Selectivity Problems in Quasi-Experimental Studies

Bengt Muthen and Karl G. Jöreskog

Eval Rev 1983 7: 139
DOI: 10.1177/0193841X8300700201

Selectivity problems can occur whenever one tries to estimate population parameters from a nonrandom sample. The sample may be nonrandom because only individuals with certain characteristics are selected into the sample (sample selection), or because individuals participate voluntarily in the sample (self-selection). Selective samples can also occur because individuals fall out of the sample for various reasons, despite an initial random sample (attrition). In such situations it is important to model the selection process as realistically as possible. Selectivity problems are discussed in terms of a general model that is estimated by the maximum likelihood method. Both single-group and multiple-group analyses are considered. The multiple-group case is related to the problem of evaluation of treatment effects in nonequivalent control group designs. The general model and the estimation procedure is illustrated by means of a simulation study. An extension of the general model to latent variable models is discussed.

SELECTIVITY PROBLEMS IN QUASI-EXPERIMENTAL STUDIES

BENGT MUTHÉN
University of California, Los Angeles
KARL G. JÖRESKOG
University of Uppsala, Sweden

1. INTRODUCTION

Selectivity problems can occur whenever one tries to estimate population parameters from a nonrandom sample. When the sample of data is nonrandom, it is important to try to model, as realistically as possible, the process by which the observed units have been selected into the sample. Selective samples may occur because only individuals

---

AUTHORS' NOTE: *This article was presented at the Conference on Experimental Research in the Social Sciences, Gainesville, Florida, January 8-10, 1981. This project, Methodology of Evaluation Research, has been supported by the Bank of Sweden Tercentenary Foundation, project director Karl G. Jöreskog. The authors thank Bengt Dahlqvist for fast and skillful programming.*

EVALUATION REVIEW, Vol. 7 No. 2, April 1983 139-174

0193-841X/83/020139-36$3.85

with certain characteristics, more or less precisely defined, are included in the sample. This may be the case in large social programs, for example, where only low-income families are eligible for the program (sample selection), or when individuals participate voluntarily in the program (self-selection). Selective samples may also occur in longitudinal studies due to attrition; that is, individuals fall out of the sample for various reasons, despite an initial random sample. Analyzing a selective sample as if it is random will result in biased and inconsistent estimates of the parameters.

Selectivity problems have been of considerable interest in recent econometric work, for example, see Stromsdorfer and Farkas (1980). Within the single-group regression framework, selectivity problems have been discussed in the context of labor force participation of married women by many writers, for example, Gronau (1974), Lewis (1974), and Heckman (1974, 1977). Selection modeling has also been applied to situations of self-selection in the choice of college education and regarding economic returns to schooling, in, for example, Griliches et al. (1978), Kenny et al. (1979), and Willis and Rosen (1979). Selectivity modeling in the analysis of longitudinal data has been considered by Hausman and Wise (1976, 1979). Selectivity problems have also been discussed in the context of evaluation of treatment effects in nonequivalent control group designs, for example by Goldberger (1972a, b), Cain (1975), in the overview by Reichardt (1979), and by Sörbom (1981).

In this article we shall discuss selectivity problems in terms of a model that in some respects is more general than those of previous writers. Selection modeling for a single group is considered in Section 2. Multiple-group issues are discussed in Section 3 and related to conventional analysis of covariance. A general model and its estimation is presented in Section 4. A simulation study is reported in Section 5, and in Section 6 an extension of the general selection model to latent variable models is discussed.

2. SELECTION IN A SINGLE GROUP

As an example from education, consider the case where y is an achievement test, x is a home background variable, and the model is

$$y = \beta_0 + \beta_1 x + \epsilon \qquad [1]$$

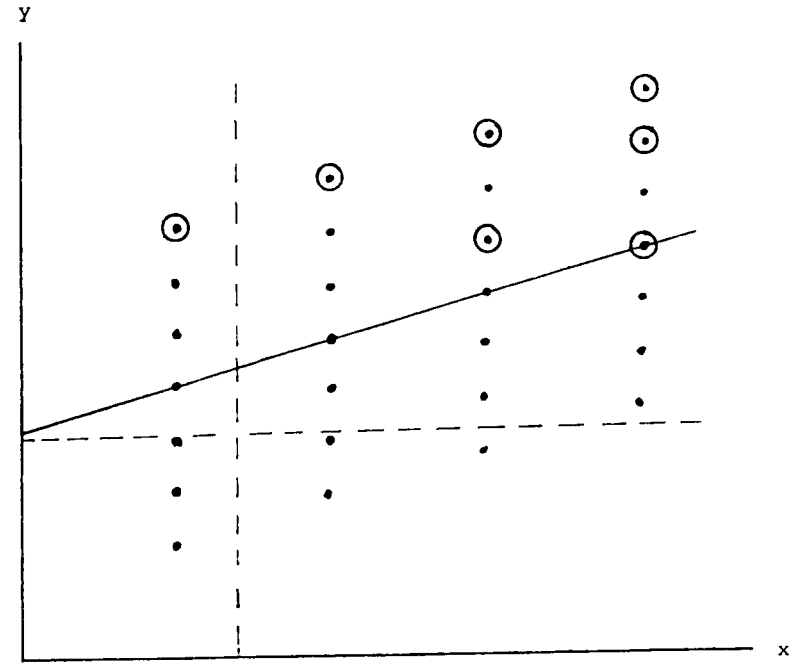


Figure 1

where ϵ is uncorrelated with x. Figure 1 shows a scatterplot of typical units in the population, say students of a certain age. (This graph is inspired by Hausman and Wise, 1976.) The straight line (equation 1) represents the population regression of y on x. If one has a random sample of observations on y and x, one can obtain unbiased estimates of β_0 and β_1 by ordinary least squares (OLS). If the sample is non-random, however, a population unit will be selected into the sample or not depending on the values of certain characteristics, which may be y, x, or other unobserved variables. If this fact is ignored, OLS will in general give biased estimates, which in turn leads to incorrect inferences for the full population. A solution to this problem is to try to model the selection process. Estimation can then be carried out for an extended model, including both the original regression relation, such as equation 1, and the selection model part. With a proper model specification, correct estimates can then be obtained for the parameters of equation 1.

In this article we shall assume that the selection process depends on a single selection variable η, and that units are selected into the sample if η exceeds a threshold value. In most cases the selection variable η is unobserved and its characteristics unknown. We shall first consider the cases when η coincides with x and y, respectively.

When η = x there is zero probability of selecting population units to the left of the vertical broken line in Figure 1. Here, students with "good" home background would be considered. However, this is of no consequence provided that units to the right of this line are selected randomly, and that the variation in x is sufficient to determine the slope β_1 of the population regression. Hence, when η = x, OLS gives unbiased estimates of β_0 and β_1.

When η = y, a unit is observed in the sample only if y exceeds a threshold. Here, only high-achieving students would actually be used in the analysis. In Figure 1 this means that population units below the horizontal broken line have zero probability of inclusion in the sample. In this case, the error term ϵ will be correlated with x in the sample, the mean of ϵ being larger for units with smaller x-values. When the threshold is zero, this corresponds to the familiar Tobit model (Tobin, 1958; Amemiya, 1973), originally proposed as a limited dependent variable model for consumption studies (no consumption if y = 0). OLS will give an estimate of the slope β_1, which is biased downward and inconsistent in large samples.

Now consider the case when η is unobserved. Units are selected if η exceeds a threshold. This is perhaps the most realistic case in the kind of application considered. Here, η may be a latent variable such as social disadvantage, where a high disadvantage results in an individual being selected. The reason for selection may be that limited funds are available, and measurement is concentrated on students that are thought to be in particular need of a certain schooling treatment. Still, the intent is to try to make inferences to the full population. The value of η represents a characteristic of the student that is not completely known to the investigator, who does not completely control the selection process. Figure 1 illustrates the probable case when η is negatively correlated with y, using encircled dots to exemplify population units that may not be included in the sample. For selectable units, the mean of ϵ is smaller for large x-values. Here there is no sharp division into selectable and nonselectable units. Ignoring selection will result in biased OLS estimates also in this case. This will be explicated below.

We will now consider the case when a proper specification of the selection process can be done, reviewing selection modeling attempted

in the literature so far. It should be noted that a weakness of selection modeling is that the general problem of misspecification may be enhanced. For instance, a misspecification of the regression relation, such as equation 1, may be mistaken for indication of selection bias (see Olsen, 1979; Stromsdorfer and Farkas, 1980: 13-41).

For the education example it may be realistic to assume that the selection variable η is linearly related to (although not completely determined by) the observed home background variable x,

$$\eta = \gamma_0 + \gamma_1 x + \delta \qquad [2]$$

We shall assume that in the total population the joint distribution of ϵ and δ is independent of x with means zero and with covariance matrix

$$\underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_{\epsilon\epsilon} & \\ \sigma_{\delta\epsilon} & \sigma_{\delta\delta} \end{bmatrix}$$

where $\sigma_{\epsilon\epsilon}$ is the variance of ϵ, $\sigma_{\delta\delta}$ is the variance of δ, and $\sigma_{\delta\epsilon}$ is the covariance between ϵ and δ. We shall also use the regression of ϵ on δ,

$$\epsilon = \omega\delta + \upsilon \qquad [3]$$

where υ is uncorrelated with δ and $\omega = \sigma_{\delta\epsilon}/\sigma_{\delta\delta}$ is the regression coefficient. Since the scale for η is arbitrary, we may without loss of generality assume that the threshold is zero and that $\sigma_{\delta\delta} = 1$, in which case $\omega = \sigma_{\delta\epsilon}$. This assumption is made throughout this section.

Consider the regression of y on x for selectable units with $\eta > 0$,

$$E(y|x, \eta > 0) = \beta_0 + \beta_1 x + E(\epsilon|\eta > 0)$$

But

$$E(\epsilon|\eta > 0) = E(\epsilon|\delta > -\gamma_0 - \gamma_1 x)$$

$$= \omega E(\delta|\delta > -\gamma_0 - \gamma_1 x)$$

so that

$$E(y|x, \eta > 0) = \beta_0 + \beta_1 x + \omega E(\delta|\delta > -\gamma_0 - \gamma_1 x) \qquad [4]$$

Let $\lambda(x) = \gamma_0 + \gamma_1 x$. Then the last term in equation 4 involves $E[\delta \mid \delta > -\lambda(x)]$, which is a monotonically decreasing function of λ denoted by $f(\lambda)$. It is clear that the ordinary least squares regression of y on x fails to give consistent estimates of β_1, unless $\sigma_{\delta\epsilon} = 0 (\omega = 0)$. This is because the ordinary regression of y on x omits the random variable $f(\lambda)$, which is correlated with x.

