10 Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment

Jerry A. Hausman and David A. Wise

10.1 Introduction

Unbiased parameter estimates, although illusory, are thought by many researchers to be the primary objective of empirical analysis in the social sciences in general and of econometric analysis in particular. In a technical sense unbiased estimation of the parameters of a behavioral model requires that the independent measured variables of the specification be uncorrelated with unmeasured variables not explicitly accounted for in the analysis, rather captured only in spirit through a stochastic (or error) term. We normally think of correlation between independent variables and the error term as arising from improperly excluded variables, simultaneous relationships, or inaccurately measured independent variables. But such correlation may also be artificially induced, often unintended, through sample selection. Sample selection is not always random; in fact it is often systematically nonrandom. Stratification based on endogenous variables is a prevalent example.

Individual data in the social sciences are often collected by survey. The selection of persons to be surveyed is often based on a stratified sample design, with random sampling within strata. The proportions of observations within strata—for example, defined by levels of income and education—do not necessarily reflect population proportions, as they would if the sample were selected randomly from the population at large. In general this does not pose a problem for empirical analysis based on survey data if stratification is based on exogenous variables only. That is, unbiased estimates of behavioral parameters may be obtained, for example, by standard regression techniques.[1] But often variables considered as endogenous to the model whose parameters are to be estimated are also the basis for stratification.

The research was performed pursuant to contract number HEW 100-76-0073 from the Department of Health, Education, and Welfare, Washington, D.C. The opinions and conclusions expressed herein are solely those of the authors and should not be construed as representing the opinions or policy of any agency of the United States government.

This study was part of continuing analysis of the Gary experiment at Mathematica Policy Research. The authors also acknowledge research support of the National Science Foundation. Research assistance was provided by G. Burtless. We have benefited from comments by Charles Manski, John Pratt, and Roy Radner and from the comments of two referees for this volume.

1. There may, however, be questions about extrapolation of the results beyond the sample range of independent variables.

Many major surveys, some conducted in conjunction with social experiments, are characterized by endogenous stratification.[2] The selection of participants in the New Jersey negative income tax experiment is an extreme example; see Hausman and Wise (1977a). No families with incomes greater than one and one-half times the poverty level were sampled. The income maintenance experiments in Gary and Seattle–Denver followed a less extreme selection procedure. Although higher income families were not excluded from these experiments, they were undersampled. In Gary this was particularly true of families with incomes greater than 2.4 times the poverty level. Even below that level families were grouped into intervals defined by multiples of the poverty level. Sampling proportions within the intervals did not necessarily reflect population proportions.[3]

We present in this chapter alternative methods of correcting for endogenous sampling in order to obtain consistent estimates of population parameters. If sample-versus-population proportions within strata are known, either weighted least squares or a more efficient maximum likelihood procedure may be used. If these proportions are not known, they may be estimated along with behavioral parameters using our proposed maximum likelihood procedure. Although possibly not immediately transparent, it should become clear that the methods we propose here are conceptually parallel to the estimation procedures proposed by Manski and Lerman (1977) and Manski and McFadden (chapter 1) under conditions of choice-based sampling. But our procedures are directed toward estimation with continuous endogenous variables, while theirs are directed at estimation in discrete (or qualitative) choice situations. The underlying problem—the likelihood that an observation is in the sample depends on the value of an endogenous (or outcome) variable—is the same, however.

2. The 1967 survey of economic opportunity also undersampled high income families. So did the University of Michigan panel study of income dynamics that resurveyed part of the survey of economic opportunity sample. The use of any of these data sets to estimate behavioral relationships that treat earnings or components of earnings (wages and hours worked) as endogenous variables will lead to biased and inconsistent estimates of population parameters.

3. In attempting to estimate the treatment effect of this experiment, we found that there were two potentially serious statistical problems: one was attrition and the other, sample selection. We found that either of these problems could be handled individually without undue complication but that treating them simultaneously, although conceptually straightforward, would present a somewhat complicated estimation problem. Thus we have used preexperimental (baseline) data, before attrition became a matter of concern, in this chapter. A primary goal was to see whether or not correction for sample selection

10.2 The Problem of Endogenous Sampling and Estimation Methods

We shall focus the conceptual formulation of the problem on the sample selection procedure followed in the Gary income maintenance experiment. The proposed method of estimation is in no way peculiar to this experiment.

Approximately 2,600 families were drawn at random from certain geographic areas in Gary, Indiana, for the experiment.[4] But only about 1,800 families actually participated. To select the 1,800, families were stratified by income as well as by exogenous variables. Five income intervals were defined by multiples of the poverty level, and families were selected at random from within the intervals; but the within interval totals were not intended to reflect population proportions.

While for estimation purposes it is necessary to assume a precise formulation of the sampling procedure, we do not in fact have a precise description of the process followed in the Gary experiment. Consequently we obtained approximate accounts from persons knowledgeable about the early phases of the experiment and assumed a process that we think represents a good approximation to the true procedure: (1) a family was selected at random, (2) the family was classified according to five income intervals, (3) the family was retained in the sample with some probability that depended on its income interval, and (4) this procedure was followed until a sample size of 1,800 was obtained. We assume that the 1,800 was fixed by the sample design rather than the 2,600. There are several other reasonable possibilities. We shall develop a statistical model and obtain empirical estimates based on these assumptions. Then we shall discuss other plausible procedures and statistical models that correspond to them.

lead to parameter estimates that were substantially different from those obtained without correction. We previously found that the extreme form of sample selection in the New Jersey experiment lead to seriously biased estimates of behavioral parameters (see Hausman and Wise 1976, 1977). The much less severe sampling procedure followed in the Gary experiment, however, does not seem to produce large bias in parameter estimates. Thus we have concluded that evaluation of experimental results without explicit corrections for sample selection would not in this case yield substantial inaccuracies. In particular we have in another paper proposed a method of correcting for attrition bias and have presented estimates based on it that do not at the same time correct for sample selection bias (see Hausman and Wise 1977b).

4. All the families were black, and there had to be at least one dependent under the age of 18 present in the household. The majority of the families were headed by feinales. For more detail see Kehrer et al. (1975).

