. However, in many cases the intercept has economic interest. Heckman (1990b) considers the essentiality of this term in the estimation of the Average Treatment Effect or the Effect of Treatment on the Treated. He supports that, from the above estimators, only Galland and Nychka's (1987) one produces consistent estimators of the density $f(u_i, v_i)$ and of the primary's equation intercept because of the strong smoothness and continuity assumptions about the distribution of the error terms they pose. In this framework then, one may estimate the intercept as:

$$\hat{\beta}_{(0)} = \frac{\sum_{i=1}^{N} \left(Y_i - X_{(1)i} \times \beta_{(1)}\right) \times D_i \times I\left(Z_i\gamma \succ s\right)}{\sum_{i=1}^{N} D_i \times I\left(Z_i\gamma \succ s\right)}$$

where X_i and β vectors are partitioned into [1: $X_{(1)i}$] and [$\beta_{(0)}$: $\beta_{(1)}$], respectively and s reflects a smoothing parameter. Thus, the basic idea is to get the average value of the deviation $Y_i - X_{(1)i} \times \beta_{(1)}$ for observations where the expected value of the errors approaches to zero as N goes to infinity.

5.14 Sample Selection Models With Alternative Censoring Rules

Many authors have considered sample selection modeling and parameter estimation. Specifically, parametric models have found interesting applications in several empirical studies. The estimators discussed so far have been limited to a dependent variable in the selection equation that takes the value zero or one. In other words, so far only sample selection models with discrete choice rules have been reviewed.

In practice, the selection mechanism is not limited to this simple case. Even if any kind of dependent variable in the selection equation can be transformed into a binary through some threshold, an extended (continuous) form for D_i may reveal additional information, not exploited otherwise. Thus, it is interesting to consider a slight generalization in the specification of the censoring function determining the selection.

At this point it is worth discussing sample selection models with alternative censoring rules. A general representation can be:

$$Y_i^{(1)} = X_i\beta + u_i; \quad i = 1,...,N \tag{5.47}$$

$$D_i^{(1)} = Z_i\gamma + v_i; \quad i = 1,...,N \tag{5.48}$$

$$D_i = h\left(D_i^{(1)}\right) \tag{5.49}$$

$$Y_i = j\left(D_i, Y_i^{(1)}\right) \tag{5.50}$$

It is obvious that the only difference with the alternative model (5.10) – (5.13) is the generic form *h(.)* of the selection mechanism and the process determining the observability of Y_i, $j(.)$. Various cases of this general model are considered below.

5.14.1 Other Tobit Type Censoring Rules

Tobit models refer to regression models in which the range of the dependent variable is constrained in some way. In economics such a model was first suggested in the pioneering work of Tobin (1958) who analyzed household expenditures on durable goods using a regression model that assumes non-negative expenditures (the dependent variable of his regression model). This kind of models as well as its various generalizations is known popularly among economists as *Tobit* or *Tobit Type censored models* because of their similarity to Probit models. Amemiya (1984) gives a complete description of Tobit models, the censoring rules that include their estimation methods and their properties. We are interested in Tobit censoring in terms of a selection bias model. Tobit Type-II censoring has already described in the previous paragraphs. Thus now, we focus attention to Tobit-Type III and Tobit-Type IV censoring.

Tobit Type-III censoring Rules

Tobit – Type III is the most commonly considered censoring type. Suppose that the dependent variable in the selection equation is partially observed above some threshold, let's say zero. Instead of observing just the sign of $D_i^{(1)}$ in the estimation process, one can also exploit its positive values. An example of this model is Tobin's (1958) expenditure model and Heckman's (1974) labor supply model. The choice process, indicated by *h(.)*, is identified as:

$$D_i = \begin{cases} D_i^{(1)} & if \quad D_i^{(l)} \succ 0 \\ \\ 0 & otherwise \end{cases} \tag{5.51}$$

and

$$Y_i = Y_i^{(1)} \times I\left(D_i \succ 0\right) \tag{5.52}$$

where I_i is an indicator function, taking the value one if Y_i is uncensored and zero otherwise. Note that this model differs from Type-II only in that the values of D_i are observed when $D_i^{(1)}$ is positive in this model.

Under Assumption 1, the appropriate way to estimate the selection equation is to maximize the log-likelihood function

$$log\, L = \sum_{D_i=0} log\left[1 - \Phi\left(Z_i\gamma/\sigma_v\right) - \frac{N_1}{2} log\, \sigma_v^2 - \frac{1}{2\sigma_v^2} \times \sum_{D_i=1} \left(D_i - Z_i\gamma\right)^2\right]$$

over γ, β and σ_v^2 to yield consistent and asymptotically normal estimates $\hat{\gamma}$, $\hat{\beta}$ and s_v^2. Olsen (1978b) proved the global concavity of Log$L$ in the Tobit model. A standard iterative method such as Newton-Raphson always converges to a global maximum of log$L$.