Let $p(z)$ denote the probability density function of δ, and let $P(x)$ denote the corresponding probability distribution function. We assume that $p(z)$ is symmetric about zero so that $p(-z) = p(z)$ and $P(-z) = 1 - P(z)$. Then

$$\begin{aligned} \Pr(\delta > -\lambda) &= 1 - \Pr(\delta \leqslant -\lambda) \\ &= 1 - P(-\lambda) \\ &= P(\lambda) \end{aligned} \qquad [5]$$

and

$$\begin{aligned} f(\lambda) &= E(\delta \mid \delta > -\lambda) \\ &= [1/P(\lambda)] \int_{-\lambda}^{\infty} z p(z)\, dz \\ &= -[1/P(\lambda)] \int_{-\infty}^{\lambda} z p(z)\, dz \end{aligned} \qquad [6]$$

Table 1 shows $\sigma_{\delta\delta}$, $p(z)$, $P(z)$, and $f(\lambda)$ for some of the well-known distributions: the normal, the logistic, the Student's t and the Laplace (this table is adapted from Goldberger, 1980). The case when δ has a standard normal distribution is of particular interest. Let $\phi(z)$ be the standard normal density function and let $\Phi(z)$ be the corresponding distribution function. Then the integral in equation 6 becomes $-\phi(\lambda)$ so that

$$f(\lambda) = \phi(\lambda)/\Phi(\lambda)$$

Therefore, in this case we have

$$E(y \mid x, \eta > 0) = \beta_0 + \beta_1 x + \omega f(\lambda) \qquad [7]$$

TABLE 1

Variance $\sigma_{\delta\delta}$ and Functions p(z), P(z), and f(λ) for Some Selected Distributions

Distribution	*Variance* $\sigma_{\delta\delta}$	*Density* $p(z)$	*Distribution Function* $P(z)$	*Truncated Mean Function* $f(\lambda)$
Normal	1	$(2\pi)^{-1/2}e^{-1/2z^2}$	$(2\pi)^{-1/2}\int_{-\infty}^{z}e^{-1/2x^2}dx$	$p(\lambda)/P(\lambda)$
Logistic	$\pi^2/3$	$e^z/(1+e^z)^2$	$1/(1+e^{-z})$	$[1/P(\lambda)]\log[1/(1-P(\lambda))]-\lambda$
Student*	$n/(n-2)$	$c_n(1+z^2/n)^{-1/2(n+1)}$	$c_n\int_{-\infty}^{z}(1+x^2/n)^{-1/2(n+1)}dx$	$[(n+\lambda^2)/(n-1)]\,p(\lambda)/P(\lambda)$
Laplace	2	$1/2\,e^{-\lvert z\rvert}$	$1/2\,e^{z}$, $z\leq 0$ $1-1/2\,e^{-z}$, $z>0$	$1-\lambda$, $\lambda\leq 0$ $(1+\lambda)/(2e^{\lambda}-1)$, $\lambda>0$

SOURCE: Adapted from Goldberger (1980).

*n is the degrees of freedom parameter and $c_n = \Gamma(1/2(n+1))/[\Gamma(1/2n)\cdot(n\pi)^{1/2}]$.

The conditional variance of y can also easily be obtained using equation 3 and known results for the truncated normal distribution (see Johnson and Kotz, 1972: 81-83).

$$\begin{aligned}\mathrm{Var}(y|x, \eta > 0) &= \mathrm{Var}(\epsilon|\eta > 0) \\ &= \omega^2 \mathrm{Var}(\delta|\delta > -\lambda) + \mathrm{Var}(v) \\ &= \omega^2 [1 - \lambda f(\lambda) - f^2(\lambda)] + \sigma_{\epsilon\epsilon} - \omega^2 \\ &= \sigma_{\epsilon\epsilon} - \omega^2 f(\lambda)\,[\lambda + f(\lambda)] \qquad [8]\end{aligned}$$

Hence, the true regression of y on x is nonlinear and heteroscedastic. Figure 2 shows $\phi(\lambda)$, $\Phi(\lambda)$, and $f(\lambda)$ for $-4 \leqslant \lambda \leqslant 4$. Figure 3 shows the mean (equation 7) and the variance (equation 8) of y as a function of x for $\beta_0 = \gamma_0 = 0$, $\beta_1 = 1$, $\gamma_1 = -1$ and $\omega = -1$. This corresponds to the third selection situation of Figure 1. It is seen that linear regression, ignoring selectivity, will here produce a downward biased estimate of β_1. Figures 4a-d show the straight line $\beta_0 + \beta_1 x$ and the true mean function for some combinations of the signs of β_1, γ_1 and ω.

Generalizing the previous model to an arbitrary number q of explanatory variables $\underset{\sim}{x}' = (x_1, x_2, \ldots, x_q)$, of which one may be the constant 1, and using vectors of regression coefficients $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$, the model

$$y = \underset{\sim}{\beta}'\underset{\sim}{x} + \epsilon \qquad [9]$$

$$\eta = \underset{\sim}{\gamma}'\underset{\sim}{x} + \delta \qquad [10]$$

$$y: \begin{cases} \text{observed if } \eta > 0, \\ \text{not observed, otherwise} \end{cases} \qquad [11]$$

with normally distributed errors can be seen as a generalized Tobit model, where the assumption $\eta = y$ has been relaxed (see Cragg, 1971). In addition to consumption and labor force studies in econometrics, where the y-variable is limited, this model has been used to model selectivity in various applications of the kind discussed in Section 1. This generalized Tobit model is the basic model we will use henceforth. For a recent survey of the statistical treatment of Tobit models, see Amemiya (1982).

The generalized Tobit model may be interpreted in two parts corresponding to the two relations in equation 10 and equation 9. With

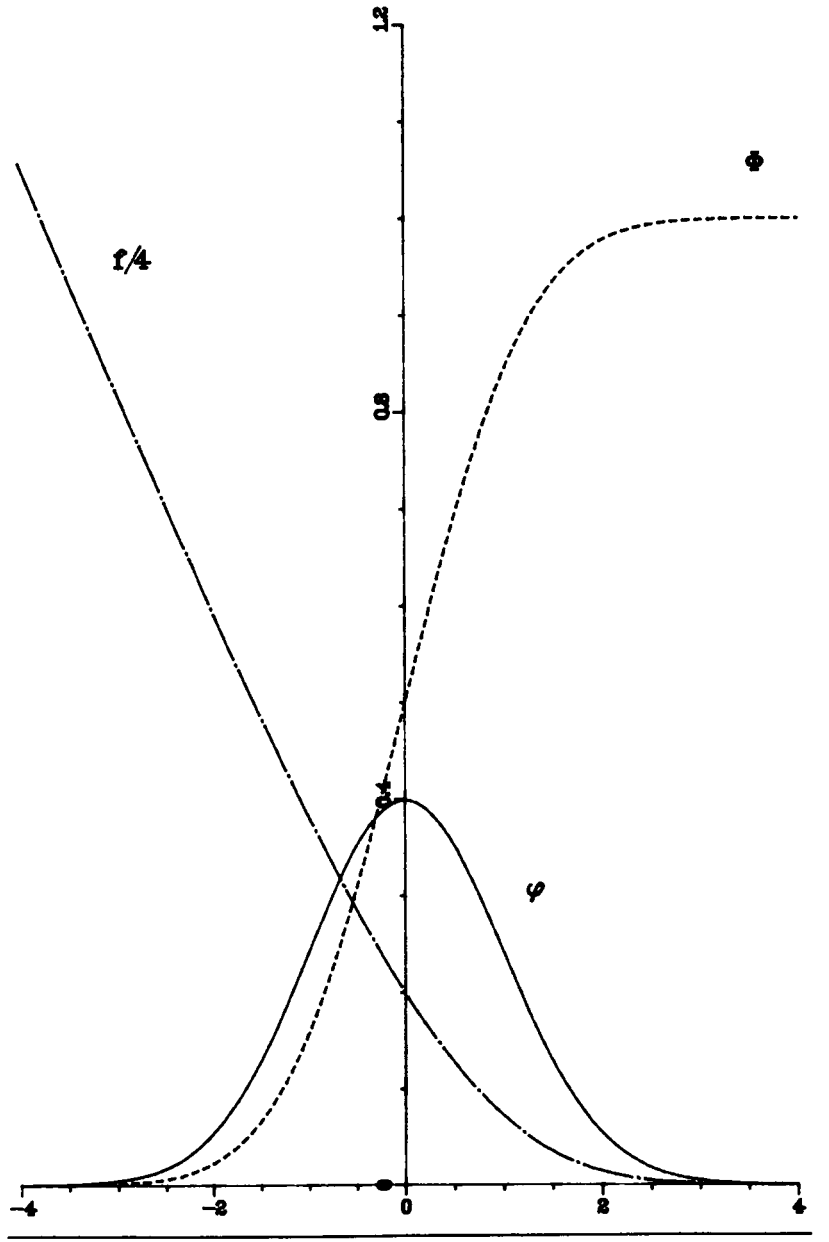


Figure 2: Functions $\phi(\lambda)$, $\Phi(\lambda)$, and $f(\lambda)$ for $-4 \leqslant \lambda \leqslant 4$