We shall begin by assuming only two income intervals. Assume that in the population, income $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, with Y given $\mathbf{X}$ distributed normally with mean $\mathbf{X}\boldsymbol{\beta}$, variance σ^2, and density function denoted by $f(Y \mid \mathbf{X})$. Assume that below some level L a proportion P_1 of a random sample of the population is in fact sampled and above L, a proportion P_2. (Note that with purely random sampling these proportions would have expected value equal to one.) These values may be thought of as the probabilities of retaining randomly sampled values. The density function h of Y given $\mathbf{X}$ in the sample can be written as

$$h(y) = \begin{cases} \dfrac{P_1 \cdot f(y)}{P_1 \cdot Pr[Y \le L] + P_2 \cdot Pr[Y > L]}, & \text{if } y \le L, \\ \dfrac{P_2 \cdot f(y)}{P_1 \cdot Pr[Y \le L] + P_2 \cdot Pr[Y > L]}, & \text{if } y > L, \end{cases} \tag{10.1}$$

where f is the normal density function $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. The distribution of Y for any given $\mathbf{X}$, say $\mathbf{X}^*$, would not be smooth like that of the normal. The distribution might look something like the one in figure 10.1, where the solid line represents a normal distribution and the dashed line the distribution in our sample. There is a discontinuity at the point L with

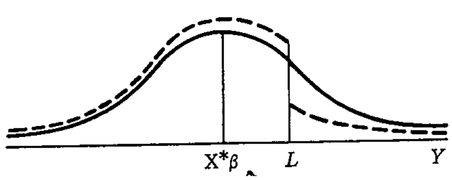

Figure 10.1

greater density relative to the normal below L and less density above. The denominator in (10.1) can be thought of as a normalizing constant, assuring that the integral over $h(y)$ with respect to y is one. Note that we cannot identify both P_1 and P_2, only their ratio. We divide through by P_1 and let $P_2/P_1 = P$ and rewrite the probabilities in the denominator, giving

$$h(y) = \begin{cases} \dfrac{f(y)}{\int_{-\infty}^{L} f(y)dy + P \cdot \int_{L}^{\infty} f(y)dy}, & \text{if } y \leq L, \\[2ex] \dfrac{P \cdot f(y)}{\int_{-\infty}^{L} f(y)dy + P \cdot \int_{L}^{\infty} f(y)dy}, & \text{if } y > L. \end{cases} \tag{10.2}$$

Note that, if no persons are sampled above L, so that $P_2 = 0$ (and thus P), equation (10.2) reduces to the density of a truncated distribution (as shown in Hausman and Wise 1977, equation 1.4). Thus complete truncation can be seen to be a special case of this more general possibility.

The expected value of Y given $\mathbf{X}$ can be obtained in a straightforward manner by integration over the density function shown in equation (10.2). It is

$$E(Y \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} - \sigma \frac{(1 - P)\phi[(L_i - \mathbf{X}_i\boldsymbol{\beta})/\sigma]}{(1 - P)\Phi[(L_i - \mathbf{X}_i\boldsymbol{\beta})/\sigma] + P}, \tag{10.3}$$

where ϕ is a unit normal density function and Φ the corresponding distribution function. Note that this expression reduces to the expected value of a truncated distribution when $P = 0$, which of course indicates complete truncation of the distribution at L. We see also that it equals $\mathbf{X}\boldsymbol{\beta}$ when sample selection is random.[5]

If we write $Pr[Y_i \leq L]$, as $\Phi[(L - \mathbf{X}_i\boldsymbol{\beta})/\sigma] = \Phi_i$, and divide the sample into N_1 persons with $Y \leq L$ and N_2 with $y > L$, we can write the log likelihood function as

5. The form of equation (10.3) suggests an estimator not discussed in the text. If P_1 and P_2 were known, a probit specification could be used to estimate $\boldsymbol{\beta}/\sigma$, allowing estimation of values for ϕ and Φ in equation (10.3). Using the fitted value of $\{(1 - P)\phi[\cdot]\}/\{(1 - P)\Phi[\cdot] + P\}$, consistent estimates of both $\boldsymbol{\beta}$ and σ could be obtained by ordinary least squares regression of Y on $\mathbf{X}$ and the fitted value. This procedure is related to those proposed by Heckman (1976) and Lee, chapter 9, for censored models. The extension of this approach to more groups (strata) is outlined in note 7.

$$
\begin{aligned}
L &= \sum_{i=1}^{N_1} \ln f(y_i) - \sum_{i=1}^{N_1} \ln(\Phi_i + P(1-\Phi_i)) \\
&\quad + \sum_{i=1}^{N_2} \ln P + \sum_{i=1}^{N_2} \ln f(y_i) - \sum_{i=1}^{N_2} \ln(\Phi_i + P(1-\Phi_i)) \\
&= \sum_{i=1}^{N} \ln f(y_i) - \sum_{i=1}^{N} \ln(P + (1-P)\Phi_i) + N_2 \ln P. \qquad (10.4)
\end{aligned}
$$

For convenience we let the index i begin at 1 in both groups, instead of letting it run from 1 to N_1 and from $N_1 + 1$ to N_2, for example. We use this convention throughout the chapter. It should be clear from the context what the more precise notation would be. Maximization of this function would lead to a maximum likelihood estimate for P as well as for $\boldsymbol{\beta}$ and σ. Or, if P were known, it could be maximized with respect to $\boldsymbol{\beta}$ and σ only.[6]

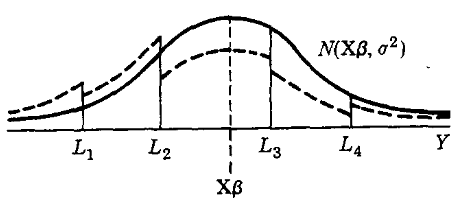

Figure 10.2

It is easy to extend the idea to more income groups. Assume that the five groups in the Gary experiment are defined by the points $L_1, L_2, \ldots, L_4$. Then, given $\mathbf{X}$, the distribution of Y might look something like the dashed representation in figure 10.2. Let the proportion of a random sample of the population below L_1, that is, selected, be P_1 between L_1 and L_2 be P_2, and so on, and above L_4 be P_5. Again we normalize by dividing each proportion by P_1. In the following we will use P_2, P_3, P_4, and P_5 but understand them to be P_2/P_1, P_3/P_1, and so on. Then the density function of Y is given by

6. N_1 and $N_2 = N - N_1$ are assumed random in the sampling procedure. $N = N_1 + N_2$ is given. An alternative sampling procedure is to fix N_1 and N_2. That is, N_1 values are sampled below L and N_2 above. This possibility is discussed in some detail in section 10.5.

$$h(y) = \begin{cases} \dfrac{f(y)}{\int_{-\infty}^{L_1} f(y)dy + P_2 \int_{L_1}^{L_2} f(y)dy + \cdots + P_5 \int_{L_4}^{\infty} f(y)dy}, & \text{if } y \le L_1, \\ \dfrac{P_2 f(y)}{\int_{-\infty}^{L_1} f(y)dy + P_2 \int_{L_1}^{L_2} f(y)dy + \cdots + P_5 \int_{L_4}^{\infty} f(y)dy}, & \text{if } L_1 < y \le L_2, \\ \vdots & \\ \dfrac{P_5 f(y)}{\int_{-\infty}^{L_1} f(y)dy + P_2 \int_{L_1}^{L_2} f(y)dy + \cdots + P_5 \int_{L_4}^{\infty} f(y)dy}, & \text{if } L_4 < y. \end{cases} \tag{10.5}$$