Under Heckman's 2-step formulation, estimation of the primary equation in a Tobit model demands computation of the correction term:

$$v_i = \left(1 - I_i\right) \times \left\{\frac{-\phi\left(Z_i\gamma^*\right)}{1 - \Phi\left(Z_i\gamma^*\right)}\right\} + I_i \times \left(D_i^{(1)} - Z_i\gamma\right)$$

that corresponds to Vella's (1993) generalized residuals. Here, $I_i = 1$ when $D_i = D_i^{(1)} \succ 0$. The parameters β and C of the primary equation

$$Y_i = X_i\beta + C \times v_i + \varepsilon_i$$

are estimated adequately with Least Squares procedure. The only difference between this method and the one of Heckman's is that the former manipulates the correction term to take into account the extended information in D_i.

Apart from parametric, a number of semi-parametric procedures relax the normality Assumption 1. To account for the selection model, Powell (1984, 1986) estimates γ semi-parametrically, includes the corresponding residuals in the primary equation and uses OLS to estimate β. In an alternative approach, Itcimura (1993) develops the *Semi-parametric Generalized Least Squares Estimator* (SGLS) to estimate γ.

Lee (1994) defines the estimator:

$$\hat{\beta} = \left\{ \sum_{i=1}^{N_1} \left[X_i - E\left(X \middle| v_i \succ -Z_i'\gamma, Z'\gamma \geq Z_i'\gamma\right) \right]' \times \left[X_i - E\left(X \middle| v_i \succ -Z_i'\gamma, Z'\gamma \geq Z_i'\gamma\right) \right]^{-1} \right\}$$
$$\times \left\{ \sum_{i=1}^{N_1} \left[X_i - E\left(X \middle| v_i \succ -Z_i'\gamma, Z'\gamma \geq Z_i'\gamma\right) \right]' \times \left[Y_i - E\left(Y \middle| v_i \succ -Z_i'\gamma, Z'\gamma \geq Z_i'\gamma\right) \right]^{-1} \right\}$$

where the estimates of expectations can be obtained via kernel smoothing. To implement this estimator, a first step estimation of γ is conducted by a procedure described in Lee (1992). Under general regularity conditions, the perceived 2-steps estimator is proved to be consistent and asymptotically normal. Lee (1992) notes that this procedure does not impose any exclusion restrictions in contrast with a semi-parametric estimation of a discrete selection model (see Chamberlain, 1986).

Honore, Kyriazidou and Urdy (1997) also provide two different 2-step semi-parametric procedures of this model that are applicable under different assumptions. The first assumes conditional symmetry on the disturbances of (u_i, v_i), that is, conditional on (X_i, Z_i) the disturbances (u_i, v_i) are distributed like $(-u_i, -v_i)$. The second estimator is based on the idea of pairwise comparisons. Both of them are consistent and asymptotically normal under conditional symmetry and independence between the errors and the regressors.

Tobit Type IV censoring

At this point it is worth mentioning another Tobit-type selection rule that is proposed by Cragg (1971). Cragg assumes that observability of Y_i (or equivalently participation decision) requires satisfaction of two censoring rules, that is:

$$Y_i^{(l)} = X_i\beta + u_i; \quad i = 1,...,N \tag{5.54}$$

$$D_{1i}^{(l)} = Z_{1i}\gamma_1 + v_{1i}; \quad i = 1,...,N \tag{5.55}$$

$$D_{2i}^{(l)} = Z_{2i}\gamma_2 + v_{2i}; \quad i = 1,...,N \tag{5.56}$$

where

$$D_{1i} = \begin{cases} D_{1i}^{(1)} & if \quad D_{1i}^{(l)} \succ 0 \\ \\ 0 & otherwise \end{cases} \qquad and \qquad D_{2i} = \begin{cases} D_{2i}^{(1)} & if \quad D_{2i}^{(l)} \succ 0 \\ \\ 0 & otherwise \end{cases}$$

and

$$Y_i = Y_i^{(1)} \times I\left(D_{1i} \succ 0, D_{2i} \succ 0\right)$$

The terms (u_i, v_{i1}, v_{i2}) are iid drawings from a trivariate normal distribution. This type of censored models are called Tobit-type IV or *double-hurdle* model and is closely related with censoring rules based on multiple indices that will be analyzed in a following paragraph. A similar idea is adopted in Deaton and Irish (1982), Blundell and Merghir (1987) and Blundell, Ham and Merghir (1987). The estimation procedure is reported in Amemiya (1984) and is slightly different from the previous.