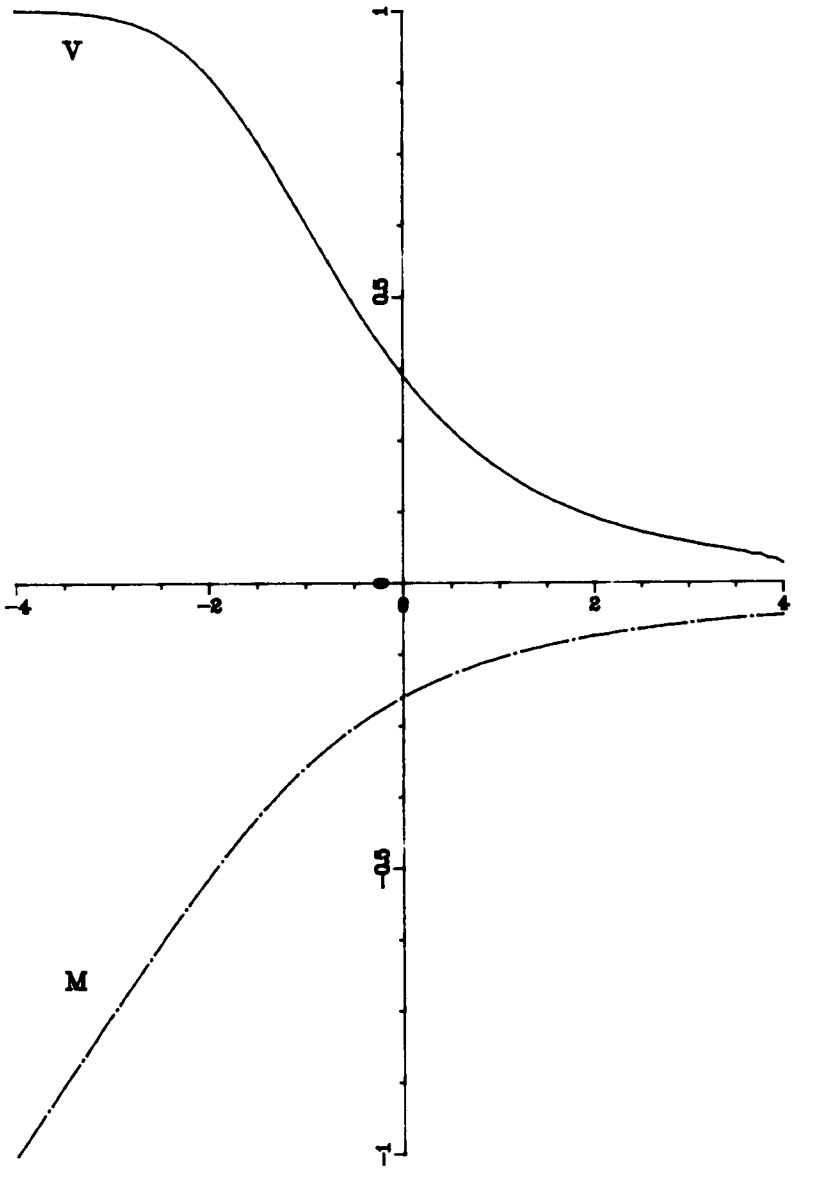


Figure 3: Mean Function $M(x) = \frac{1}{4}[\beta_0 + \beta_1 x + \omega f(\lambda)]$
Variance Function $V(x) = \sigma_{\epsilon\epsilon} - \omega^2 f(\lambda)[\lambda + f(\lambda)]$ for
$\beta_0 = \gamma_0 = 0, \beta_1 = 1, \gamma_1 = -1, \omega = -1$ and $\sigma_{\epsilon\epsilon} = 1$ and $-4 \leqslant x \leqslant 4$

FIGURE 4A

$\beta = 1, \gamma = -1, \omega = -1$

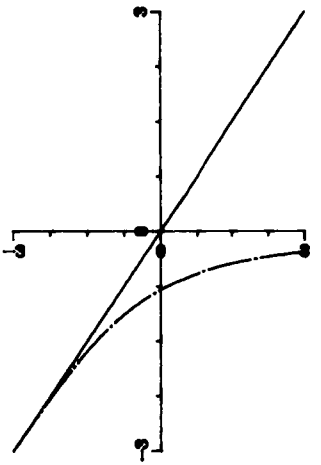


FIGURE 4B

$\beta = 1, \gamma = 1, \omega = 1$

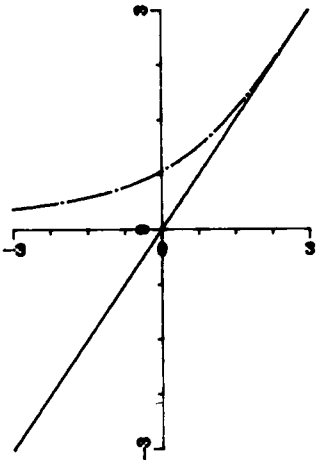


FIGURE 4C

$\beta = -1, \gamma = -1, \omega = 1$

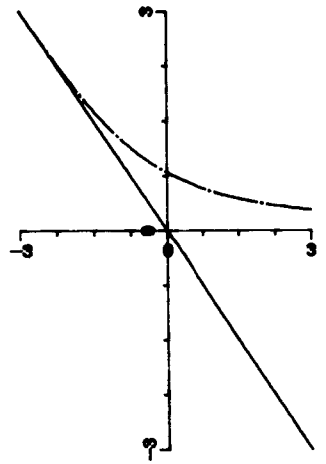
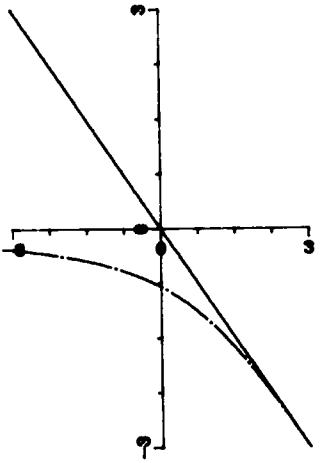


FIGURE 4D

$\beta = -1, \gamma = 1, \omega = -1$



Figures 4A-D: Linear Function $L(x) = \beta x$ and Mean Function $M(x) = \beta x + \omega f(\gamma x)$ for Four Combinations of β, γ, and ω and $-3 \leqslant x \leqslant 3$.

$\sigma_{\delta\delta} = 1$, the probability of the event y observed ($\eta > 0$) follows a Probit model,

$$\Pr(y \text{ observed } |\underset{\sim}{x}) = \Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}), \quad [12]$$

With $\Pr(y \text{ not observed } |\underset{\sim}{x}) = 1 - \Phi(\underset{\sim}{\gamma}'\underset{\sim}{x})$. The second part describes the distribution of y given $\underset{\sim}{x}$ and $\eta > 0$,

$$E(y|\underset{\sim}{x}, \eta > 0) = \underset{\sim}{\beta}'\underset{\sim}{x} + \omega f(\underset{\sim}{\gamma}'\underset{\sim}{x}). \quad [13]$$

Two types of samples must be distinguished. In the Probit model (equation 12), it is assumed that the sample includes units for which $\underset{\sim}{x}$ is recorded also for those with $\eta \leqslant 0$. This will be referred to as the *censored* case. When such units cannot occur in the sample, we have the *truncated* case.

Several techniques have been proposed for the estimation of equations 9, 10, and 11, using maximum-likelihood methods (e.g., see Griliches et al., 1978; Hausman and Wise, 1979), and various two-stage estimators applicable to the censored case only (e.g., see Heckman, 1979; Maddala and Lee, 1976). We will consider maximum-likelihood estimation, but the Heckman estimator will also be reported in the simulation study in Section 5.

In the first step of the Heckman estimator, $\underset{\sim}{\gamma}$ is estimated by maximum-likelihood Probit analysis. In the second step, OLS is applied to equation 13 in the truncated sample using the estimated $f(\lambda)$ as an additional x variable.

Recent contributions to selection modeling in the single-group case include studies pertaining to the robustness against deviations from the assumed functional form and error structure (e.g., see Goldberger, 1980; Hurd, 1979; Nelson, 1979; Olsen, 1979; Ray et al., 1980), and generalizations to more than one selection relation (e.g., see Tunali et al., 1980; Venti and Wise, 1980).

3. MULTIPLE-GROUP COMPARISONS

In this section we consider the analysis of treatment (intervention) effects for nonequivalent group designs. Such quasi-experimental designs are common in the evaluation of social programs or social experiments. Of particular concern is the case where outcome measurements are made before and after the intervention, for a control group,

and one or several treatment groups. In practice, randomization is infrequently accomplished, and the problem is how to separate the potential treatment effect(s) from group differences only produced by the way individuals were assigned to the different groups. Although originally intended for experimental settings, analysis of covariance (ANCOVA) is frequently used in this situation. Such ANCOVA applications have been strongly criticized, and attempts have been made to adjust the technique to fit the quasi-experimental setting. For a good summary of the problems and the various adjustment techniques, see Reichardt (1979) and Weisberg (1979).

As in the previous section, the nonequivalence of the control and treatment groups may be due to the investigator choosing to treat a certain subset of individuals (such as particularly needy ones) or due to self-selection by the individuals (such as volunteers in a new program). Nonequivalent groups may also arise due to attrition, despite initial randomization.

Data of this sort may be viewed as samples from different groups (populations) to be compared (see Thorndike, 1942). However, for one or several of the groups the sample(s) is (are) nonrandom or selective in the sense defined previously. Contrary to the ANCOVA approach, this article argues for the explicit modeling of the selection processes in order to avoid bias due to comparisons of nonequivalent groups.

Consider a quasi-experiment related to the education example discussed in Section 2. Say that there is one experimental group (E), one control group (C), and a single posttest y. Complete randomization has not been undertaken. It is suspected that the groups are different with respect to certain background characteristics, of which an important part is the variable x, say. The ANCOVA approach is to consider the regressions

$$\left.\begin{aligned} y^C &= \mu + \beta^C x^C + \epsilon^C \\ y^E &= \mu + \alpha + \beta^E x^E + \epsilon^E \end{aligned}\right\} \qquad [14]$$

assumed to hold for each of the two populations (groups). Given $\beta^C = \beta^E$, α is taken as the treatment effect.

With random sampling from each population (i.e., complete randomization), ANCOVA analysis, given $\beta^C = \beta^E$, consistently estimates the treatment effect α. If in fact x does not influence y, there would be no point in including it in equation 14, resulting in a simple analysis of variance. If x does influence y, its inclusion increases precision.

However, in many quasi-experimental studies there may be nonrandom selection of units to both controls and experimentals, and it is not necessarily the same selection variable that governs the selection process for controls and experimentals. A more appropriate model for this situation is

Controls:
$$y^C = \mu + \beta^C x^C + \epsilon^C \quad [15]$$
$$\eta^C = \gamma_0^C + \gamma_1^C x^C + \delta^C \quad [16]$$
$$y^C: \begin{cases} \text{observed if } \eta^C > 0 \\ \text{not observed, otherwise} \end{cases} \quad [17]$$

Experimentals:
$$y^E = \mu + \alpha + \beta^E x^E + \epsilon^E \quad [18]$$
$$\eta^E = \gamma_0^E + \gamma_1^E x^E + \delta^E \quad [19]$$
$$y^E: \begin{cases} \text{observed if } \eta^E > 0 \\ \text{not observed, otherwise} \end{cases} \quad [20]$$

Independent random sampling is assumed from the two populations in this new model. The specification is completed by a choice of bivariate distributions for the two sets of error terms ϵ and δ, given x. The selection relations (equations 16 and 19) are the needed auxiliaries to the causal relations (equations 15 and 18), in order to obtain unbiased treatment effect estimates.