If we let $\Phi[(L_1 - \mathbf{X}\boldsymbol{\beta})/\sigma] = \Phi_1$, and so on, we can again write the expected value of Y given $\mathbf{X}$ as

$$E(Y \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} - \sigma \frac{(1 - P_2)\phi_1 + (P_2 - P_3)\phi_2 + (P_3 - P_4)\phi_3 + (P_4 - P_5)\phi_4}{(1 - P_2)\Phi_1 + (P_2 - P_3)\Phi_2 + (P_3 - P_4)\Phi_3 + (P_4 - P_5)\Phi_4 + P_5}. \tag{10.6}$$

The last term of course is an indicator of sample selection bias; if all proportions were 1, the expected value would be $\mathbf{X}\boldsymbol{\beta}$.[7] We see that it is no longer straightforward to evaluate the direction of the bias. It depends on the relative values of the proportions $P_2, \ldots, P_5$. They may have

7. The estimator suggested in note 6 could be extended to this more complex situation as well, if P_1 through P_5 were known. Let the bias term in equation (10.6) be $\sigma\mathbf{V}$. Then the elements of $\mathbf{V}$ could be estimated using ordered probit analysis, and next $\boldsymbol{\beta}$ and σ by regressing Y on $\mathbf{X}$ and the fitted value of $\mathbf{V}$. Probit analysis can be used here, even though there are several groups, because the groups are ordered. Let

$$\Pi_1 = Pr(y \text{ is observed and } -\infty < y \le L_1 \mid \mathbf{X})$$

offsetting effects, for example. But we can observe that, if the sampling proportions decrease consistently with income, the bias will be negative.[8]

In practice both y and L would be indexed by i to indicate the ith family. If we let $\Phi[(L_{1i} - \mathbf{X}_i\boldsymbol{\beta})/\sigma] = \Phi_{1i}$, and so on, and N_1 be the number of observations below L_1, N_2 the number between L_1 and L_2, and so on, we can write the log likelihood function for this case as

$$\begin{aligned} L = \sum_{i=1}^{N} \ln f(y_i) - \sum_{i=1}^{N} \ln[(1 - P_2)\Phi_{1i} + (P_2 - P_3)\Phi_{2i} \\ + (P_3 - P_4)\Phi_{3i} + (P_4 - P_5)\Phi_{4i} + P_5] \\ + N_2 \ln P_2 + N_3 \ln P_3 + N_4 \ln P_4 + N_5 \ln P_5. \end{aligned} \tag{10.7}$$

As in the first example maximization of L would yield estimates for $\boldsymbol{\beta}$, σ^2. and the proportions P_2 through P_5, if they were unknown.

Note that one could test the hypothesis of population proportions under the normality assumptions by testing the hypothesis that P_2 through P_5 are all equal to one. Each could also be tested individually. If the hypothesis that they are all equal to one cannot be rejected, then it seems reasonable to assume random sampling.

The proportions P_2 through P_5 are often known, at least approximately. For example, for each income interval of the Gary data it is possible to calculate the approximate ratio of the number of persons in the actual sample to the number in a larger random sample. In some instances one could make use of population-versus-sample frequencies within intervals.

$$= P_1 \cdot \int_{-\infty}^{(L_1 - X\boldsymbol{\beta})/\sigma} \phi(u)\,du,$$

$$\Pi_2 = Pr(y \text{ is observed and } L_1 < y \leq L_2 \mid \mathbf{X})$$

$$= P_2 \cdot \int_{(L_1 - X\boldsymbol{\beta})/\sigma}^{(L_2 - X\boldsymbol{\beta})/\sigma} \phi(u)\,du,$$

and so on through Π_5. Construct the likelihood function

$$L = \sum_{i=1}^{N} \sum_{j=1}^{5} \ln \Pi_{ij}^{y_{ij}} ,$$

where y_{ij} equals 1 if y is in the jth income interval and 0 otherwise. Maximization of L will yield estimates of $\boldsymbol{\beta}/\sigma$, which can in turn be used to estimate the elements of $\mathbf{V}$.

8. Equation (10.6) does not of course indicate the magnitude of the bias in individual elements of the vector of parameters $\boldsymbol{\beta}$. In practice, however, the bias in individual parameters tends toward 0.

But if the proportions are known, it seems intuitively plausible that weighted least squares would also yield consistent estimates. The idea can be motivated by referring again to figure 10.1. Consider a particular value of **X**, say **X***. Because a random sample would include approximately two times as many observations above L as there in fact are, we would like to fill in these missing points. This can be done by using inverse sampling weights. If P is the proportion of observations above L in a random sample actually sampled, then $1/P$ would be the expected number of observations in a random sample. If again we divide the sample into two groups, with N_1 below and N_2 above L, we minimize the expression

$$S = \sum_{i=1}^{N_1} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \frac{1}{P} \sum_{i=1}^{N_2} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2. \tag{10.8}$$

More generally, if for each value Y_i we associate a sampling proportion P_i—by determining in which income interval Y_i falls—we minimize the expression

$$S = \sum_{i=1}^{N} \frac{1}{P_i} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2, \tag{10.9}$$

which is of course equivalent to weighted least squares with the weights given by $1/\sqrt{P_i}$. As with the maximum likelihood estimates we can normalize by dividing each P_i by the value of P associated with values of Y below L_1, P_1 in our terminology. To draw an explicit comparison with the likelihood function in equation (10.7), the sample can be broken into five groups and S written as

$$S = \sum_{j=1}^{5} \sum_{i=1}^{N_j} \frac{1}{P_j} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2, \tag{10.10}$$

where $1/P_1 = 1$. The log likelihood function analogous to this expression is[9]

9. Estimates of $\boldsymbol{\beta}$ could be obtained by weighted least squares in the usual way or by maximizing this function.

$$L = \sum_{j=1}^{5} \frac{1}{P_j} \sum_{i=1}^{N_j} \ln f(y_i)$$
$$= \sum_{j=1}^{5} \left\{ -\frac{N_j}{P_j} \ln \sqrt{2\pi} - \frac{N_j}{P_j} \ln \sigma - \sum_{i=1}^{N_j} \frac{1}{2P_j} \left(\frac{Y_i - \mathbf{X}_i \boldsymbol{\beta}}{\sigma} \right)^2 \right\}. \tag{10.11}$$

The relative efficiency of maximum likelihood versus weighted least squares estimates is discussed in the next section. Both are consistent. We shall show, however, that the maximum likelihood estimates based on the density function shown in equation (10.5), and the corresponding likelihood function of equation (10.7), are likely to be more efficient than the weighted least squares estimates, or identical estimates obtained from maximization of the analogous likelihood function shown in equation (10.11). If the sampling ratios are not known, as at first appeared to be the case with the Gary data, a maximum likelihood procedure must of course be used, with the weights estimated along with the behavioral parameters.

The maximum likelihood procedure proposed in this section can easily be extended to accomodate two time periods, or a two-equation—wages and hours worked—model. But because our empirical results based on pre-experimental data do not suggest substantial sample selection bias, we have not extended the analysis here. We have, however, sketched out the density and corresponding likelihood functions applicable to these extensions in the appendix.

