INSTITUTE FOR RESEARCH ON POVERTY DISCUSSION PAPERS

287-75

ESTIMATING EARNINGS FUNCTIONS FROM TRUNCATED SAMPLES

David L. Crawford

UNIVERSITY OF WISCONSIN - MADISON

ESTIMATING EARNINGS FUNCTIONS FROM TRUNCATED SAMPLES

David L. Crawford

University of Wisconsin

July 1975

The research reported here was supported in part by funds granted to the Institute for Research on Poverty pursuant to the provisions of the Economic Opportunity Act of 1964. The author gratefully acknowledges the comments and suggestions received from Glen Cain, Lewis Evans, Steven Garber, Arthur Goldberger, Donald Hester, Bengt Muthén, and Mead Over.

ABSTRACT

Over the last several years, we have witnessed the accumulation of several micro-data sets, collected for the purpose of analyzing potential responses to income maintenance schemes. These samples are not random samples from the total U.S. population because they were selected to represent those families whose incomes are below certain poverty thresholds. This paper is an examination of the ways in which such data can be used to study behavior in the total U.S. population. This examination focuses upon the estimation of conventional earnings functions for male heads of households where earnings are assumed to be a function of education, IQ, and several demographic variables.

ESTIMATING EARNINGS FUNCTIONS FROM TRUNCATED SAMPLES

1. Introduction

During the last several years, we have witnessed the accumulation of several sets of microeconomic data, collected for the purpose of analyzing potential responses to income maintenance schemes. These samples are not random samples from the total U.S. population because they were selected to represent those families whose incomes are below certain poverty thresholds. This paper examines techniques for the estimation of linear models when the sample design excludes observations when the dependent variable exceeds some "truncation" value.[1] These techniques are illustrated using a national cross-sectional sample and a subsample which has been truncated when income exceeds some predetermined level. These samples are taken from the five year panel data set collected by the Survey Research Center of the Institute for Social Research at the University of Michigan (Survey Research Center, 1972).

This paper is composed of seven sections. In the next section, the problems encountered when one uses linear regression to estimate linear models from truncated samples are examined. In the third section, two techniques which yield consistent parameter estimates from truncated samples are reported. In the fourth section, a simple earnings model is presented. In the following two sections, we discuss the data to be used and estimate the model from a random sample using linear regression and from a truncated sample using linear regression as well as the two consistent techniques presented in the

third section. The final section reports conclusions regarding the use of truncated samples and some proposals for future research.

2. Linear Regression in Truncated Samples

In this section, biases in linear regression coefficient estimates obtained from truncated samples are examined. The focus is on the comparison of the population linear regression functions in the full and truncated populations. In a joint probability distribution $p(y,\underline{x})$, the conditional expectation function $E(y|\underline{x})$, which traces out the conditional mean of y given $\underline{x}$, may be nonlinear in $\underline{x}$. The best (in a least squares sense) linear approximation to $E(y|\underline{x})$ is the population linear regression function

$$L(y|\underline{x}) = \alpha + \underline{x}'\underline{\gamma} \ . \tag{1}$$

where

$$\begin{aligned} \underline{\gamma} &= [V(\underline{x})]^{-1}\underline{C}(\underline{x},y) \\ \alpha &= E(y) - \underline{\gamma}'E(\underline{x}) \end{aligned} \tag{2}$$

and where $V(\underline{x})$ is the variance matrix of $\underline{x}$ and $\underline{C}(\underline{x},y)$ is a column vector of the covariances of individual x's with y. If $E(y|\underline{x})$ is linear then $L(y|\underline{x})$ coincides with it.

The examples which follow are special cases of the classical regression model

$$y = \underline{x}'\underline{\beta} + \varepsilon \ , \quad E(\varepsilon|\underline{x}) = 0 \tag{3}$$

where $\underline{x}$ is a k x 1 vector of exogenous variables. The conditional

expectation function is linear and therefore coincides with the population linear regression function

$$E(y|\underline{x}) = L(y|\underline{x}) = \underline{x}'\underline{\beta} \ . \tag{4}$$

It is well known that linear regression slope estimates, $\underline{b}$, based on a random sample from the full population characterized by (3) have expected values equal to the slopes of $L(y|\underline{x})$.

$$E(\underline{b}) = [V(\underline{x})]^{-1}\underline{C}(\underline{x},y) = \underline{\beta} \ . \tag{5}$$

Since a truncated sample is a random sample from a truncated population, we know that linear regression slope estimates, $\underline{b}^*$, based upon a truncated sample have expected values equal to the slope of $L^*(y|\underline{x})$ (* indicating the truncated population). That is,

$$E(\underline{b}^*) = [V^*(\underline{x})]^{-1}\underline{C}^*(\underline{x},y) = \underline{\beta}^* \ . \tag{6}$$

So the question of bias in linear regression estimates from truncated samples can be reduced to a comparison of (5) and (6); that is, linear regression coefficient estimates based on a truncated sample are biased whenever $\underline{\beta}^*$ is not equal to $\underline{\beta}$.