The random error terms δ^C and δ^E include variables other than x, influencing selection. It is an important advantage of selection modeling that such variables need not be explicitly included, as in the ANCOVA model. Given equations 15 through 20 as the true model, the ANCOVA covariate x in the analysis of equation 14 does not completely control for the existing selectivity.

Again, if $\beta^C = \beta^E$, it is reasonable to take α as a measure of the treatment effect. Hence, we need a technique to analyze data from the two groups simultaneously under selection models in which some parameters are constrained to be equal in the two groups. These equality constraints should not be taken for granted, however, but must be tested by means of the data.

With the education example, γ^E and the correlation between ϵ^E and δ^E are presumably both negative. Individuals with an unusually high

social disadvantage score, that is, a high δ^E value, are those more likely to be selected, and they are also the ones likely to have lower achievement scores, or low ϵ^E value. For the selectable experimentals, the true regression of y on x is then nonlinear and of the same shape as in Figure 3. Ignoring selectivity, ANCOVA for the experimentals and controls will give biased results. This will be studied further in Section 5.2, in a similar artificial data example.

Barnow et al. (1980) and Goldberger (1979) formulated a special form of selection, where for one experimental group and one control group,

$$y^E: \begin{cases} \text{observed if } \eta^E > 0 \\ \text{not observed, otherwise} \end{cases}$$

$$y^C: \begin{cases} \text{observed if } \eta^C \leqslant 0 \\ \text{not observed, otherwise} \end{cases}$$

where $\eta^E = \eta^C$. They illustrated their model by the well-known Head Start Compensatory Education program, so that y is the posttest achievement score. Here, a single selection variable (related to family income of the child) defines group membership. In terms of our model, their model implies that γ parameters and error covariance parameters differ only in signs between controls and experimentals. This specification seems too restrictive.

4. A GENERAL MODEL AND ITS ESTIMATION

In the previous section we formulated selection modeling for a single explanatory variable, x, and for one or two groups, the emphasis being on the basic ideas of the model. In this section we generalize the model to an arbitrary number of exogenous (explanatory) variables and to an arbitrary number of groups. The data will be regarded as sampled from G groups or populations, and for each group a univariate regression relation and a single selection relation is assumed. This formulation is chosen for simplicity; it may be generalized to a multivariate system (a structural equation system) for each group, to multivariate selection relations for each group, and also to categorical response variables.

It is essential to distinguish between two parts of the model: the causal relation and the selection relation. For each group g, g = 1, 2, . . ., G, we assume a causal relation of the form:

$$y^{(g)} = \underset{\sim}{\beta}^{(g)\prime}\underset{\sim}{x}^{(g)} + \epsilon^{(g)} \qquad [21]$$

where $\underset{\sim}{\beta}^{(g)}$ is a q × 1 vector of parameters, $\underset{\sim}{x}^{(g)}$ is a vector of random explanatory variables, and $\epsilon^{(g)}$ is a random residual, independent of $\underset{\sim}{x}^{(g)}$. We do not assume random sampling from the populations of equation 21. Instead, in addition to equation 21 we assume the selection relations for g = 1, 2, . . ., G

$$\eta^{(g)} = \underset{\sim}{\gamma}^{(g)\prime}\underset{\sim}{x}^{(g)} + \delta^{(g)} \qquad [22]$$

$$y^{(g)}: \begin{array}{l} \text{observed if } \eta > 0 \\ \text{not observed, otherwise} \end{array} \qquad [23]$$

where $\eta^{(g)}$ is a latent selection variable, $\underset{\sim}{\gamma}^{(g)}$ is a q × 1 vector of parameter, $\underset{\sim}{x}^{(g)}$ is as before, and $\delta^{(g)}$ is a random residual, independent of $\underset{\sim}{x}^{(g)}$. Let $\sigma^{(g)}_{\epsilon\epsilon}$, $\sigma^{(g)}_{\delta\epsilon}$, $\sigma^{(g)}_{\delta\delta}$ be the variance of $\epsilon^{(g)}$, the covariance between $\epsilon^{(g)}$ and $\delta^{(g)}$, and the variance of $\delta^{(g)}$, respectively. We assume for each group a bivariate normal distribution for $\epsilon^{(g)}$ and $\delta^{(g)}$, g = 1, 2, . . ., G. The groups are assumed to be independent, and for each group g we consider random sampling from the population given by equations 21, 22, and 23.

Each parameter of the model will be allowed to be any of three kinds: a free parameter, a parameter fixed to a certain value, or a parameter constrained to be equal to another parameter. For instance, both $\underset{\sim}{\beta}^{(g)}$ and $\underset{\sim}{\gamma}^{(g)}$ may contain parameters fixed to zero so that the same exogenous variables do not necessarily operate in equations 21 and 22. Also, group invariance of parameters can be tested by applying equality restrictions.

Group level differences are captured by $\beta^{(g)}$ parameters corresponding to unit x variables. The ANCOVA model, with possible group invariance of slopes and residual variances, is a special case of the above formulation, where $\sigma^{(g)}_{\delta\epsilon} = 0$ for g = 1, 2, . . ., G. Then the relation (equation 22) is inconsequential (see also the likelihood expressions that follow). A selection process may operate ($\sigma^{(g)}_{\delta\epsilon} \neq 0$) in one or more of the groups, and may operate differently in different groups.

Consider the bivariate distribution of y and η given $\underset{\sim}{x}$ and $\eta > 0$. For simplicity, the group index is omitted. Let $\phi(z; \mu, \sigma^2)$ denote the normal density for a variable z with mean μ and variance σ^2, let $\phi_{y\eta}$ denote the bivariate normal density for y and η given x, and let $\Phi(a)$ denote the probability that a standard normal variable falls below a. Let $\sigma_\delta = \sqrt{\sigma_{\delta\delta}}$.

The probability that $\eta > 0$, given $\underset{\sim}{x}$, may then be written as $\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1})$. The density for the singly truncated bivariate normal distribution for y and η, given $\underset{\sim}{x}$, is

$$p_{y\eta} = \begin{cases} 0, \text{ if } \eta \leqslant 0 \\ \phi_{y\eta}/\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1}), \text{ otherwise.} \end{cases} \quad [24]$$

From equation 24 we obtain the marginal distribution for y as

$$p_y = \int_0^\infty \phi_{y\eta}\, d\eta/\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1})$$

$$= \phi(y; \underset{\sim}{\beta}'\underset{\sim}{x}, \sigma_{\epsilon\epsilon}) \times \Phi(\mu_{\eta\cdot y}\sigma_{\eta\cdot y}^{-1})/\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1}) \quad [25]$$

where

$$\mu_{\eta\cdot y} = \underset{\sim}{\gamma}'\underset{\sim}{x} + \sigma_{\delta\epsilon}\,\sigma_{\epsilon\epsilon}^{-1}\,(y - \underset{\sim}{\beta}'\underset{\sim}{x}) \quad [26]$$

$$\sigma_{\eta\cdot y}^2 = \sigma_{\delta\delta} - \sigma_{\delta\epsilon}^2\,\sigma_{\epsilon\epsilon}^{-1} \quad [27]$$

Equation 25 gives the density of y, given $\underset{\sim}{x}$, in a truncated sample. The event $\eta > 0$ has probability one in such a sample. This means that population units with $\eta \leqslant 0$ cannot be included in the sample; not only do we not observe y, but we do not observe $\underset{\sim}{x}$ either. In the contrary case of a censored sample, a unit for which y is not observed ($\eta \leqslant 0$), given $\underset{\sim}{x}$, occurs with the probability $\Phi(-\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1})$. In this case, the density of y given $\underset{\sim}{x}$ when $\eta > 0$ is that of equation 25 except that the denominator cancels out.

The likelihood for both the truncated and the censored case may be summarized in the following way. For g = 1, 2, . . . , G, let $N^{(g)}$ denote

the sample size, let $N_t^{(g)}$ denote the number of sample units for which $\underset{\sim}{x}^{(g)}$ is observed, but not $y^{(g)}$, let

$$s^{(g)} = \begin{cases} 1, \text{ if selection occurs in group g} \\ 0, \text{ otherwise (random sample assumed)} \end{cases} \quad [28]$$

and let

$$t^{(g)} = \begin{cases} 1, \text{ if } \eta \leqslant 0\text{-units cannot occur in the} \\ \text{sample for group g (truncated sample)} \\ 0, \text{ otherwise (censored sample)} \end{cases} \quad [29]$$

The log likelihood for independent samples from the G groups may then be written

$$\log L = \sum_{g=1}^{G} \left[s^{(g)}(1 - t^{(g)}) \sum_{i=1}^{N_t^{(g)}} \log f_1(\underset{\sim}{x}_i^{(g)}) + \sum_{i=N_t^{(g)}+1}^{N^{(g)}} \log f_2(y_i^{(g)}, \underset{\sim}{x}_i^{(g)}) \right] \quad [30]$$

where

$$f_1(\underset{\sim}{x}_i^{(g)}) = \Phi(-\underset{\sim}{\gamma}^{(g)\prime} \underset{\sim}{x}_i^{(g)} \sigma_\delta^{-1}) \quad [31]$$

$$f_2(y_i^{(g)}, \underset{\sim}{x}_i^{(g)}) = \phi(y_i^{(g)}; \underset{\sim}{\beta}^{(g)\prime} \underset{\sim}{x}_i^{(g)}, \sigma_{\epsilon\epsilon}^{(g)}) \times \left\{ \Phi\left(\mu_{\eta \cdot y_i}^{(g)} \sigma_{\eta \cdot y_i}^{(g)-1} \right) \right\}^{s^{(g)}} \times \left\{ \Phi\left(\underset{\sim}{\gamma}^{(g)\prime} \underset{\sim}{x}_i^{(g)} \sigma_\delta^{(g)-1} \right) \right\}^{-t^{(g)} s^{(g)}} \quad [32]$$

Maximum-likelihood (ML) estimates are obtained from equation 30 in a straightforward fashion. The numerical optimization may

however be nontrivial, since the shape of the likelihood function can be complicated, yielding convergence problems. This may be particularly pressing in cases where the model does not fit well, with small samples or with poor starting values. In the censored case, reasonable starting values may be obtained from a separate Probit regression and an OLS regression in the truncated sample. The truncated case is relatively more difficult since separate Probit type information is not available for the estimation of the $\underset{\sim}{\gamma}$-parameters. We note that for the truncated case when $\sigma_{\delta\epsilon} = 0$ holds exactly, $\underset{\sim}{\gamma}$ will in fact be indeterminate.