10.3 Relative Efficiencies of Weighted Least Squares versus Maximum Likelihood Estimates

Only when the sampling proportions P_i are known can weighted least squares estimates be obtained. Thus only in this case does it make sense to compare the variances of weighted least squares with maximum likelihood estimates. In practice these proportions or their approximate values are likely to be known.

Although the weighted least squares estimates do not depend on distributional assumptions, because of the stratification of the endogenous variable, both their expected value and variance do. Not surprisingly, standard errors calculated from a weighted least squares regression are not consistent estimates of the true standard errors, even asymptotically.[10] On

10. We have reported the calculated weighted least squares standard errors in the results, however.

the other hand, the maximum likelihood estimates themselves depend on distributional assumptions. Given that the distributional assumptions are correct, however, the maximum likelihood estimates are more efficient than weighted least squares. And standard errors are provided easily by applying asymptotic distribution theory relevant to maximum likelihood estimates. But because the weighted least squares estimates are distribution free, they provide a check on the distributional assumptions used to obtain maximum likelihood estimates. Both estimates are asymptotically unbiased.

As we will see, the relative efficiencies of these two estimators depend on several parameters. To obtain some idea of the orders of magnitude that one might expect, we consider a case with only two strata and in which only the expected value of the dependent variable is to be estimated, without variables X. This gives a reasonable indication of relative efficiencies when the expected value depends on a vector of parameters.

The model we consider is of the form

$$Y_i = \beta + \varepsilon_i, \tag{10.12}$$

where the ε_i are independently distributed as $N(0, \sigma^2)$. Assume that Y is divided into two strata defined by point L as shown in figure 10.1 and as discussed in section 10.2. For convenience we continue to use both P_1 and P_2 (as opposed to their ratio P). Recall that because ordinary least squares, OLS, yields an unbiased estimate of the sample mean, the expectation of the OLS estimate, $\bar{Y}$, is

$$E\hat{\beta}_{\text{OLS}} = \beta - \sigma \frac{(P_1 - P_2)\phi((L - \beta)/\sigma)}{P_2 + (P_1 - P_2)\Phi(L - \beta/\sigma)}, \tag{10.13}$$

which is analogous to equation (10.2) with P relaced by P_2/P_1. Note that, if P_2 is greater than P_1 (values of Y greater than L are oversampled), the bias is positive. It is negative if values less than L are oversampled, so that P_1 is greater than P_2.

The weighted least squares, WLS, estimate of β is found by minimizing the expression

$$S = \sum_{i=1}^{N} \frac{1}{P_i}(Y_i - \beta)^2 = \sum_{i=1}^{N_1} \frac{1}{P_1}(Y_i - \beta)^2 + \sum_{i=1}^{N_2} \frac{1}{P_2}(Y_i - \beta)^2, \tag{10.14}$$

where N_1 of the values in the sample fall below L and N_2 above. This is a special case of equation (10.9). The estimate is given by

$$\hat{\beta}_{\mathrm{WLS}} = \left(\frac{N_1}{P_1} + \frac{N_2}{P_2}\right)^{-1} \left[\sum_{i=1}^{N_1} \frac{Y_i}{P_1} + \sum_{i=1}^{N_2} \frac{Y_i}{P_2}\right]. \tag{10.15}$$

If the Y_i are normally distributed, the expected value of Y_i for $Y_i \leq L$ is given by $\beta - \sigma(\phi/\Phi)$, and for $Y_i > L$ by $\beta + \sigma(\phi/(1 - \Phi))$, where both ϕ and Φ are functions of $(L - \beta)/\sigma$. Although the weighted least squares estimator of β is not unbiased, in general it is consistent and asymptotically unbiased. Given N_1 (and thus $N_2 = N - N_1$), its expected value is given by

$$E(\hat{\beta}_{\mathrm{WLS}} \mid N_1) = \beta + \left(\frac{N_1}{P_1} + \frac{N_2}{P_2}\right)^{-1} \left[-\frac{N_1}{P_1}\sigma\frac{\phi}{\Phi} + \frac{N_2}{P_2}\sigma\frac{\phi}{1 - \Phi}\right]. \tag{10.16}$$

As N gets large, N_1 goes to $P_1\Phi N/D$, and $N_2 = N - N_1$ to $P_2(1 - \Phi)N/D$, where $D = P_1\Phi + P_2(1 - \Phi)$. The second term of (10.16) is zero when these values are substituted for N_1 and N_2.[11]

The derivation of the variance of the weighted least squares estimator is complicated by the fact that N_1 (and N_2), as well as the Y_i are random variables. If we use the property that the variance of $\hat{\beta}$ is equal to $E[\mathrm{Var}(\hat{\beta} \mid N_1)] + \mathrm{Var}[E(\hat{\beta} \mid N_1)]$, we can write it as

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_{\mathrm{WLS}}) &= E\left\{\left(\frac{N_1}{P_1} + \frac{N_2}{P_2}\right)^{-2} \left[\frac{N_1}{P_1^2}\left(\sigma^2 - \sigma^2\left(\frac{L - \beta}{\sigma}\cdot\frac{\phi}{\Phi} + \frac{\phi^2}{\Phi^2}\right)\right)\right.\right. \\
&\quad \left.\left. + \frac{N_2}{P_2^2}\left(\sigma^2 + \sigma^2\left(\frac{L - \beta}{\sigma}\cdot\frac{\phi}{1 - \Phi} - \frac{\phi^2}{(1 - \Phi)^2}\right)\right)\right]\right\} \\
&\quad + \mathrm{Var}\left\{\beta + \left(\frac{N_1}{P_1} + \frac{N_2}{P_2}\right)^{-1} \left[-\frac{N_1}{P_1}\sigma\frac{\phi}{\Phi} + \frac{N_2}{P_2}\sigma\frac{\phi}{1 - \Phi}\right]\right\} \\
&= E\{f(N_1)\} + \mathrm{Var}\{g(N_1)\},
\end{aligned} \tag{10.17}$$

where f and g are defined by the last equality. By using appropriate asymptotic Taylor expansions of both f and g, the variance can be approximated by

11. By expanding the second term in (10.16) around the expected value of N_1, it can be shown that the expected value of this term with respect to N_1 goes to zero at the rate of $1/N$.

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_{\mathrm{WLS}}) = {} & \frac{D\Phi}{P_1 N}\left[\sigma^2 - \sigma^2\left(\frac{L-\beta}{\sigma}\cdot\frac{\phi}{\Phi} + \frac{\phi^2}{\Phi^2}\right)\right] \\ & + \frac{D(1-\Phi)}{P_2 N}\left[\sigma^2 + \sigma^2\left(\frac{L-\beta}{\sigma}\cdot\frac{\phi}{1-\Phi} - \frac{\phi^2}{(1-\Phi)^2}\right)\right] \\ & + \frac{P_1\Phi\cdot P_2(1-\Phi)\cdot\phi^2\cdot\sigma^2}{N}\left(\frac{1-\Phi}{P_1} + \frac{\Phi}{P_2}\right)^2\left(\frac{1}{\Phi} + \frac{1}{1-\Phi}\right)^2,\end{aligned} \tag{10.18}$$

where the first two terms come from $E\{f(N_1)\}$ and the third from $\operatorname{Var}\{g(N_1)\}$. The last term essentially represents the variance in the estimate due to the randomness of N_1.