Example 1. In our first example there is a single x and a population consisting of the nine equi-probable points displayed in Table 1.

Table 1

x	-1	-1	-1	0	0	0	1	1	1
y	-2	-1	0	-1	0	1	0	1	2

The conditional expectation function for y given x is

$$E(y|x) = x \ . \tag{7}$$

Since the conditional expectation function is linear, it coincides with $L(y|x)$ which has slope

$$\frac{C(x,y)}{V(x)} = 1 \ . \tag{8}$$

Now form a subpopulation by deleting those points where y is greater than zero. In this truncated population consisting of 6 equi-probable points, the linear conditional expectation function is also the population linear regression function

$$E^*(y|x) = L^*(y|x) = -.5 + .5x \ . \tag{9}$$

Figure 1, shows the nine original points, $L(y|x)$, and $L^*(y|x)$. Now, consider drawing random samples from the full and the truncated populations and computing the linear regression of y on x. Note that estimates obtained with samples from the truncated population will be biased estimates of the slope of $L(y|x)$ because they are unbiased for the slope of $L^*(y|x)$, which is not equal to the slope of $L(y|x)$.

Example 2. In the second example, there is again a single x, but the population consists of the twelve equi-probable points displayed in Table 2.

Table 2

x	-2	-2	-2	-1	-1	-1	0	0	0	1	1	1
y	-3	-2	-1	-2	-1	0	-1	0	1	0	1	2

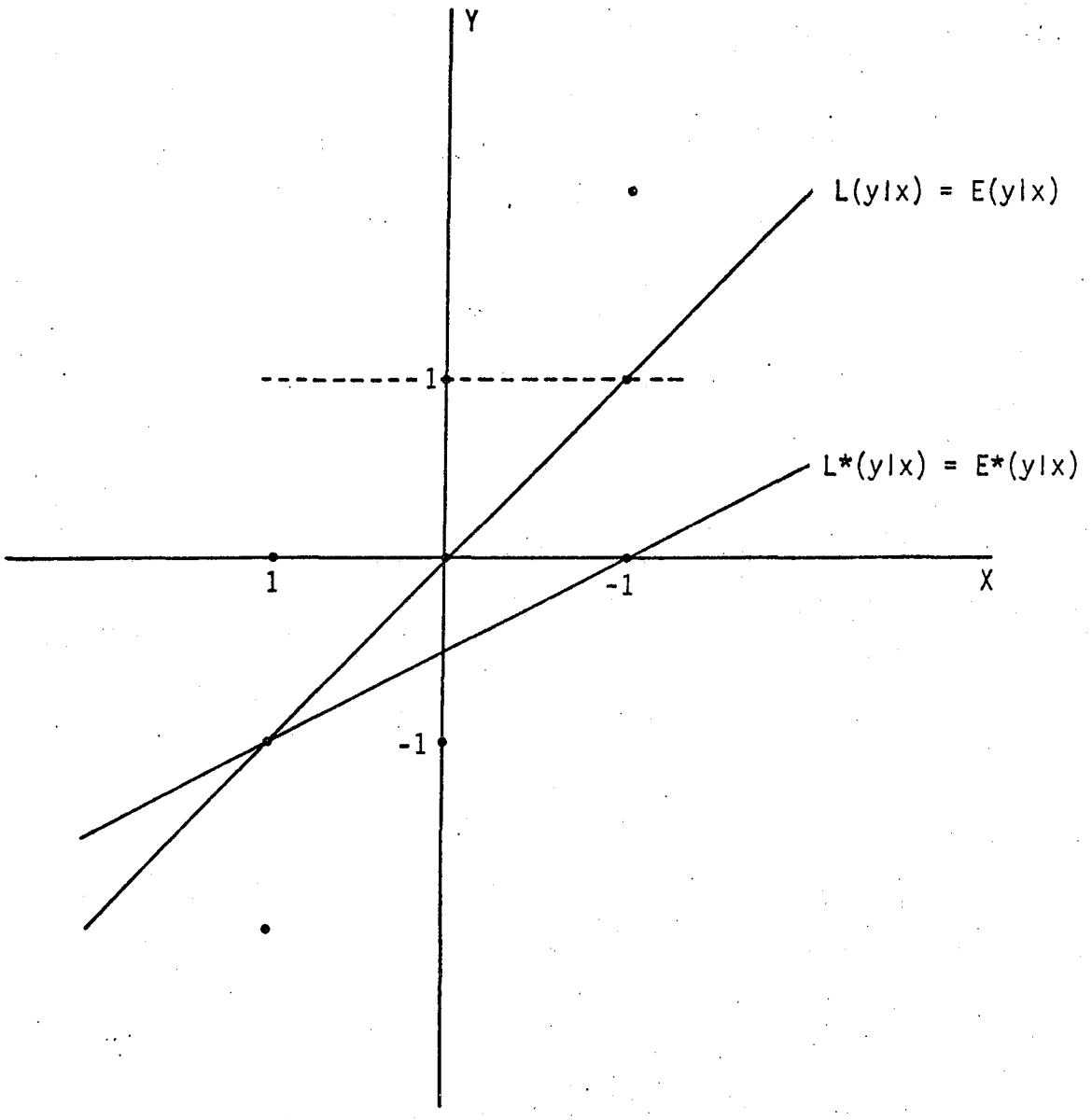

Figure 1

The conditional expectation function is again

$$E(y|x) = x = L(y|x) \tag{10}$$

which appears as the line AC in Figure 2. Again form the truncated population by deleting points at which y is greater than zero. This time, the conditional expectation function in the truncated population $[E^*(y|x)]$ is not linear. This function is labeled ABD in Figure 2. We can readily calculate $L^*(y|x)$, the best linear approximation to $E^*(y|x)$. Using the results

$$E^*(y) = -1.11 \ , \quad E^*(x) = -.889 \ , \quad C^*(x,y) = .679 \ ,$$
$$V^*(x) = .988 \tag{11}$$

we conclude that

$$L^*(y|x) = -.500 + .688x \tag{12}$$

which is graphed in Figure 2 as EF.

As in the first example, linear regression in a sample from the truncated population yields a biased estimate of the slope of $L(y|x)$. If b^* is such a linear regression estimate, we know that

$$E(b^*) = .688 \neq 1 \ . \tag{13}$$

Next, we turn to a simple case of the classical normal regression model where there is a single regressor z with a unit coefficient:

$$y = z + \varepsilon$$
$$\varepsilon \sim N(0,\sigma^2) \ . \tag{14}$$

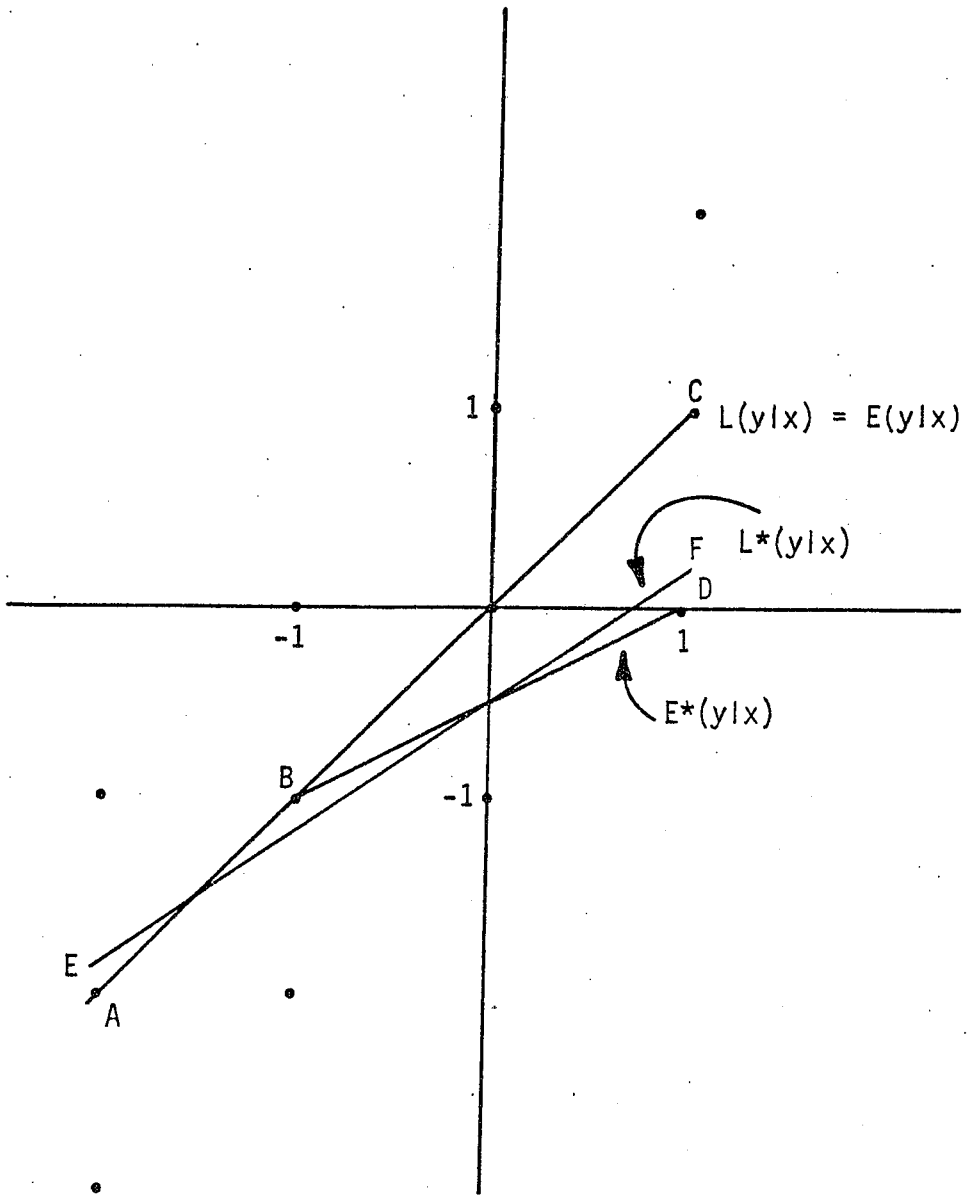

Figure 2

This model implies that the conditional distribution of y given z is normal with mean z and variance σ^2. The probability density function for y is

$$g(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-z)^2}{2\sigma^2}\right) . \tag{15}$$

The conditional expectation function is linear in z so it coincides with $L(y|z)$:

$$E(y|z) = L(y|z) = z . \tag{16}$$

Now, consider the truncated subpopulation which consists of all points at which y is less than some value H. In that subpopulation, the density function for y is

$$g^*(y) = \begin{cases} g(y)/G(H) & \text{for } y < H \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where G is the cumulative normal distribution function

$$G(H) = \int_{-\infty}^{H} g(y)\,dy . \tag{18}$$

In the subpopulation, the conditional expectation of y given z is

$$\begin{aligned} E^*(y|z) &= \int_{-\infty}^{H} y g^*(y)\,dy \\ &= \frac{1}{G(H)} \int_{-\infty}^{H} y g(y)\,dy \\ &= \frac{1}{G(H)} \int_{-\infty}^{H} \left(g(y)z - \sigma^2 \frac{dg(y)}{dy} \right) dy \end{aligned}$$