Of some importance is the choice of parameterization in the actual computations. To ensure positive values for the variance expressions in the likelihood, we use the following parameterization. Consider the new parameter $\sigma^*_{\epsilon\epsilon}$, defined by $\sigma_{\epsilon\epsilon} = e^{\sigma^*_{\epsilon\epsilon}}$ yielding positive $\sigma_{\epsilon\epsilon}$. Also, the indeterminacy of the scale of η is used to set $\sigma^2_{\eta\cdot y} = 1$; that is, $\sigma_{\delta\delta}$ is not a free parameter, but is restricted as $\sigma_{\delta\delta} = 1 + \sigma_{\delta\epsilon} e^{-\sigma^*_{\epsilon\epsilon}}$. Hereby, all parameters in the optimization are free to vary from minus to plus infinity. In the actual reporting of the estimates, however, we find it convenient to revert to the more conventional parameterization with $\sigma_{\epsilon\epsilon}$ and $\sigma_{\delta\delta} = 1$. Also, instead of $\sigma_{\delta\epsilon}$, we will report the correlation between the errors, denoted by ρ. Standard errors will be given for this latter set of parameter estimates.

Let d_i be defined such that

$$\log L = \sum_{g=1}^{G} \sum_{i=1}^{N^{(g)}} d_i^{(g)}$$

and let

$$\underset{\sim}{A} = \sum_{g=1}^{G} \sum_{i=1}^{N^{(g)}} (\partial d_i^{(g)}/\partial \underset{\sim}{\theta} \times \partial d_i^{(g)}/\partial \underset{\sim}{\theta}')_{\underset{\sim}{\theta}=\underset{\sim}{\hat{\theta}}} \qquad [33]$$

where $\underset{\sim}{\hat{\theta}}$ is the ML estimate of the parameter vector $\underset{\sim}{\theta}$. The squared, asymptotic standard errors of $\hat{\theta}$ may then be found on the diagonal of $\hat{A}^{-1}$ (see also Griliches et al., 1978).

For the iterative optimization involved, the so-called FLEPOW algorithm is used (see Gruvaeus and Joreskog, 1970). This algorithm is based on a rapidly converging quasi-Newton method that makes use of first-order derivatives of the likelihood function, and a positive definite weight matrix $\underset{\sim}{E}$ that is built up during the iterations to approximate the

inverse of the Hessian matrix at the solution point. Starting values for the parameter estimates must be provided and also a starting value of $\underset{\sim}{E}$. For reasonably good starting values, a starting value of $\underset{\sim}{E}$ may be obtained by evaluating $\underset{\sim}{A}$ at that point and using $\underset{\sim}{E} = \underset{\sim}{A}^{-1}$. A similar algorithm was presented in Berndt et al. (1974) and has been applied, for example, in work by Hausman and Wise (1976, 1979).

5. ANALYSIS OF SIMULATED DATA

The aim of this section is to illustrate the models and selectivity issues discussed in previous sections by analysis of data sets generated from three different basic models. No more than two samples of different size will be drawn in each case, hence this study is more limited in scope than a rigorous Monte Carlo investigation. A similiar study was carried out by Wales and Woodland (1980) for the original Tobit model in the single group case.

5.1. SINGLE-GROUP DATA

Two basic models will be used here. Model 1 is specified with a single x,

$$y = 0.0 + 1.0x + \epsilon \qquad [34]$$

$$\eta = 0.0 - 1.0x + \delta \qquad [35]$$

$\sigma_{\delta\epsilon} = \sigma_{\delta\delta} = 1.0$, $\rho = -0.5$, and the mean and variance of x are chosen as $\mu_x = 0.0$, $\sigma_{xx} = 1.0$. Also, x is taken to be normally distributed.

Model 2 is specified with three x variables, where not all variables are operating in both the regression and selection relation,

$$y = 1.2 + 2.0x_1 + 1.0x_2 + 0.0x_3 + \epsilon \qquad [36]$$

$$\eta = 1.0 + 1.0x_1 + 0.0x_2 + 2.0x_3 + \delta \qquad [37]$$

where $\sigma_{\epsilon\epsilon} = 1.44$, $\sigma_{\delta\delta} = 1.0$, $\rho = 0.5$, and $\mu_{x_1} = \mu_{x_2} = \mu_{x_3} = 0.0$, $\sigma_{x_1x_1} = 1.0$, $\sigma_{x_2x_2} = 0.04$, $\sigma_{x_3x_3} = 0.01$, $\rho_{x_1x_2} = 0.8$, $\rho_{x_1x_3} = 0.3$, $\rho_{x_2x_3} = 0.5$. The x variables are taken to be trivariate normal.

Given the two basic models, trivariate and five-variate normal vectors corresponding to (y x η) for Model 1 and (y x_1 x_2 x_3 η) for Model 2 are generated. Two basic sample sizes are used in each case, N = 1000 and N = 4000. For each model and basic sample, a truncated subsample is created by including only selectable units, for which $\eta > 0$. In the corresponding censored case, the sample size is maintained as 1000 or 4000, where units with $\eta \leqslant 0$ are considered as nonselected, lacking observed y values. The truncated and censored cases may be viewed as corresponding to different real-life situations, where different amounts of data information are available.

For each basic model and sample size, several analyses are made. Using only the truncated sample, ordinary regression ignoring selection is reported. This is compared with the correct model formulation, that is, the truncated case of the respective basic model estimated by ML according to Section 4. With the censored sample case, ML Probit regression for the estimation of γ will be reported together with the Heckman estimator for β and ω. This is compared with the full model formulation, the censored case of the respective basic model, estimated by ML according to Section 4. We can also study the gain in precision of the estimates, comparing the truncated and censored case, estimated under the correct model formulation.

The results are given in Table 2 for Model 1 and in Table 3 for Model 2. For Model 1 there is strong selectivity, where only about half of the full population consists of selectable units. For Model 2 the corresponding figure is about three-quarters. Hence, we find overall more markedly biased estimates from ordinary regression for the first model. The columns "Truncated Case" and "Censored Case" give ML estimates in accordance with Section 4. In the truncated case, this estimator performs well, and the estimates are in no case more than twice the standard errors from the true values. For N = 1000, some of the standard errors are, however, rather large.

With information corresponding to the censored case, the Probit estimator for γ works extremely well in all cases. It is in fact comparable to the also high performance full information ML estimator (censored case), also with respect to precision in the estimates. The Heckman estimator for the β parameters performs very well, and is also comparable to the ML estimator. Note that $\sigma_{\epsilon\epsilon}$ is not consistently estimated (underestimated) and that the standard errors that are given are only approximate and too low, since these quantities are obtained via

TABLE 2
Parameter Estimates for Data Simulated According to Model 1*

Parameter	Population Value	Regression	Probit	Heckman	Truncated Case	Censored Case
			$N_t = 496$, $N = 1000$			
β_0	.0	−.373		.101	−.209	.074
		(.054)		(.278)	(.119)	(.179)
β_1	1.0	.788		1.048	.931	1.033
		(.052)		(.062)	(.095)	(.114)
$\sigma_{\epsilon\epsilon}$	1.0	.985		.979	.982	1.126
		(.065)		(.064)	(.076)	(.131)
ω				−.587		
				(.333)		
γ_0	.0		.001		.991	.013
			(.046)		(1.599)	(.046)
γ_1	−1.0		−1.033		−3.448	−1.040
			(.067)		(4.542)	(.068)
ρ	−.5			−.593	−.248	−.522
					(.413)	(.164)
			$N_t = 1963$, $N = 4000$			
β_0	.0	−.435		.002	−.223	.013
		(.027)		(.149)	(.137)	(.083)
β_1	1.0	.807		1.059	.965	1.065
		(.027)		(.088)	(.084)	(.054)
$\sigma_{\epsilon\epsilon}$	1.0	.916		.911	.978	1.054
		(.029)		(.028)	(.056)	(.062)
ω				−.539		
				(.183)		
γ_0	.0		.020		.851	.021
			(.023)		(.723)	(.023)
γ_1	−1.0		−1.040		−1.277	−1.043
			(.032)		(.346)	(.032)
ρ	−.5		−.542		−.521	−.538
					(.122)	(.078)

*Standard errors in parentheses.

TABLE 3
Parameter Estimates for Data Simulated According to Model 2*

Parameter	*Population Value*	*Regression*	*Probit*	*Heckman*	*Truncated Case*	*Censored Case*
			N_t = 239, N = 1000			
β_0	1.2	1.462 (.046)		1.270 (.123)	1.388 (.064)	1.192 (.076)
β_1	2.0	1.848 (.076)		1.979 (.106)	1.923 (.088)	2.033 (.092)
β_2	1.0	.397 (.396)		.415 (.397)	.450 (.431)	.414 (.415)
β_3	0.0	.442 (.477)		.710 (.497)	.482 (.537)	.834 (.529)
$\sigma_{\epsilon\epsilon}$	1.44	1.323 (.071)		1.318 (.070)	1.351 (.084)	1.475 (.115)
ω				.522 (.314)		
γ_0	1.0		1.007 (.060)		3.075 (.998)	1.009 (.057)
γ_1	1.0		.994 (.104)		2.126 (.922)	.982 (.107)
γ_2	0.0		−.090 (.497)		.038 (3.570)	−.009 (.509)
γ_3	2.0		1.897 (.603)		1.214 (3.652)	1.754 (.613)
ρ	0.5			.454	.670 (.263)	.602 (.113)
			N_t = 984, N = 4000			
β_0	1.2	1.432 (.023)		1.171 (.066)	1.265 (.074)	1.160 (.039)
β_1	2.0	1.773 (.039)		1.966 (.058)	1.942 (.073)	1.973 (.047)
β_2	1.0	1.049 (.197)		1.028 (.196)	.939 (.276)	1.030 (.204)
β_2	.0	−.156 (.251)		.280 (.268)	.058 (.346)	.303 (.268)
$\sigma_{\epsilon\epsilon}$	1.44	1.356 (.035)		1.348 (.034)	1.446 (.057)	1.506 (.055)
ω				.690 (.167)		

(continued)

TABLE 3 (Continued)

Parameter	Population Value	Regression	Probit	Heckman	Truncated Case	Censored Case
γ_0	1.0		.900		1.547	.991
			(.029)		(.478)	(.029)
γ_1	1.0		.994		1.352	1.000
			(.051)		(.307)	(.050)
γ_2	.0		−.083		−.779	−.104
			(.252)		(1.323)	(.246)
γ_3	2.0		2.054		1.502	2.023
			(.314)		(1.639)	(.308)
ρ	.5			.594	.590	.585
					(.093)	(.057)

*Standard errors in parentheses.

ordinary regression (see Heckman, 1979; with corrections in Stromsdorfer and Farkas, 1980: ch. 2, where a consistent estimator for $\sigma_{\varepsilon\varepsilon}$ and appropriate standard errors are given).