5.14.2 Multiple Alternatives – Ordered Censoring Rules

Another commonly encountered censoring rule for the alternative selection bias model (5.47) – (5.50) is formulated by defining an ordered choice set. In this case censoring function *h(.)* generates a series of ordered outcomes through the following rule:

$$D_i = 0 \text{ if } -\infty \prec D_i^{(1)} \le 0; \quad D_i = 1 \text{ if } 0 \prec D_i^{(1)} \le m_1;$$

$$D_i = 2 \text{ if } m_1 \prec D_i^{(1)} \le m_2; \ldots; \; D_i = k \text{ if } m_{k-1} \prec D_i^{(1)}$$

where m_i, $i = 1, \ldots, j-1,$ denotes separation points satisfying $m_1 < m_2 < \ldots < m_{j-1}$. The outcomes Y_i are ordered as:

$$\begin{aligned} Y_{0i} &= X_i\beta_0 + u_{0i} \quad \text{if} \quad D_i = 0 \\ Y_{1i} &= X_i\beta_1 + u_{1i} \quad \text{if} \quad D_i = 1 \\ &\ldots \\ Y_{ki} &= X_i\beta_k + u_{ki} \quad \text{if} \quad D_i = k \end{aligned} \tag{5.58}$$

where X_i affects the return to each choice (attributes), β_{ik} is the vector of parameters to be estimated and u_i represents the vector of the unobserved error terms. This term allows for heterogeneous preferences, so that to affect the return to each choice differently.

Garen (1984) exploits the ordering in the selection equation with the development of a continuous selection model. He considers the case where $D_i^{(1)}$ is continuously observed (Tobit Type-III selection rule) and, under Assumption 1, suggests an appealing estimation procedure. He applies this procedure at data from a "return from schooling" study. The censoring variable $D_i^{(1)}$ is the "years of schooling" and is being treated as a continuous variable. Obviously, this formulation exploits more information than the model with a binary choice variable where one defines $D_i = 1$ if the individual takes schooling education and $D_i = 0$ otherwise.

Garen derives and estimates the model. His most interesting result, similar to Willis and Rosen (1979), is twofold. First, expected lifetime earnings influence the decision to attend college. Second, those who did not attend college would have earned less than measurably similar people who did attend, while those who attend college would have earned less as high school graduates than measurably similar people who stopped after high school. Willis and Rosen (1979) have called this characteristic relationship as "comparative advantage".

Vella (1993) considers the same selection rule from a different perspective. Analogically to the conventional selection model (5.10) – (5.13), the analyst may observe Y_{1i} for a specific value of D_i. Let's say that D_i takes values from 1 to k-1 according to the interval $D_i^{(1)}$ belongs to. Without loss of generality, Y_i is observed if and only if $m_{k\text{-}p} \leq D_i^{(1)} \leq m_{k\text{-}l} \Leftrightarrow D_i = 1$. In this framework, relationship (5.52) is of the form:

$$Y_i = Y_i^{(1)} \times I\left(D_i = 1\right)$$

Vella studies a general methodology of estimating the selection model (5.47) – (5.50) with this selection rule, under Assumption 1. In the first step he estimates γ by Ordered Probit (for a description see Appendix 1) and then computes the generalized or Probit residuals of Pagan and Vella (1989) for each outcome:

$$v_i = \frac{\phi\left(m_{k-1} - Z_i\hat{\gamma}^*\right) - \phi\left(m_k - Z_i\hat{\gamma}^*\right)}{\Phi\left(m_k - Z_i\hat{\gamma}^*\right) - \Phi\left(m_{k-1} - Z_i\hat{\gamma}^*\right)} \tag{5.59}$$

and includes them in the primary equation to account for selectivity. The equation of interest is then:

$$Y_i = X_i\beta + C \times v_i + \varepsilon_i$$

where β and C can be estimated by Ordinary Least Squares.

To relax Assumption 1, Vella (1993) suggests capturing departures from normality in the disturbances u_i by powering up the generalized residuals by the index $Z_i\gamma$ and its higher powers. Further discussion of this topic is given in Lee (1982).

5.14.3 Multiple Alternatives-Unordered Censoring Rules (Polychotomous model)

In many applications, ordering in outcomes is not possible. Thus, the above selection rule is implausible and an alternative model has to be considered. Extending the work of McFadden (1973) on the discrete choice models, Lee (1983) and later Dubin and McFadden (1984) developed two separate parametric approaches to deal with selectivity bias for the case where individuals are faced with more than two mutually exclusive alternatives and they have to choose only one.