The correct variance is not given when a standard regression program is used to calculate the weighted least squares estimates. The correct variance is also considerably more difficult to calulate than the maximum likelihood variance.

A maximum likelihood, ML, estimate for β may be obtained by maximization of a function analogous to (10.4).[12] The variance of the asymptotic distribution of the maximum likelihood estimator of β is given by

$$\operatorname{Var}(\hat{\beta}_{\mathrm{ML}}) = \frac{\sigma^2}{N}\left[1 - \frac{(P_1-P_2)\phi}{P_2 + (P_1-P_2)\Phi}\left(\frac{L-\beta}{\sigma} + \frac{(P_1-P_2)\phi}{P_2 + (P_1-P_2)\Phi}\right)\right]^{-1}, \tag{10.19}$$

using the appropriate term from the information matrix. Note that it equals σ^2/N when $P_1 = P_2$ and increases as P_2 approaches zero.

To compare the relative efficiencies of maximum likelihood and weighted least squares estimates, we need to make some simplifications, since both estimators depend on L, β, and σ, as well as P_1 and P_2. As discussed in the previous section, we need only consider the ratio of P_1 and P_2; we accomplish this by setting $P_1 = 1$. We also set $L = 0$ and $\sigma = 1$. Ratios of the weighted least squares to the maximum likelihood variance for various values of P_1 and β are given in table 10.1.

12. The case of known P_i yields a likelihood function very similar to those corresponding to tobit and standard truncation situations. Amemiya's (1973) proofs of the properties of maximum likelihood estimators can be altered in a straightforward way and applied here. His proofs, however, cannot be as easily extended to cover the case of unknown P_i.

Table 10.1
Relative efficiency of maximum likelihood versus weighted least squares estimates for selected values of P_2 and β

P_2	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
0.01	9.906	5.439	3.818	4.506
0.10	1.736	1.524	1.696	1.748
0.20	1.291	1.256	1.379	1.325
0.30	1.148	1.149	1.225	1.172
0.50	1.045	1.053	1.080	1.054
0.70	1.012	1.015	1.022	1.014
0.80	1.005	1.006	1.009	1.006
0.90	1.001	1.001	1.002	1.001
0.99	1.000	1.000	1.000	1.000
1.00	1.000	1.000	1.000	1.000

The gain in efficiency from using maximum likelihood instead of weighted least squares is small, as long as P_2 is greater than 0.5. But the relative efficiency of maximum likelihood becomes substantial if P_2 is less than 0.3, say. It should be emphasized, however, that the maximum likelihood asymptotic variance is much easier to calculate for any values of P_2 (as well as the other parameters), although weighted least squares estimates themselves may be easier to obtain. The relative efficiencies are not affected very much by β except at very low levels of P_2. One might conclude that weighted least squares would give a good indication of the importance of endogenous stratification, when compared to least squares estimates, as long as P_2 were not very small.

Relative efficiencies in the more general case with $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta}$ is a vector of parameters, would be developed analogously, with β replaced by the conditional expectation $\mathbf{X}_i\boldsymbol{\beta}$. But in this case the conditional expectation depends on $\mathbf{X}$ as well as $\boldsymbol{\beta}$, and the variance calculations depend on the values of $\mathbf{X}$ in the sample. Thus it is impossible to present simple comparisons like those in table 10.1. Presumably the same considerations apply, however. For P_2 close to one, weighted least squares should entail little loss of efficiency. The analysis also could be extended to more strata and associated values P_i. There seems to be no straightforward way to compare efficiencies in this case either, and in addition our results for two strata may provide a less reliable guide for this more complicated case. However, one might suspect that very low values of P_i with respect to any strata tend to lower the relative efficiency of weighted least squares.

10.4 Empirical Results of the Selection Bias in the Gary Income Maintenance Experiment

We will first describe briefly the sample selection procedure followed in the Gary experiment. (For a more detailed discussion see Kehrer et al. 1975.) Then we will compare maximum likelihood and weighted least squares estimates—that correct for sample selection—with least squares estimates obtained with no attempt to correct for selection bias.

Recall that after some preliminary screening a random sample of approximately 2,600 black families was selected from specified geographic areas of Gary. The families had to include at least one dependent child under the age of 18; more than half were female-headed households. This group was stratified by income as well as by exogenous variables. Families were then selected at random within the income intervals defined by multiples of the poverty level, which depends on family size. The government poverty line for a family of four was $4,275 in 1972, when the experiment began.

Only the male-headed households, of which about 730 were selected for the experiment, are used in our analysis. The proportions of the random sample of male-headed households in each income interval that were included in the selected sample of 730 are shown in the third column of table 10.2. Normalized ratios—the P values in equations (10.5) through (10.7)—are shown in the last column. The first two columns define the income groups. (The last four numbers are the values of P_2 through P_5, which follow the definitions applied in equation 10.4, in particular the bias expression in equation 10.6.) We see that relative to families in the first income interval those in the third and fourth intervals are slightly oversampled, while those in the highest income group are substantially undersampled. In a random sample we would expect to find about two and one-half times as many observations in the high income group as we in fact have. By referring back to equation (10.6), we see that of the four terms in the numerator of the bias expression the first is approximately zero, while the second and third are positive, and the fourth is negative. Thus we cannot a priori evaluate the direction of the bias. But it seems clear that the bias should be much less than in the New Jersey negative income tax experiment. The New Jersey sample excluded altogether families in the two highest income intervals as defined here. As one might expect, the estimates indicate much less bias using the Gary sample than was found using data from the New Jersey experiment.

Table 10.2
Sampling proportions and ratios

Group	Income (Y) interval (multiple of poverty line)	Proportion of random sample selected for experiment	Normalized ratios
1	$Y \leq 0.5$	0.7273	$P_1 = 1.0000$
2	$0.5 < Y \leq 1.0$	0.7381	$P_2 = 1.0148$
3	$1.0 < Y \leq 1.5$	0.8061	$P_3 = 1.1083$
4	$1.5 < Y \leq 2.4$	0.8594	$P_4 = 1.1816$
5	$2.4 < Y$	0.2966	$P_5 = 0.4078$

To evaluate the extent of sample selection bias in the Gary data, we estimated earnings equations using pre-experimental (baseline) data rather than experimental data, because a large number of families dropped out of the experiment over time. We wanted to avoid the somewhat more complex specification that would be required to correct for sample selection and attrition bias simultaneously. The results indicate that sample selection bias was not severe. Therefore in evaluating the extent of attrition bias (Hausman and Wise 1977b), we have not at the same time made a correction for sample selection bias.