For Model 2 the zero population coefficients result in poorly estimated β_2, β_3 parameters for the N = 1000 case. Here, the regression relation is misspecified, since x_3 is included. It seems as if the nonlinear influence of x_3 via the selection relation is picked up by a linear term in the regression relation. Fixing $\beta_3 = 0$ gives an improved overall picture, with β_2 estimate in the censored case .713(.363).

In the N = 4000 cases, the gain in precision when using information from nonselected units is well reflected in the lower standard errors for the columns "Censored Case" as compared to the columns "Truncated Case." Going from N = 1000 to N = 4000, we would expect a reduction of standard errors to about half the size. This pattern holds for the censored case, but not for the truncated case; suggesting that the large-sample approximation of the standard errors is rather poor in the truncated case for the smaller sample sizes used.

We finally report the computing time for the ML estimator in the censored case and for Model 2, starting with the regression and Probit estimates, and $\rho = 0$. On the IBM 370/158, the time used was about two minutes for N = 1000 and about nine minutes for N = 4000.

5.2. DATA FOR TWO GROUPS

By means of a model for two groups, Model 3, we will now illustrate the Section 3 issues regarding the estimation of treatment effects using nonequivalent groups. We chose a model for a control group and an experimental group in line with the example of Section 3. For the control group, Model 3 states

$$y^C = -0.4 + 0.8x^C + \epsilon^C \quad [38]$$

and that random sampling is the case. Here, $\mu_x^C = 0.0$, $\sigma_{xx}^C = 1.0$, $\sigma_{\epsilon\epsilon}^C = 0.9$. For experimentals, Model 3 is the same as Model 1,

$$y^E = 0.0 + 1.0x^E + \epsilon^E \quad [39]$$

$$\eta^E = 0.0 - 1.0x^E + \delta^E \quad [40]$$

with means, variances, and covariances as before. In fact, the same sample will be used for this group. We will only study the case of $N^{(g)} = 4000$, where $N^{(g)}$ is the basic sample size in each of the groups. We may view the full population regressions (equations 38 and 39) in the following way. We first consider a single parent population. Treatment produces two new subpopulations, one for controls, which is the same as before; while, as is seen for experimentals, the treatment affects both the intercept and the slope. Thus, there is a positive true treatment effect for large x values ($x > -2$), and the effect increases with increasing x.

We first study ANCOVA, which, ignoring selectivity, is carried out on the truncated sample. In this case, there are still 4000 controls, but only 4000 - 1963 = 2037 experimentals. Ordinary ANCOVA assumes group-invariant slopes and residual variances, so that the treatment effect is taken to be the difference in intercepts. Testing this invariance hypothesis for the data at hand, we obtain $\chi^2(2) = .047$, using a standard likelihood-ratio test. Due to selectivity, ANCOVA is therefore unable to reject this hypothesis. The results are presented in Table 4. The treatment effect, $\beta_0^E - \beta_0^C$, is estimated as $-.016$, but is not significant, $\chi^2(1) = .356$. Hence, selection of the most needy ones to the experimental group masks the positive true treatment effect. This is because the ANCOVA covariate, x, does not completely control for the non-

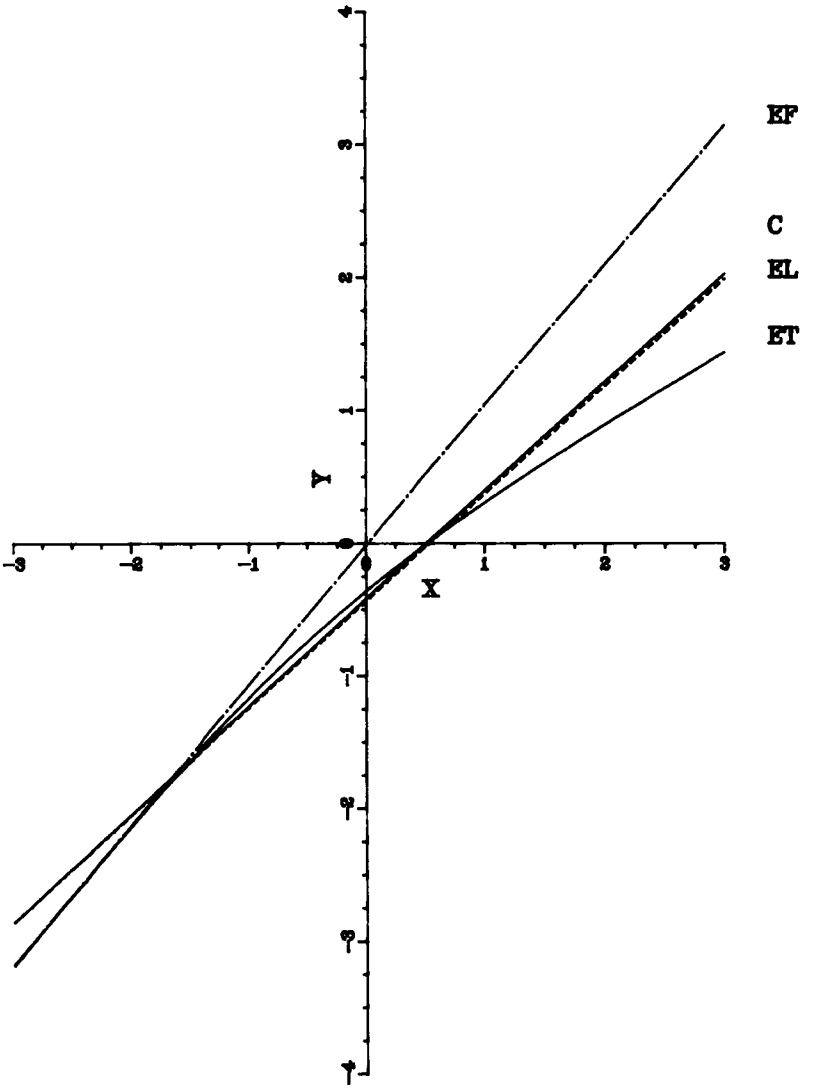
TABLE 4
Parameter Estimates for Control Group and Experimental Group Data Simulated According to Model 3*

Parameter	Population Value	ANCOVA	Truncated Case	Censored Case
		Controls		
β_0^C	−.4	−.416 (.015)	−.416 (.015)	−.416 (.015)
β_1^C	.8	.812 (.013)	.813 (.014)	.813 (.014)
$\sigma_{\epsilon\epsilon}^C$	.9	.915 (.017)	.914 (.020)	.914 (.020)
		Experimentals		
β_0^E	.0	−.432 (.023)	−.223 (.137)	.013 (.083)
β_1^E	1.0	.812 (.013)	.965 (.084)	1.065 (.054)
$\sigma_{\epsilon\epsilon}^E$	1.0	.915 (.017)	.978 (.056)	1.054 (.060)
γ_0^E	.0		.851 (.723)	.021 (.023)
γ_1^E	−1.0		−1.277 (.346)	−1.043 (.032)
ρ^E	−.5		−.521 (.122)	−.538 (.078)

*Standard errors in parentheses.

equivalence of the groups due to selectivity. Figure 5 shows this situation graphically using the different estimated regressions.

Allowing for selectivity in the experimental group, the ML estimator of Section 4 was applied to the same data. The test of group-invariant slopes and error variances resulted in $\chi^2(2) = 7.002 (p < .05)$ for the truncated case and $\chi^2(2) = 13.462 (p < .005)$ in the censored case. In both analyses, the hypothesis is correctly rejected. The estimates are given in the two right-most columns of Table 4. In both cases the estimated regression lines correspond well to the true lines.



EF = Regression line for experimentals, estimating the full population regression.
ET = Regression curve for experimentals, estimating the regression in the selected subpopulation.
EL = Regression line for experimentals, estimating the linear approximation to the regression in the selected subpopulation.
C = Regression line for control, estimating the full population.

Figure 5: Estimated Regressions for Control Group and Experimental Group Data Simulated According to Model 3

6. LATENT EXOGENOUS VARIABLES

We will now discuss the analysis of selective samples in the context of latent (unobserved) variable models. A general approach to the study of latent variable models has been given, for example, by Jöreskog (1977; see also Sörbom and Jöreskog, 1981). For simplicity we will here limit ourselves to the case where the latent variables occur on the right-hand side in the regression relation of interest.

6.1. GENERAL RESULTS ON SELECTION

It will be useful to review some classical results on selection in multivariate distributions due to Pearson (1912) and Lawley (1943-1944), and applied to factor analysis models by Meredith (1964).