Suppose a simple canonical model with a single outcome and multiple alternatives. Let us say that each individual is faced up with M alternatives $k = 1, \ldots, M$, $k \in A$. For each of N observations (individuals), assume that the observed outcome for each k alternative, Y_k, is modeled as:

$$Y_{ik} = X_{ik}\beta_k + u_{ik} \tag{5.60}$$

The choice model is specified as:

$$D_{ik}^{(1)} = Z_{ik}\gamma_k + v_{ik} \tag{5.61}$$

A different parameter vector for each outcome characterizes this model that is known in the literature as *polychotomous selection model*. The variables X_{ik} and Z_{ik} are distinct vectors of observed characteristics and assumed to be exogenous. The unobserved variables (u_{ik}, v_{ik}) are independently and identically distributed (iid) across k alternatives with support $\Re^{M+1}$, continuous cumulative distribution function (cdf) F and continuous probability density function (pdf) f. The marginal cdf's of u_i and v_i respectively are $F(u_i)$ and $F(v_i)$ with corresponding pdf's $f(u_i)$ and $f(v_i)$.

Lee's (1983) approach

For the selection model (5.60) and (5.61), Lee recasts the selection rule (5.48) to assume that an alternative s is chosen if and only if:

$$D_{is}^{(l)} \succ \max_{k=1,\ldots,M,\ k\neq s} D_{ik}^{(l)} \tag{5.60}$$

Then, closely to Vella (1993), he defines a polychotomous variable I to indicate a specific alternative and a binary variable D_i. In terms of I and D_i, s alternative is chosen under the following rule:

$$I = s \Leftrightarrow \mathrm{D_{is}} = 1 \quad \text{if and only if} \quad v_s \succ -Z_{is}\gamma_s \tag{5.61}$$

where $v_s \equiv \max_{k=1,\ldots,M,\ k\neq s} D_{ik}^{(l)} - \varepsilon_{is}$

Although the selection rule seems well defined now, one observes that in equation (5.61) v_s are not independently and identically distributed (iid). Applying translation method an iid transformation for v_i's is constructed. Suppose that J_v is some continuous univariate cdf. Then the transformed v_i is:

$$v_i^{tr} = J_v^{-1}\{H_i[Max(D_i)]\} \tag{5.62}$$

where $H_i = F(Z_i\gamma + v_i)$. The joint distribution of (u_i, v_i^{tr}) is then:

$$F(u_i, v_i^{tr}) = J\{u_i, J_v^{-1}[F(v_i)]\} \tag{5.63}$$

To estimate the primary equation, Lee (1983) limits his analysis to the special case discussed by McFadden (1973) where the stochastic part v_i^{tr} of the selection equation (5.59) is assumed to be iid Gumbel (Extreme Value – I) distributed:

$$P\left(v_{ik}^{tr} \le v^{tr}\right) = exp\left(-e_{ik}\right) \tag{5.64}$$

As shown by McFadden (1973) and Domencich and McFadden (1975), in this case the selection probabilities are computed from:

$$P(D_i = s) = \frac{exp(Z_i\gamma)}{\sum_k exp(Z_k\gamma)} \tag{5.65}$$

And the distribution function of v_s is given by:

$$F\left(v^{tr}\right) = P\left(v_{ik}^{tr} \prec v^{tr}\right) = \frac{exp\left(v^{tr}\right)}{exp\left(v^{tr}\right) + \sum_{\substack{k=1,2... \\ k \neq j}} exp(Z_i\gamma)} \tag{5.66}$$

This is the *Multinomial Logit model (MNL)* of McFadden (1973). Application of the Multinomial model for estimation of selection probabilities as well as of parameter γ is the result of the comparison of several latent variables as indicated by formula (5.60). On the contrary, in the simple case of ordered outcomes where a single latent variable was responsible for the observed outcome, the structure of the model is simpler and the appropriate estimation procedure is Ordered Probit.

Under Multinomial Logit estimation, by assuming that the marginal distribution of u_{ik} is standard normal N(0, 1) and denoting $v_i^{tr} = J_v = \Phi^{-1}\,[F(v_i)]$, the 2-step procedure requires estimation of:

$$Y_{ik} = X_{ik}\beta_k + \sigma_{uk}\rho^2{}_k\,\phi\left[J_k\left(Z_{ik}\hat{\gamma}\right)\right] / F\left(Z_{ik}\hat{\gamma}\right) + n_{ik} \tag{5.67}$$

where $\sigma_{uk} = \sqrt{Var(u_k)}$, ρ_k is the correlation coefficient between u_j and v_k^{tr} and $\hat{\gamma}$ is obtained by the Multinomial Logit model. OLS procedure provides adequate estimates of β_k and $C_k = \sigma_{uk} \times \rho_k$.