We will compare four sets of estimates:

1. ordinary least squares,
2. weighted least squares using inverse sample ratios,
3. maximum likelihood with known sample ratios,
4. maximum likelihood with unknown sample ratios.[13]

The weights used in the second and third approaches are those listed in the last column of table 10.2. Estimates using the last two methods are obtained by maximization of a likelihood function similar to equation (10.7), but the third uses the known normalized sample ratios and estimates only $\boldsymbol{\beta}$ and σ^2, while the fourth estimates the ratios P_2 through P_5 along with $\boldsymbol{\beta}$ and σ^2. If the maintained assumptions of the model are in accord with empirical evidence—in particular that given $\mathbf{X}$ income is distributed

13. The maximum likelihood estimates were obtained using the Berndt, Hall, Hall, and Hausman modified scoring algorithm. The costs were approximately twice the cost of the weighted least squares estimates.

log normal—then the estimated ratios should be close to the those shown in the table.

The dependent variable in each case is the logarithm of labor income. The independent variables are

Constant
Education: years of education,
Experience: years of work experience,
Income: log of nonlabor family income, including foodstamps, AFDC payments, public assistance, and earnings of other family members,
Union: a dummy variable that is one for union members and zero otherwise.

A total of 585 black males were used in the analysis, comprising both controls and experimentals (persons ultimately assigned an experimental income guarantee and tax rate).[14] Note the limits $L_1, \ldots, L_4$ that define the income intervals depend on family size and must therefore be calculated for each observation. The ratios, however, are the same over all observations. The limits have been adjusted slightly because they pertain to family income, whereas our data pertain to earnings of the male head only.[15]

Parameter estimates are presented in table 10.3. Ratios of the other estimates to least squares estimates are shown in table 10.4. In this case we find that least squares estimates do not differ substantially from those that correct for sample selection bias and do not seem to be systematically biased in one direction or another. In general the estimates obtained by weighted least squares and by the two maximum likelihood methods agree rather closely with one another. The estimated coefficients on income, however, differ substantially, with the weighted least squares estimates tending considerably closer to zero than the maximum likelihood estimates.

14. Although there were about 730 male-headed households in the sample, we had complete data for only 585 of them. This could of course affect the relationship between the estimated P values and the sampling proportions as shown in table 10.2. Based on evidence reported elsewhere (Hausman and Wise 1977b), we believe that these missing observations would not have a substantial affect on the parameter estimates.

15. Other income was assumed given, and the limits were related to the earnings of the male head by determining individual earnings limits corresponding to the specified limits. For example, if other family income is 0.5 times the poverty level, then the male head would have to earn between 0.5 and 1.0 times the poverty level to be in group 3. In practice family income not included in labor income of the male head was very small on the average.

Table 10.3
Parameter estimates (and standard errors) by method of estimation

Variable	Least squares	Weighted least squares[a]	Maximum likelihood (ratios known)	Maximum likelihood (ratios unknown)
Constant	5.916	5.8424	5.9300	5.7355
	(0.0879)	(0.0899)	(0.1047)	(0.1196)
Education	0.0190	0.0270	0.0252	0.0281
	(0.0068)	(0.0068)	(0.0079)	(0.0083)
Experience	0.0042	0.0048	0.0050	0.0053
	(0.0018)	(0.0018)	(0.0020)	(0.0023)
Income	−0.0162	−0.0056	−0.0189	−0.0231
	(0.0068)	(0.0069)	(0.0092)	(0.0101)
Union	0.2596	0.2314	0.2021	0.2881
	(0.0519)	(0.0407)	(0.0386)	(0.0647)
P_2	—	—	—	1.4831
	—	—	—	(0.1836)
P_3	—	—	—	0.8267
	—	—	—	(0.1983)
P_4	—	—	—	2.3916
	—	—	—	(0.2361)
P_5	—	—	—	0.2429
	—	—	—	(0.0570)

[a]The standard errors shown in this column are those reported from a regression program. They understimate the true standard errors.

Table 10.4
Ratios of other estimates to least squares estimates by method of estimation

Variable	Weighted least squares	Maximum likelihood (ratios known)	Maximum likelihood (ratios unknown)
Constant	0.99	1.00	0.97
Education	1.42	1.33	1.48
Experience	1.14	1.19	1.26
Income	0.35	1.17	1.43
Union	0.89	0.78	1.11

The estimated ratios P_2 through P_5 do differ from their sample counterparts shown in table 10.2, but the general pattern of the estimates is similar to the sample ratios. That is, they suggest that persons in the highest income group were undersampled relative to persons in the other groups. But they also suggest a much larger oversampling of persons in the fourth interval than the sample ratios indicate and a somewhat larger oversampling in the second interval. We can see nonetheless that, if these estimates were in fact accurate reflections of empirical ratios, the undersampling in the highest interval would tend to be offset by oversampling in the fourth.[16] Thus the bias resulting from one is offset by the other.

We note also that these sampling ratios are estimated with considerable precision. Each of the standard errors is less than one-fourth of the corresponding estimate. We have not formally tested the hypothesis that all are equal to one, indicating random sampling, but it is clear from the standard errors that this hypothesis would be rejected.

Finally we emphasize that in principle it is not possible to distinguish deviations from random sampling from deviations from normality in the population. We have maintained the hypothesis of a log normal distribution of income in the population, given **X**. On the other hand, it seems unlikely that deviations from normality in the population would follow the pattern of the estimates we have obtained.

10.5 Alternative Sampling Procedures

The sampling process assumed in the foregoing analysis was intended to reflect as closely as possible the process actually used in the Gary experiment, as we understand it. There are of course several other possibilities. We will briefly discuss two others, to the point of presenting appropriate likelihood functions. Recall that we assumed that the total number of observations retained was fixed, but the number of retained observations within each stratum was random. One alternative is to stratify the population (or a random sample from the population) and draw a fixed number of observations within each stratum. A second alternative is to *fix the number of retained plus unretained observations*, letting both the number of observations within each stratum and the number of observations actually observed be random.

16. We note that the calculated ratios are subject to error.

Consider first sampling within strata, with N_1 observations taken from the first, N_2 from the second, and so on. The density function for observations in the sample is given by

$$h(y) = \begin{cases} \dfrac{f(y)}{\Phi_1}, & \text{if } y \leq L_1, \\ \dfrac{f(y)}{\Phi_2 - \Phi_1}, & \text{if } L_1 < y \leq L_2, \\ \vdots & \vdots \\ \dfrac{f(y)}{1 - \Phi_4}, & \text{if } L_4 < y. \end{cases} \tag{10.20}$$

Within each stratum the values of y have truncated normal density functions. The appropriate log likelihood function is then

$$L = \sum_{j=1}^{5} \left\{ \sum_{i=1}^{N_j} \ln f(y_i) - \sum_{i=1}^{N_j} \ln (\Phi_{ji} - \Phi_{j-1,i} \right\}, \tag{10.21}$$

where $\Phi_0 = 0$ and $\Phi_5 = 1$. This is a straightforward generalization of our earlier work (Hausman and Wise 1977a).