Pearson and Lawley considered influences of selection on a random vector variable $\underset{\sim}{z}$, say. For a set of selection variables, here denoted by $\underset{\sim}{\eta}$, selection is of a general type, changing the density of $\underset{\sim}{z}$, p_z, into p_z^*. Given that the regression of $\underset{\sim}{z}$ on $\underset{\sim}{\eta}$ in the total population is linear and homoscedastic, it is shown that

$$\underset{\sim}{\mu}_z = \underset{\sim}{\mu}_z^* - \underset{\sim}{\Sigma}_{z\eta}^* \underset{\sim}{\Sigma}_{\eta\eta}^{*-1} (\underset{\sim}{\mu}_\eta^* - \underset{\sim}{\mu}_\eta) \qquad [41]$$

$$\underset{\sim}{\Sigma}_{zz} = \underset{\sim}{\Sigma}_{zz}^* - \underset{\sim}{\Sigma}_{z\eta}^* (\underset{\sim}{\Sigma}_{\eta\eta}^{*-1} - \underset{\sim}{\Sigma}_{\eta\eta}^{*-1} \underset{\sim}{\Sigma}_{\eta\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{*-1}) \underset{\sim}{\Sigma}_{\eta z}^* \qquad [42]$$

$$\underset{\sim}{\Sigma}_{\eta z} = \underset{\sim}{\Sigma}_{\eta\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{*-1} \underset{\sim}{\Sigma}_{\eta z}^* \qquad [43]$$

where we use the general notation $\underset{\sim}{\mu}_u$ for the mean vector of the random variable u and $\underset{\sim}{\Sigma}_{uv}$ for the covariance matrix of the random vectors $\underset{\sim}{u}$ and $\underset{\sim}{v}$. Quantities with asterisks refer to the distribution in the subpopulation of selectables and the corresponding quantities without asterisks to the total population. This gives

$$\underset{\sim}{\mu}_z^* = \underset{\sim}{\mu}_z + \underset{\sim}{\Sigma}_{z\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{\cong 1} (\underset{\sim}{\mu}_\eta^* - \underset{\sim}{\mu}_\eta) \qquad [44]$$

$$\underset{\sim}{\Sigma}_{zz}^* = \underset{\sim}{\Sigma}_{zz} + \underset{\sim}{\Sigma}_{z\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{-1} (\underset{\sim}{\Sigma}_{\eta\eta}^* - \underset{\sim}{\Sigma}_{\eta\eta}) \underset{\sim}{\Sigma}_{\eta\eta}^{-1} \underset{\sim}{\Sigma}_{\eta z} \qquad [45]$$

We note that the selection situations studied in the previous sections are of this type. We considered the conditional distribution of y and η

for given $\underset{\sim}{x}$. Due to the bivariate normality of the errors, linearity and homoscedasticity is ensured, so that the mean and variance of y, given $\underset{\sim}{x}$ and $\eta > 0$, can be obtained by equations 44 and 45.

Now consider the factor analysis model (see Lawley and Maxwell, 1971)

$$\underset{\sim}{z} = \underset{\sim}{\nu} + \underset{\sim}{\Lambda}\underset{\sim}{\xi} + \underset{\sim}{\zeta} \qquad [46]$$

where $\underset{\sim}{\nu}$ is a vector of location parameters, $\underset{\sim}{\Lambda}$ is a matrix of factor loadings, $\underset{\sim}{\xi}$ is a vector of latent factors, and $\underset{\sim}{\zeta}$ is a vector of residuals (unique variables or measurement errors). Following Sörbom (1974), let $\underset{\sim}{\mu}_{\xi} = \underset{\sim}{\theta}$. With the usual assumptions.

$$\mu_z = \nu + \Lambda\theta \qquad [47]$$

$$\Sigma_{zz} = \Lambda\Phi\Lambda' + \Psi \qquad [48]$$

where Φ is the covariance matrix of the factors, and Ψ is the covariance matrix of the residuals, usually assumed to be diagonal.

Assume that a set of selection variables $\underset{\sim}{\eta}$ are directly related to $\underset{\sim}{\xi}$ only, but not to $\underset{\sim}{\zeta}$ or $\underset{\sim}{z}$ ($\underset{\sim}{\eta}$ is indirectly related to $\underset{\sim}{z}$). Applying the Pearson-Lawley formulas, it is found that the factor analysis model holds in the selected population and that $\underset{\sim}{\nu}$, $\underset{\sim}{\Lambda}$, and $\underset{\sim}{\Psi}$ are unaffected by the selection (see Meredith, 1964; Olsson, 1978), and that

$$\underset{\sim}{\mu}^*_z = \underset{\sim}{\nu} + \underset{\sim}{\Lambda}\underset{\sim}{\theta}^* \qquad [49]$$

$$\underset{\sim}{\Sigma}^*_{zz} = \underset{\sim}{\Lambda}\underset{\sim}{\Phi}^*\underset{\sim}{\Lambda}' + \underset{\sim}{\Psi} \qquad [50]$$

where

$$\underset{\sim}{\theta}^* = \underset{\sim}{\theta} + \underset{\sim}{\Sigma}_{\xi\eta}\underset{\sim}{\Sigma}^{-1}_{\eta\eta}(\underset{\sim}{\mu}^*_{\eta} - \underset{\sim}{\mu}_{\eta}) \qquad [51]$$

$$\underset{\sim}{\Phi}^* = \underset{\sim}{\Phi} + \underset{\sim}{\Sigma}_{\xi\eta}\Sigma^{-1}_{\eta\eta}(\underset{\sim}{\Sigma}^*_{\eta\eta} - \underset{\sim}{\Sigma}_{\nu\nu})\Sigma^{-1}_{\eta\eta}\underset{\sim}{\Sigma}_{\eta\xi} \qquad [52]$$

Invariance properties of this kind are utilized in multiple-group factor analyses and structural equation modeling where different subpopulations are compared in a simultaneous analysis (see Jöreskog, 1971; Sörbom, 1974; Jöreskog and Sörbom, 1980; and Sörbom and Jöreskog, 1981).

6.2. SELECTION MODELING WITH LATENT EXOGENOUS VARIABLES

In Sörbom (1978, 1981; see also Sörbom and Jöreskog, 1981), the multiple-group factor analysis is extended to handle ANCOVA situations with latent variables. Of particular interest here is the case where the covariates (the exogenous variables) are imperfectly measured, assuming a factor analytic measurement model. Allowing for measurement error in the covariates avoids biased results (e.g., Reichardt, 1979, and references therein). We will consider a simple case of this type of model and introduce the added complication of selective samples.

Consider the following model for groups g = 1, 2, . . . , G,

$$y^{(g)} = \beta_0^{(g)} + \underset{\sim}{\beta}^{(g)\prime}\underset{\sim}{\xi}^{(g)} + \epsilon^{(g)} \qquad [53]$$

$$\eta^{(g)} = \gamma_0^{(g)} + \underset{\sim}{\gamma}^{(g)\prime}\underset{\sim}{\xi}^{(g)} + \delta^{(g)} \qquad [54]$$

$$\underset{\sim}{x}^{(g)} = \underset{\sim}{\nu} + \underset{\sim}{\Lambda}\underset{\sim}{\xi}^{(g)} + \underset{\sim}{\zeta}^{(g)} \qquad [55]$$

$$y^{(g)}, \underset{\sim}{x}^{(g)}: \begin{cases} \text{observed, if } \eta^{(g)} > 0 \\ \text{not observed, otherwise} \end{cases} \qquad [56]$$

where $\epsilon^{(g)}$ and $\delta^{(g)}$ have covariance $\sigma_{\delta\epsilon}^{(g)}$, and both $\epsilon^{(g)}$ and $\delta^{(g)}$ are assumed to be independent of $\xi^{(g)}$ and $\zeta^{(g)}$. Here, equation 55 shares the assumptions of equation 46. Note that $\underset{\sim}{x}^{(g)}$ is included in equation 56, since in this case the model also restricts the marginal distribution of $\underset{\sim}{x}^{(g)}$, so that we consider the joint $y^{(g)}$, $\underset{\sim}{x}^{(g)}$ distribution, not only the conditional distribution of $y^{(g)}$ given $\underset{\sim}{x}^{(g)}$, as before. Here we consider the truncated case only.

With $\sigma_{\delta\epsilon}^{(g)} = 0$ for g = 1, 2, . . . , G, relation (equation 54) is inconsequential and we obtain a special case of Sörbom (1978). The measurement model (equation 55) is assumed to have group-invariant $\underset{\sim}{\nu}$ and $\underset{\sim}{\Lambda}$. The groups may consist of a control group and several treatment groups. The latent variable vector $\underset{\sim}{\xi}^{(g)}$ contains the covariates. Given group-invariant slope parameters $\underset{\sim}{\beta}^{(g)}$, treatment effects are obtained as differences in the $\beta_0^{(g)}$ (g = 1, 2, . . . , G) parameters.

Now consider the full model given by equations 53 through 56 for each group g. Note that the specification allows separate exogenous variables to operate in equations 53 and 54, and that those in equation

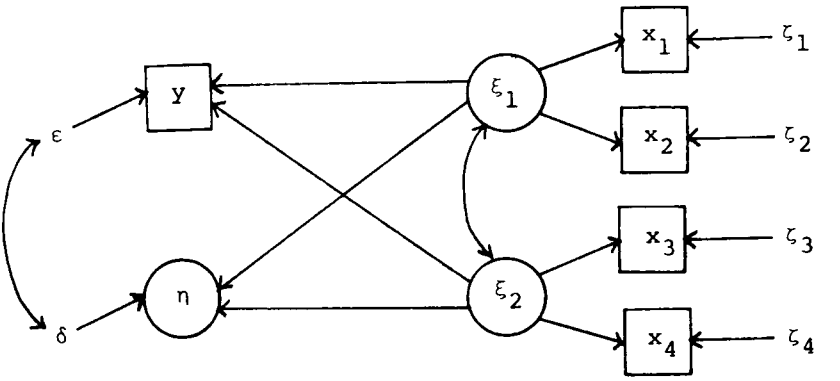


Figure 6: Path Model for Selection in the Presence of Two Latent Exogenous Variables

54 can be specified to be directly observed variables measured without error. For each group g, there is selectivity in the structural regression (equation 53) of $y^{(g)}$ on $\xi^{(g)}$ whenever the error covariance $\sigma_{\delta\epsilon}^{(g)}$ is nonzero.

We will illustrate the selectivity issues by means of the model of Figure 6, where squares denote observed variables and circles denote latent variables.

To continue the example of Section 3, y and ξ_1 represent achievement post- and pretest scores, where for simplicity the posttest is taken to be measured without error, while ξ_1 is the pretest true score. (Random measurement error in y causes no bias since it is absorbed in the residual ϵ.) Here, ξ_2 may represent, say, true home background score. True scores rather than actually observed scores are assumed to influence the selection variable η. Assume that the prerequisites for the Pearson-Lawley formulas hold in a certain population, for example, by multivariate normality for all variables involved. Consider the analysis of a random sample from the subpopulation $\eta > 0$, that is, a selective, truncated, sample from the full population. We note that by the Pearson-Lawley results, the factor analysis model (the measurement model) for the marginal distribution of x_1, x_2, x_3, x_4 in this subpopulation will hold with invariant $\underline{\nu}$, $\underline{\Lambda}$, and $\underline{\Psi}$, since η is indirectly related to $x_1 - x_4$ via ξ_1, ξ_2. In the distribution of y, x_1, x_2, x_3, x_4, we note that y can be considered as an additional measure of both ξ_1 and ξ_2 in a five-variate factor analysis model. Since the selection variable is directly related to y, this model will not hold in the subpopulation of

selectables. Estimating the structural regression of y on ξ_1 and ξ_2 by the methods of Sörbom (1978), Sörbom and Jöreskog (1981) gives biased results in a way analogous to previous sections.