Dubin and McFadden's (1984) approach

The selection rule of the model (5.59) – (5.60) implies that an alternative *s* is chosen when:

$$D_s = 1 \Leftrightarrow D^*_s > 0 \Leftrightarrow Z_s\gamma + v_s > Z_k\gamma + v_k \Leftrightarrow$$
$$Z_s\gamma - Z_k\gamma > v_k - v_s, \text{ for } k = 1, \ldots M \text{ and } k \neq s \qquad (5.68)$$

This formulation considers polychotomous choice model as a model with M – 1 binary decision rules with partial observations. In order to estimate the parameters of the regression model, one has to correct for the selectivity bias due to that causes $E(u_i|v_i) \neq 0$. In this case the correction is expressed as:

$$\lambda_h = E(u_{ik}|v_i) = \frac{1}{F(v_i)} \times \int_{-\infty}^{c_1} \ldots \int_{-\infty}^{c_m} u_{ik} \times f(v_i) dv_i \qquad (5.69)$$

Almost all applications of this approach have modeled $F(v_i)$ as multinomial logit. In this case the expectations in (5.68) has a closed-form expressions. Specifically, the parameter γ of the choice model is estimated then by Logit model. However, any distribution for which the expectations in (5.68) exist is admissible, including Multivariate Normal. In this case γ would be estimated simply by a Probit model.
To obtain an estimate of β, the corresponding model is:

$$Y_{ik} = X_{ik}\beta_k + C_k \times \lambda_{ik} + n_{ik} \qquad (5.70)$$

In Dubin and McFadden (DM) approach, one has to estimate first the parameter γ by using, more often, the conditional logit model. Then, substituting γ with the estimate $\hat{\gamma}$ estimate β and C_j from (5.69) by Least Squares method. This procedure is repeated for each variable Y_k. Maddala (1983, pages 275 – 278) compares the above two approaches and results that the second approach (DM) is more cumbersome. For this reason Lee's method has become popular in applied research with polychotomous selection models. On the other hand, we have to mention that Lee's approach, although simpler, is less

robust than DM method since the latter allows for extra parameter flexibility in describing selectivity patterns. Schmertmann (1994) shows that the inexistence of a restrictive assumption such as (5.62) in DM estimator leads to better estimates of β and γ.

5.14.4 Censoring Rules Based on Multiple Indices

There are several practical instances where selectivity occurs in terms of several sources, rather than just one, as considered until now. This case must be separated from the ones described on the ordered and unordered selection rules where one could observe or not Y_i according to some value of $D_i^{(l)}$. Here, observability of Y_i depends on two or more criteria and thus a specific decision and the amount of Y_i are not so intimately related. This case is described by model (5.54) – (5.56). An illustrative example is provided in Maddala (1983).

Let us assume the simple case where the observability of Y_i depends on the value of $D_{1i}^{(l)}$ and $D_{2i}^{(l)}$. Two crucial features are posed here:

F1. Whether the selection model is a *joint decision* or a *sequential decision* model. The first supposes that $D_{1i}^{(l)}$ and $D_{2i}^{(l)}$ are defined over the entire set of observations. The second restricts the analyst to define $D_{2i}^{(l)}$ only for that individuals with $D_{1i}^{(l)} > 0$.

F2. Whether choices are *partially* or *completely* observed. In the first case, one observes D_{1i} and D_{2i} separately, while in the second he observes directly a combination of these two indices $D_i = D_{1i}^{(l)} \times D_{2i}^{(l)}$.

Most of the papers referred to that kind of censoring proceed to the estimation of the corresponding selection model under Assumption 1. However, the generalization made by Lee (1982, 1983) is straightforward.

Estimation of the selection bias model under the different specifications stated (F1 and F2) is described in Poirier (1980) who works with a joint decision model with partial observability, Abowd and Farber (1982) that examine a sequential decision model with partial observability and Maddala (1983) who considers a bivariate joint probit model with complete observability.

Estimation of the primary equation of a selective model under sequential decision and complete observability is not referred to the literature. In fact selection bias analysis under this framework is more complex than it seems theoretically. If one specifies the distribution of $D_{2i}^{(1)}$ and estimate γ_2 only for those with $D_{1i} = 1$ by the Probit method, he has the opportunity to examine the self-selection behavior in the second stage where people decide whether to participate ($D_{2i}^{(1)} > 0$) or not ($D_{2i}^{(1)} \leq 0$). However, it is not clear how one examines the self-selection behavior in both stages.

The case of multiple indices censoring seems to be plausible in practical situations where several factors may affect participation decision in terms of a social program. For example, returns from schooling is usually not only determined by the years of schooling of each person but also from other factors like:

- ✓ A college committee decision to accept an applicant for college education, or
- ✓ Financial difficulties of the student that, after acceptance, may force him to resign from the program.

A returns from schooling study without taking into account these important factors may lead in rather misleading results since selectivity is not determined only by one factor. On the other hand, a multiple indices approach requires a great amount of data for the additional information needed to evaluate the multiple factors and examine their affection to outcomes. Intuitively, rich data raise the financial and time costs of the evaluation study.

5.14.5 Other Types of Censoring

Until now, we have focused on two-equation models where the dependent variable in the selection equation is censored and in primary equation is continuous. Various combinations of the above rules would yield different types of censoring. Here, we introduce the case where the selection equation includes a continuous dependent variable and the primary equation a censored dependent variable.