A variant of this possibility arises if we know the proportions of the population with values of y within each of the strata. Let the values be $Q_1, \ldots, Q_5$. Then, for example,

$$Q_1 \equiv \int_{\mathbf{X}} \Phi_1 \cdot g(\mathbf{X}) d\mathbf{X},$$

where $g(\mathbf{X})$ is the multivariate density function defined over population values of the vector $\mathbf{X}$. Similar expressions apply to Q_2 through Q_5. Presumably more efficient estimates could then be obtained by maximizing (10.21) subject to these identity constraints. In general, however, they seem to be intractable.[17]

17. A possible exception is to suppose that $\mathbf{X}$ is distributed multivariate normal with mean μ_X and covariance matrix Σ_X. Then, for example,

$$Q_1 \equiv \Phi\left[\frac{L_1 - \mu_X \beta}{1 + \beta' \Sigma_X \beta}\right],$$

and analogous expressions define Q_2 through Q_5. Alternatively $g(\mathbf{X})$ could be replaced by

Next suppose that the total sample size is fixed but that some values of y are unobserved. Each value of $\mathbf{X}$ is observed. Whether y is observed or not depends on the stratum in which the randomly selected observation of y falls. If it falls in the first stratum, it is retained in the sample with probability P_1. It is retained with probability P_2 if it falls in the second, and so forth.

To develop a likelihood function for this case, consider the pairs of values $(y, \mathbf{X})$ and the sample selection probability P. The likelihood that a pair $(y, \mathbf{X})$ will be in the sample is given by

$$l(y,\mathbf{X}) \cdot P(y) = f(y \mid \mathbf{X}) g(\mathbf{X}) P(y). \tag{10.22}$$

It is the likelihood that the pair $(y, \mathbf{X})$ is randomly selected multiplied by the probability that it is retained in the sample, once selected. The probability of retention depends only on y. As shown in (10.22), $P(y)$ equals P_1 for $y < L$, P_2 for $L_1 < y < L_2$, and so on.

The likelihood of a pair $(y, \mathbf{X})$ with y unobserved is given by

$$l(y,\mathbf{X})(1 - P(y)) = f(y \mid \mathbf{X}) g(\mathbf{X})(1 - P(y)). \tag{10.23}$$

But only the stratum of y is known if y is unobserved, not y itself. The likelihood of observing $\mathbf{X}$ with unobserved y in the first stratum, for example, is given by

$$\begin{aligned} \int_{-\infty}^{L_1} f(y \mid \mathbf{X}) g(\mathbf{X})(1 - P(y))\,dy &= \Phi\left[\frac{L_1 - \mathbf{X}\beta}{\sigma}\right] g(\mathbf{X}) \cdot P_1 \\ &= \Phi_1 g(\mathbf{X}) P_1. \end{aligned}$$

The analogous expression for unobserved y in the second stratum is given by

$$(\Phi_2 - \Phi_1) g(\mathbf{X}) P_2,$$

and so forth.

If there are N_1 observed values of y in the first stratum, and $N_{1'}$ unobserved, N_2 observed values in the second and $N_{2'}$ unobserved, and so forth, the likelihood function for N observations is

weights corresponding to empirical observations and (10.21) maximized with respect to these weights as well as the other parameters of the likelihood function (see Cosslett 1977).

$$L = \sum_{j=1}^{5} \left\{ N_i \ln P_j + \sum_{i=1}^{N_j} \ln (f(y_i \mid \mathbf{X}_i) + N_{j'} \ln (1 - P_j) \right.$$
$$\left. + \sum_{i=1}^{N_{j'}} \ln (\Phi_{ji} - \Phi_{j-1,i}) \right\}, \tag{10.24}$$

where $\Phi_0 = 0$ and $\Phi_5 = 1$.

The term $\Sigma_{i=1}^{N} \ln g(\mathbf{X}_i)$ has been deleted because it does not include any of the parameters of (10.24). Presumably this formulation would provide more efficient estimates of P_1 through P_5, because within any stratum the number of unobserved as well as observed values of y is known. The expected value of each of these numbers is determined by the corresponding value of P. The estimates of the β parameters would also be more efficient because, although given $N_1 + \cdots + N_5$, we have the same number of observations of y; the observations with y unobserved are represented explicitly by $N_{1'} + \cdots + N_{5'}$ probit functions in (10.24). Each indicates the probability that y falls in the indicated interval, given $\mathbf{X}$, and provides additional information on the value of β.

A variant of this case is to suppose that for y unobserved, $\mathbf{X}$ is also unobserved; the observations are completely missing. This was the assumption in sections 10.2 through 10.4. But we assume somewhat more information here, namely, the number of observations that are discarded in each stratum, and we make explicit use of this information.

The likelihood for retained observations is the same as in equation (10.22). But to get expressions for the likelihoods of unobserved values, we must integrate out $\mathbf{X}$ as well as y, since neither is observed in this case. For example, the probability of an unobserved pair $(y, \mathbf{X})$ with y in the first stratum is given by

$$\int_{\mathbf{X}} \int_{-\infty}^{L_1} f(y \mid \mathbf{X}) g(\mathbf{X})(1 - P(y)) dy d\mathbf{X} = \int_{\mathbf{X}} \Phi_1 g(\mathbf{X})(1 - P_1) d\mathbf{X}$$
$$\equiv (1 - P_1) Q_1.$$

Note this expression is identical to $(1 - P_1)Q_1$. Similar expressions pertain to unobserved pairs in the other strata.

The log likelihood function for N observations would be

$$L = \sum_{j=1}^{5} \left\{ N_j \ln P_j + \sum_{i=1}^{N_j} \ln f(y_i \mid \mathbf{X}_i) + \sum_{i=1}^{N_j} \ln g(\mathbf{X}_i) \right.$$

$$\left. + N_{j'} [\ln (1 - P_j) + \ln Q_j] \right\}, \tag{10.25}$$

where

$$Q_1 = \int_{\mathbf{X}} \Phi_1 g(\mathbf{X}) d\mathbf{X}, \tag{10.26}$$

$$Q_2 = \int_{\mathbf{X}} (\Phi_2 - \Phi_1) g(\mathbf{X}) d\mathbf{X},$$

$$\vdots$$

$$Q_5 = \int_{\mathbf{X}} (1 - \Phi_4) g(\mathbf{X}) d\mathbf{X}.$$

One could think of this function as the likelihood of $N_1, N_2, \ldots, N_5$ observed values in the five strata; $N_{1'}, N_{2'}, \ldots, N_{5'}$ unobserved values; $y_{11}, \ldots, y_{1N_1}$ in strata 1; $y_{21}, \ldots y_{2N_2}$ in strata 2; and so on to $y_{51}, \ldots, y_{5N_5}$ in strata 5. Because more information is retained in this procedure than the one described in section 10.3, it yields more efficient estimates. In fact, if the Q_i were known, (10.25) could be maximized subject to (10.26) and considered as a series of identity constraints. Maximization of this likelihood function, with or without the constraint, appears to be impractical at this time, however.