A simple ad hoc estimator using the methods of Section 4 seems possible, however. Since the measurement model for $x_1 - x_4$ holds in the subpopulation, we may use the truncated sample to estimate factor scores $\hat{\xi}_1$ and $\hat{\xi}_2$, properly scaled to approximate the covariance matrix Φ^* (see Lawley and Maxwell, 1971: ch. 8). In a second step, y and η are regressed on $\hat{\xi}_1$ and $\hat{\xi}_2$ according to the truncated case of the selection model in Section 4. This enables us to obtain an approximate test for selectivity and approximately estimate the full population regression coefficients.

If, however, any of the observed exogenous variables $x_1 - x_4$ influences η directly (see also Goldberger, 1972a), the measurement model will also be incorrectly specified. Hence, ignoring selectivity may give seriously biased results from analyses by the methods of Sörbom (1978), Sörbom and Jöreskog (1981).

With the model of equations 53 through 56, any of the above selection situations can be specified.

7. CONCLUSION

In this article we have shown the statistical and computational feasibility of correctly analyzing selective samples by selection modeling. Analysis methods originally proposed in econometric studies have been shown to be of potential use in general quasi-experimental studies, particularly regarding selectivity in the context of treatment effect evaluation with nonequivalent groups. In such evaluations, randomization is seen not to be essential to unbiased treatment effect estimation.

> Indeed, the critical difference for avoiding bias is not whether the assignments are random or nonrandom, but whether the investigator has *knowledge of and can model* this selection process [Cain, 1975: 304].

This suggests that the investigator should gather extensive information on the selection processes involved, and seek to be in control of

them by systematic selection in a consistent manner. For instance, in the context of our model, we have seen the benefits of using censored sample information rather than truncated sample information. Say that nonrandom selection into the experimental (treatment) group is desirable from an ethical point of view. With the language of our model we may take an initial random sample for which $\underset{\sim}{x}$ is observed. From this sample, units can be selected in a nonrandom but consistent way, resulting in censored sample information. Indeed, we may consider the selectivity problem in the opposite way. Instead of trying to *find* the correct selection model, we could select *according to* a prescribed model, attempting to determine by design the selection variable η in terms of a set of background variables $\underset{\sim}{x}$. Of course, the selectivity problem of attrition will remain.

Admittedly, the selection modeling of this article may certainly be an oversimplification for many practical quasi-experimental situations. Selectivity problems, however, seem unavoidable by design, implying that more flexible statistical specifications should be investigated.

REFERENCES

AMEMIYA, T. (1982) Tobit Models: A Survey. Palo Alto, CA: Rhodes Associates.

———(1973) "Regression analysis when the dependent variable is truncated normal." Econometrica 41, 6: 997-1016.

BARNOW, B. S., G. G. CAIN, and A. S. GOLDBERGER (1980) "Issues in the analysis of selection bias, in E. Stromsdorfer and G. Farkas (eds.) Evaluation Studies Review Annual, Vol. 5. Beverly Hills, CA: Sage.

BERNDT, E. B., B. HALL, R. HALL, and J. A. HAUSMAN (1974) "Estimation and inference in non-linear structural models." Annals of Economic and Social Measurement 3: 653-656.

CAIN, G. C. (1975) "Regression and selection models to improve nonexperimental comparisons," pp. 297-317 in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment, Some Critical Issues in Assessing Social Programs. New York: Academic Press.

CRAGG, J. G. (1971) "Some statistical models for limited dependent variables with application to the demand for durable goods." Econometrica 39: 829-844.

GOLDBERGER, A. S. (1980) "Abnormal selection bias." Social Systems Research Institute, University of Wisconsin, Madison.

———(1979) "Methods for eliminating selection bias." Department of Economics, University of Wisconsin, Madison.

———(1972a) "Selection bias in evaluating treatment effects: some formal illustrations." Discussion paper 123-72. Madison, WI: Institute for Research on Poverty.

———(1972b) "Selection bias in evaluating treatment effects: the case of interaction." Discussion paper 129-72. Madison, WI: Institute for Research on Poverty.

GRILICHES, Z., B. H. HALL, and J. A. HAUSMAN (1978) "Missing data and self-selection in large panels." Annales de l'INSEE 30-31: 137-176.

GRONAU, R. (1974) "Wage comparisons—a selectivity bias." J. of Pol. Economy 82: 1119-1144.

GRUVAEUS, G. T. and K. G. JÖRESKOG (1970) "A computer program for minimizing a function of several variables." Research Bulletin 70-14. Princeton, NJ: Educational Testing Service.

HAUSMAN, J. A. and D. A. WISE (1979) "Attrition bias in experimental and panel data: the Gary income maintenance experiment." Econometrica 47: 455-474.

———(1976) "The evaluation of results from truncated samples: the New Jersey income maintenance experiment." Annals of Economic and Social Measurement 5: 421-445.

HECKMAN, J. (1979) "Sample selection bias as a specification error." Econometrica 47: 153-161.

———(1977) "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)." University of Chicago. (mimeo)

———(1974) "Shadow prices, market wages, and labor supply." Econometrica 42: 679-694.

HURD, M. (1979) "Estimation in truncated samples when there is heteroscedasticity." J. of Econometrics 11: 247-258.

JOHNSON, N. and S. KOTZ (1972) Distributions in Statistics: Continuous Multivariate Distributions. New York: John Wiley.

JÖRESKOG, K. G. (1977) "Structural equation models in the social sciences: specification, estimation and testing," pp. 265-286 in P. R. Krishnaiah (ed.) Applications of Statistics. Amsterdam: North-Holland.

———(1971) "Simultaneous factor analysis in several populations." Psychometrika 36: 409-426.

———and D. SÖRBOM (1980) Simultaneous Analysis of Longitudinal Data from Several Cohorts. Research Report 80-5. Department of Statistics, University of Uppsala.

KENNY, L. W., L. LEE, G. S. MADDALA, and R. P. TROST (1979) "Returns to college education: an investigation of self-selection bias based on the Project Talent data." Int. Econ. Rev. 20, 3: 775-789.

LAWLEY, D. N. (1943-1944) "A note on Karl Pearson's selection formulae." Proceedings of the Royal Society Edinburgh, Section A (Mathematics and Physics Section) 62, 1: 28-30.

——and A. E. MAXWELL (1971) Factor Analysis as a Statistical Method. London: Butterworth.

LEWIS, H. G. (1974) "Comments on selectivity biases in wage comparisons." J. of Pol. Economy: 1145-1155.

MADDALA, G. S. and L-F. LEE (1976) "Recursive models with qualitative endogenous variables." Annals of Economic and Social Measurement 5: 525-545.

MEREDITH, W. (1964) "Notes on factorial invariance." Psychometrika 29: 177-185.

NELSON, F. D. (1979) "The effect of and a test for misspecification in the censored-normal model." Social Science Working Paper 291, California Institute of Technology.

OLSEN, R. J. (1979) "Tests for the presence of selectivity bias and their relation to specifications of functional form and error distribution." Working paper 812, Yale University.

OLSSON, U. (1978) "Selection bias in confirmatory factor analysis." Research Report 78-4. Department of Statistics, University of Uppsala.

PEARSON, K. (1912) "On the general theory of the influence of selection on correlation and variation." Biometrika 8: 437-443.

RAY, S. C., R. A. BERK, and W. T. BIELBY (1980) "Correcting sample selection bias for bivariate logistic distribution of disturbances." University of California.

REICHARDT, C. S. (1979) "The statistical analysis of data from non-equivalent group designs," in T. D. Cook and D. T. Campbell (eds.) Quasi-Experimentation: Design & Analysis for Field Settings. Chicago: Rand McNally.

SÖRBOM, D. (1981) "Structural equation models with structured means," in K. G. Jöreskog and H. Wold (eds.) Systems Under Indirect Observation: Causality, Structure, Prediction. Amsterdam: North-Holland.

———(1978) "An alternative to the methodology for analysis of covariance." Psychometrika 43: 381-396.

———(1974) "A general method for studying differences in factor means and factor structures between groups." British J. of Mathematical and Statistical Psychology 27: 229-239.

———and K. G. JÖRESKOG (1981) "The use of structural equation models in evaluation research." Presented at the conference on Experimental Research in Social Sciences, Gainesville, Florida, January 8-10.

STROMSDORFER, E. and G. FARKAS (1980) Evaluation Studies Review Annual, Vol. 5. Beverly Hills, CA: Sage.

THORNDIKE, R. L. (1942) "Regression fallacies in the matched groups experiment." Psychometrika 7: 85-102.

TOBIN, J. (1958) "Estimation of relationships for limited dependent variables." Econometrica 26: 24-36.

TUNALI, F. I., J. R. BEHRMAN, and B. L. WOLFE (1980) "Identification, estimation and prediction under double selection." Presented at the 1980 Joint Statistical Meetings of American Statistical Association and Biometric Society, Houston, Texas.

VENTI, S. and D. A. WISE (1980) "Test scores, educational opportunities, and individual choice." Discussion Paper Series, Kennedy School of Government, Harvard University.

WALES, T. J. and A. D. WOODLAND (1980) "Sample selectivity and the estimation of labor supply functions." Int. Econ. Rev. 21, 2: 437-468.

WEISBERG, H. I. (1979) "Statistical adjustments and uncontrolled studies." Psych. Bull. 86: 1149-1164.

WILLIS, R. J. and S. ROSEN (1979) "Education and self-selection." J. of Pol. Economy 87, 5: 7-36.

Bengt Muthén is Assistant Professor of Education at the Graduate School of Education, UCLA. Dr. Muthén was formerly a researcher in the Jöreskog group of the Statistics Department, University of Uppsala, where he earned his Ph.D. in Statistics. Dr. Muthén's

research involves latent variable structural equation modeling with categorical data, including the factor analysis of dichotomous variables.

Karl G. Jöreskog is Professor and Chairman of the Department of Statistics at the University of Uppsala. He is a past president of the Psychometric Society and has published numerous articles on factor analysis, covariance structure analysis, structural equation models, and analysis of longitudinal data.