In this framework, the general model can be written as:

$$Y_i^{(l)} = X_i\beta + u_i; \qquad i = 1,...,N \tag{5.71}$$

$$D_i = Z_i\gamma + v_i; \qquad i = 1,...,N \tag{5.72}$$

$$Y_i = l\left(Y_i^{(l)}\right) \tag{5.73}$$

$$Y_i = h\left(D_i\right) \tag{5.74}$$

Smith and Blundell (1986) consider Tobit type-III censoring for variable Y_i. Rivers and Vuong (1988) study this model under binary censoring (conventional model). In either type of censoring single ML estimation of (5.71) will produce inconsistent estimates for β due to the endogeneity of D_i. Thus, alternative methods have to be considered.

Vella (1992) examines the above model when the primary equation has a binary outcome variable and the selection equation has a dependent variable that is partially observed (Tobit Type-III censoring). He estimates the Tobit residuals for the subsample corresponding to $D_i \succ 0$ which simply take the form $\hat{v}_i = D_i - Z_i\hat{\gamma}$ where the hats denote the Tobit estimates. Then, he estimates the primary equation by Probit over the subset, satisfying $D_i \succ 0$ while including v_i as an explanatory variable. Vella (1998) also describes the *conditional Maximum Likelihood* estimation steps in terms of Heckman's (1978) endogenous variable model by firstly employing the bivariate normality assumption for the error terms of equations (5.71) and (5.72). Under Assumption 1:

1. Rewrite (5.71) as $Y_i^{(l)} = X_i\beta + \theta D_i + C\hat{v}_i + \delta_i$
2. Obtain OLS estimates of γ from (5.72) and compute the residuals $\hat{v}_i = D_i - Z_i\gamma$,
2. Estimate $Y_i^{(l)} = X_i\beta + \theta D_i + C\hat{v}_i + \delta_{1i}$ by Maximum Likelihood where $\delta_{1i} = \delta_i + C \times \left(v_i - \hat{v}_i\right)$ is normally distributes with zero mean. The normality is retained as $\hat{v}_i$ is a linear transformation of normally distributed random variables.

Although these are some general steps to proceed in the case where alternative censoring rules are considered, there are several other types of censored models that can be estimated, depending on the form of $l(\cdot)$ in (5.73), provided they require normality. Extension of the relative theory to non-normal case is not referred in the present literature.

5.15 The E.M. Algorithm Approach

So far, it has been stated that the evaluation problem is a missing data problem. All approaches attempt to solve it by estimating, or better by replacing, the missing information using matching methods, randomized experiments or other econometric methods.

An alternative statistical approach that unites a variety of approaches in econometrics and statistics replaces missing data, or functions of missing data, with expectations of the lost information computed with respect to the available data. This technique applies the *Expectation-Maximization Algorithm* or else *EM algorithm* of Dempster, Laird and Rubin (1977).

The E step of the algorithm is the one that replaces the missing data, $(Y_{1i} \mid D_i = 0)$ or $(Y_{0i} \mid D_i = 1)$, that is the missing information that has to be estimated. By assuming parametric functional forms for distributions of unobservables and parametric functional forms for behavioral functions the E step replaces the missing data or functions of them with its expectations. Then, the M step replaces the functions of missing data by their expectations and maximizes the complete data likelihood. Using parameters generated in M step, E step is repeated until the algorithm converges to a unique set of parameter values. Schematically, E.M. algorithm proceeds as follows:

Figure 5.2: Description of the EM Algorithm

The E step: Given the current value $\theta^{(t)}$ of the parameter vector, the E step computes the expected value of the complete data log-likelihood, given the observed data and the current parameters, which is called the "objective function":

$$Q\left(\theta \middle| \theta^{(t)}\right) = \int l(\theta, Y) \times f\left(Y^m \middle| Y^o, \theta^{(t)}\right) dY^m$$
$$= E\left\{ l(\theta | Y) \middle| Y^o, \theta^{(t)} \right\}$$

The M step: This step determines $\theta^{(t+1)}$, that is the parameter vector maximizing the log-likelihood of the complete data (i.e. the complete data log-likelihood). Formally, $\theta^{(t+1)}$ satisfies:

$$Q\left(\theta^{(t+1)} \middle| \theta^{(t)}\right) \geq Q\left(\theta \middle| \theta^{(t)}\right)$$

One iterates between the E and M steps until the algorithm converges.

Under some conditions, specified by McLachlan and Krishnan (1997), the algorithm converges to a stationary point, which however, may not be the maximum of the likelihood function. This constitutes the problem of local maxima of the EM algorithm. In any case, when E.M. converges, estimates of the parameters are obtained. Heckman (1990a) provides a complete illustration of this method. Amemiya (1984) enables this method to censored Tobit regression models.