10.6 Conclusion

Sample selection for panel surveys is often based on a stratified sample design with random sampling within strata. The proportions of observations within strata do not necessarily correspond to population proportions. This is usually not a serious problem if stratification is based on exogenous variables. But often variables we would like to treat as endogenous are also the basis for stratification. For example, the New Jersey negative income tax experiment excluded entirely all families with income greater than 1.5 times the poverty level. The income maintenance experiments in Gary and Seattle–Denver, although not excluding higher income families, undersampled them. The 1967 survey of economic opportunity also undersampled high income families, as did the University of Michigan panel study of income dynamics that resurveyed part of the SEO sample. Any uses of these data sets that treat earnings or components of earnings as endogenous variables in behavioral relationships will lead to biased estimates of population parameters.

We have presented alternative methods of correcting for endogenous sample selection when faced with rather general stratification designs. If relevant sample versus population proportions are known, either weighted least squares or a more efficient maximum likelihood procedure can be used. If the proportions are not known, they can be estimated along with behavioral parameters using our proposed maximum likelihood procedure. We have demonstrated the technique through estimation of earnings functions using data from the Gary income maintenance experiment. In this case we find that, although the sampling was not random, undersampling of the highest income families tended to be offset by oversampling of the next highest group. Thus parameter estimates were not seriously biased. All of the methods of correcting for endogenous sampling produced similar results.

We note in particular that estimation of sampling ratios when they are unknown is quite practical. In fact we obtained very precise estimates. The general pattern of the estimates was consistent with a priori knowledge about actual ratios, although some of the individual estimates differed significantly from their empirical counterparts.

10.7 Appendix: Extension of the Analysis to Two Time Periods and to Two Equations

The idea embodied in equations (10.5) and (10.7) can easily be extended to earnings, for example, in two time periods such as before and during an experiment, by noting that, if Y_1 and Y_2 are jointly normal, with $Y_1 = \mathbf{X}_1\boldsymbol{\beta} + \varepsilon_1$ and $Y_2 = \mathbf{X}_2\boldsymbol{\beta} + \varepsilon_2$, the appropriate density function would now be

$$h(y_1, y_2) = \begin{cases} \dfrac{f(y_1, y_2)}{\int_{-\infty}^{L_1} f_1(y_1)dy_1 + P_2 \int_{L_1}^{L_2} f_1(y_1)dy_1 + \ldots + P_5 \int_{L_4}^{\infty} f_1(y_1)dy_1}, & \text{if } y_1 \leq L_1, \\ \vdots & \\ \dfrac{P_5 f(y_1, y_2)}{\int_{-\infty}^{L_1} f_1(y_1)dy_1 + P_2 \int_{L_1}^{L_2} f_1(y_1)dy_1 + \ldots + P_5 \int_{L_4}^{\infty} f_1(y_1)dy_1}, & \text{if } L_4 < y_1, \end{cases} \tag{10.27}$$

where $f(y_1, y_2)$ is a bivariate normal density function with appropriate mean vector and covariance matrix and $f_1(y_1)$ is the marginal density function for y_1. This would lead to a log likelihood function of the form

$$\begin{aligned} L = &\sum_{i=1}^{N} f(y_{1i}, y_{2i}) \\ &- \sum_{i=1}^{N} \ln [(P_1 - P_2)\Phi_{1i} + (P_2 - P_3)\Phi_{2i} \\ &+ (P_3 - P_4)\Phi_{3i} + (P_4 - P_5)\Phi_{4i} + P_5] \\ &+ N_2 \ln P_2 + N_3 \ln P_3 + N_4 \ln P_4 + N_5 \ln P_5, \end{aligned} \tag{10.28}$$

where the functions $\Phi_{1i}, \Phi_{2i}, \ldots$, are defined as shown and refer to the cumulative distribution function of Y_1.

The extension of this approach to estimation of simultaneous hourly wage rate and hours-worked equations is also straightforward. Without going through the details here, it can be shown that the appropriate density function would be of the form

$$f(\ln w, \ln h) = \begin{cases} \dfrac{\tilde{\phi}(\ln w, \ln h)}{(P_1 - P_2)\Phi[d_1] + (P_2 - P_3)\Phi[d_2] + \ldots + (P_4 - P_5)\Phi[d_4] + P_5}, & \text{if } y \leq L_1, \\ \dfrac{P_5 \cdot \tilde{\phi}(\ln w_i, \ln h_i)}{(P_1 - P_2)\Phi[d_1] + (P_2 - P_3)\Phi[d_2] + \ldots + (P_4 - P_5)\Phi[d_4] + P_5}, & \text{if } L_4 \leq y, \end{cases} \tag{10.29}$$

where $\tilde{\phi}(\cdot)$ is a bivariate normal density function, and

$$d_1 = \frac{\ln L_1 - X_1\delta_1 - X_1\delta_1\beta - X_2\delta_2}{\sqrt{\omega_{11} + \omega_{22} + 2\omega_{12}}},$$

$$d_2 = \frac{\ln L_2 - X_1\delta_1 - X_1\delta_1\beta - X_2\delta_2}{\sqrt{\omega_{11} + \omega_{22} + 2\omega_{12}}},$$

$$d_4 = \frac{\ln L_4 - X_1\delta_1 - X_1\delta_1\beta - X_2\delta_2}{\sqrt{\omega_{11} + \omega_{22} + 2\omega_{12}}}.$$

The notation is the same as that in Hausman and Wise (1977a), and the development leading to equation (10.29) is analogous to the approach followed there. Again the resulting likelihood function has a rather simple form.

Finally, extension to two equations and two time periods is also straightforward but somewhat tedious and therefore not carried out here.

References

Amemiya, T. 1973. Regression Analysis when the Dependent Variable is Truncated Normal. *Econometrica.* 41: 997–1016.

Cosslett, Stephen R. 1977. Choice-Based Sampling and Disaggregate Demand Forecasting. Mimeograph. Department of Economics, University of California, Berkeley.

Hausman, J. A., and Wise, D. A. 1976. The Evaluation of Results from Truncated Samples: The New Jersey Negative Income Tax Experiment. *Annals of Economic and Social Measurement.* 5: 421–445.

Hausman, J. A., and Wise, D. A. 1977a. Social Experimentation, Truncated Distributions, and Efficient Estimation. *Econometrica.* 45: 319–339.

Hausman, J. A., and Wise, D. A. 1977b. Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. Discussion paper 47D. John F. Kennedy School of Government, Harvard University.

Heckman, J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*. 5: 475–492.

Kehrer, K. C., E. K. Bruml, G. T. Burtless, and D. N. Richardson. 1975. The Gary Income Maintenance Experiment: Design, Administration, and Data Files Mimeograph.

Manski, C. F., and S. R. Lerman. 1977. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica*. 45: 1977–1988.