Molenberghs, Bijnens and Shaw (1997) mention that the major drawbacks of the E.M. algorithm are its typically slow rate of convergence and the inability to provide automatically precision estimates. In the light of these observations, Newton-Raphson (N-R) algorithm may be better than the E.M. However, as Meilijson (1989) notes, the E.M. algorithm may be more beneficial than N-R in some cases since the latter exhibits a tendency to converge to values outside the allowable parameter space. Louis (1982) and Meng and Rubin (1991) proposed alternative methods to overcome the limitations of the E.M. algorithm.

5.16 Identification of the Distribution of Impacts

As it has been stressed, evaluation of a social program is not limited to estimation of structural models and mean parameters. Knowledge of the distribution of outcomes is also of great economic interest. Various features of this distribution can be estimated in order to attain a complete picture of the benefits from program participation.

To this extend, Heckman (1990a) provides a theorem for identification of the joint distribution of outcomes. Provided that sufficient variation in individuals' attributes exists and that participation is based solely on outcome maximization, he demonstrates that $F\left(Y_{0i}, D_i \middle| X_{0i}\right)$, $F\left(Y_{1i}, D_i \middle| X_{1i}\right)$ and $F\left(u_{0i}, u_{1i}\right)$ can be identified non-parametrically, but the joint distribution $F\left(Y_{0i}, Y_{1i}, D_i \middle| X_{1i}\right)$ cannot. No arbitrary parametric structure on the outcome equations or on the distribution of the unobservables generating outcomes needs to be imposed. The relative theorem and its proof are provided in Appendix 2, along with a useful generalization of Heckman and Smith (1998). To identify the full

joint distribution of outcomes through bounding, the assumption of same dependence across quantiles of Y_{Di} have to be adopted.

5.17 Discussion on the Non-Experimental Methods

Similar to experimental estimators, econometric estimators have also been the subject of criticisms for many years. Each approach has advantages over the alternative and this created a continuous juxtaposition between the experimenters (e.g. Burtless, Holland, LaLonde, Rosenbaum, Rubin), which are mainly statisticians, and the econometricians (e.g. Heckman, Itchimura, Robb, Smith). Several papers have been based on this subject. Here, we provide some discussion on the non-experimental estimators, reviewed in this dissertation.

Effectiveness of econometric estimators

Several applications have been conducted to compare the effectiveness of the experimental against the non-experimental methods. LaLonde's (1986) and Fraker and Maynard's (1987) ones are the most usually referred to in the literature. These studies are witnessed to have a strong influence in promoting the use of experiments to evaluate social programs in general, and employment and training programs in particular. The relative authors use an experimental evaluation of the National Supported Work Demonstration (NSW) as a benchmark against which to compare non-experimental estimates. The NSW experimental treatment group is employed in conjunction with matching comparison groups to estimate the impact of training. Several commonly used non-experimental estimators are then considered and a wide variety of impact estimates are obtained, most of which differ substantially from the corresponding experimental estimates. Based on these results the authors claim that only experiments provide correct estimates since they are based on the indisputable free from bias method of randomization. Thus, econometric estimators cannot entirely eliminate selection bias.

Heckman, LaLonde and Smith (1999) characterize LaLonde (1986) influential study as misunderstood. Once analysts define bias clearly, compare comparable people, take

into account the unemployment histories of the trainees and comparison group individuals, administer them the same questionnaire and place them in the same local labor market, much of these bias in using non-experimental estimators is attenuated. It is also suggested shifting the emphasis in program evaluation away from specifying econometric methods for selection bias and toward more careful construction and weighting of comparison groups.

Experimenters also argue that apart from the academic interest, there is also practical-policy interest in choosing a particular estimator (experimental or non-experimental). LaLonde (1986) produced several non-experimental estimators and then conducted a limited set of model selection tests that failed to eliminate the models that produce this variability. This failure led him stressing the inexistence of a way to choose among competing non-experimental estimators. On the contrary, experiments produce a consensus, that is one estimate rather than the bewildering array of econometric estimates. This statement constitutes the main criticism against econometric methods of evaluation.

Heckman and Hotz (1989) reanalyze the data and demonstrate that experimental claims are somewhat exaggerated. The authors of the pessimistic studies did not perform standard model specification tests, reviewed in Heckman and Hotz (1989). When these tests are performed, they eliminate all but the non-experimental models that produce the inference obtained by experiments. But even when the analyst cannot result in a single number (estimator), this does not mean that experiments have to be preferred to econometric methods. Heckman and Smith (1995) indicate that appearance of a consensus view is a consequence of only one interpretation of the data being given. Intuitively, this seems to limit the evaluation analysis.

Burtless and Orr (1986) criticize an assumption of the econometric methods, that once valid estimates from a program are produced these estimates are also valid for all similar programs. They claim that it is very dangerous to generalize from the effects of a particular combination of education, employment and training services to another. Moreover, they consider that it is unclear how one could extrapolate the estimated effects of an existing program to a completely new program. Following these well-founded statements of the authors, econometricians can similarly argue on the superiority of non-

experimental method. Since experiments usually produce only one estimate without invoking any distributional assumptions, it seems dangerous to generalize this estimate in special situations where some specific assumptions have to be considered.

Non-response Bias, Hawthorne Effects and Limited Duration Bias

Burtless and Orr (1986) claim that several of the experimental limitations also occur in econometric studies. Non-response bias and Hawthorne effects can also occur in non-experimental methods. Specifically, Hawthorne effects are referred to be more serious in non-experimental evaluations while Limited duration bias is proven to be present in the non-experimental demonstration of CETA program (1973).

Although these disadvantages may occur in both evaluation methods it is unlikely that they have more severe effects in econometric methods. Intuitively, in an experimental setting, where participants are informed of being subjects of study, the behavior is naturally affected. As for non-response bias, Burtless and Orr (1986) do not mention a convincing reason to support the advantage of experiments on this subject.

Choice of a Specific Non-Experimental Estimator

Holland (1989) indicates that causal inference in nonrandomized studies requires more data and more assumptions than in randomized ones. The latter argument constitutes one of the main strengths of experiments, namely the absence of distributional assumptions for an adequate analysis. He also supports that econometric methods produce a variety of estimators from where it is difficult to choose the most appropriate one and the usual strategy of the econometricians is to look over the "menu" of estimators, select the assumptions that make the most sense for the non-experimental setting at hand and then obtain the corresponding estimate. Instead of this naïve approach they could perform a sensitivity analysis (see Rosenbaum and Rubin, 1983a) to display the sensitivity of a particular estimator on (a) the distribution of u_i, (b) the dependence on u_i of the conditional distribution of D_i, given u_i and the other covariates, and on (c) the dependence on u_i of the conditional distribution of Y_i, given u_i and the other covariates.

Heckman and Hotz (1989, rejoinder) support that econometric literature has not ignored the issues (a) – (c) addressed by sensitivity analysis. Heckman and Robb (1985a)

approach all of these issues by assuming different functions for the error terms u_i and different estimators. They also claim that sensitivity analysis must not be considered as an essential tool in evaluation studies. Sensitivity analysis creates the illusion to the unwary that robust results are produced under plausible behavioral assumptions. However, this kind of analysis is a logical and computational impossibility, because the infinite number of possible assumptions that might be made relative to the available data. Instead of this, econometrics literature has focused on the identification of participants' impacts with alternative configuration of data and under alternative plausible identifying restrictions.

Cost of econometric methods

Burtless (1995) characterizes experimentation as a very expensive evaluation method. Experiments usually consume a great deal of real resources, especially in comparison with econometric analysis of existing data sources. However, this is not always true. As a convincing argument we pose that collection of high quality data entails great financial costs. Existing general survey data, which are inexpensively obtained, often contain either too few participants or non-participants, or contain too little information on individuals' characteristics. This information is, naturally, important for conducting better non-experimental evaluations and is usually obtained only by collecting costly new survey data.

Estimation of the conditional means

An apparent advantage of non-experimental methods over experimental ones is that the latter can estimate only one mean impact (e.g. effect of treatment on the treated), given one randomization is implemented. Estimation of additional mean parameters on a study can be conducted only by econometric methods that increase the overall cost of the study.

Furthermore, it is known that, by design, social experiments balance the bias between treatment and control group and eliminate them by the subtraction in estimating the TT parameter. Suppose that an analyst has the following structural models of treatment and control group persons:

$$E\left(Y_{0i}\middle|D_i=1,X_i,Z_i\right)=g_0(X_{0i})+E\left(u_{0i}\middle|D_i=1,X_{0i},Z_i\right)$$
$$E\left(Y_{1i}\middle|D_i=1,X_{1i},Z_i\right)=g_1(X_{1i})+E\left(u_{1i}\middle|D_i=1,X_{1i},Z_i\right)$$

Social experiments does not allow to separately identify the structure of $g_0(X_i)$ and $g_1(X_i)$ from the conditional error terms since bias and dependence between u_i and v_i are not eliminated, causing $E(u_i|v_i) \neq 0$. Specific assumptions have to be invoked for this identification, posed only by econometric methods.

At this point it is worth indicating that, as Heckman, Smith and Clements (1997) has proven, selectivity is neither the main nor the only type of bias in evaluation studies. Particularly, an alternative kind of bias, not encountered in the present thesis, arises due to measurement errors in the binary variable D_i when misclassification in the states $D_i = 0$ and $D_i = 1$ occurs. Skinner (1998) proposes a number of alternative estimators to reduce *misclassification bias* that considers of great importance. Based on the *measurement error model* of Chua and Fuller (1987) he suggests inference procedures to adjust for measurement errors.