$$= \frac{z}{G(H)} \int_{-\infty}^{H} g(y)dy - \frac{\sigma^2}{G(H)} \int_{-\infty}^{H} dg(y)$$

$$= z - \sigma^2 \frac{g(H)}{G(H)} \,.^3 \tag{19}$$

To study $E^*(y|z)$ it is convenient to transform variables. Let

$$a = (H-z)/\sigma \tag{20}$$

so

$$z = H - \sigma a \ . \tag{21}$$

Then

$$\sigma \frac{g(H)}{G(H)} = \frac{f(a)}{F(a)} = r(a) \ , \tag{22}$$

say, where $f(\cdot)$ and $F(\cdot)$ are the standard normal density and cumulative functions respectively. Then (19) can be rewritten as

$$E^*(y|z) = z - \sigma r(a) \ . \tag{23}$$

Evans (1975) has calculated $r(a)$ for various values of a. His graph of $r(a)$ is reproduced in Figure 3 and is used to plot $E^*(y|z)$ in Figure 4. Note that $E^*(y|z)$ is a positive, convex function of z which asymptotically approaches $y = z$ as $a \to -\infty$ and $y = L$ as $a \to \infty$.[4]

The best linear approximation to $E^*(y|z)$ is $L^*(y|z)$ which has slope equal to $C^*(z,y)/V^*(z)$. Using

$$C^*(z,y) = C^*[z,E^*(y|z)] \tag{24}$$

and (23), we obtain

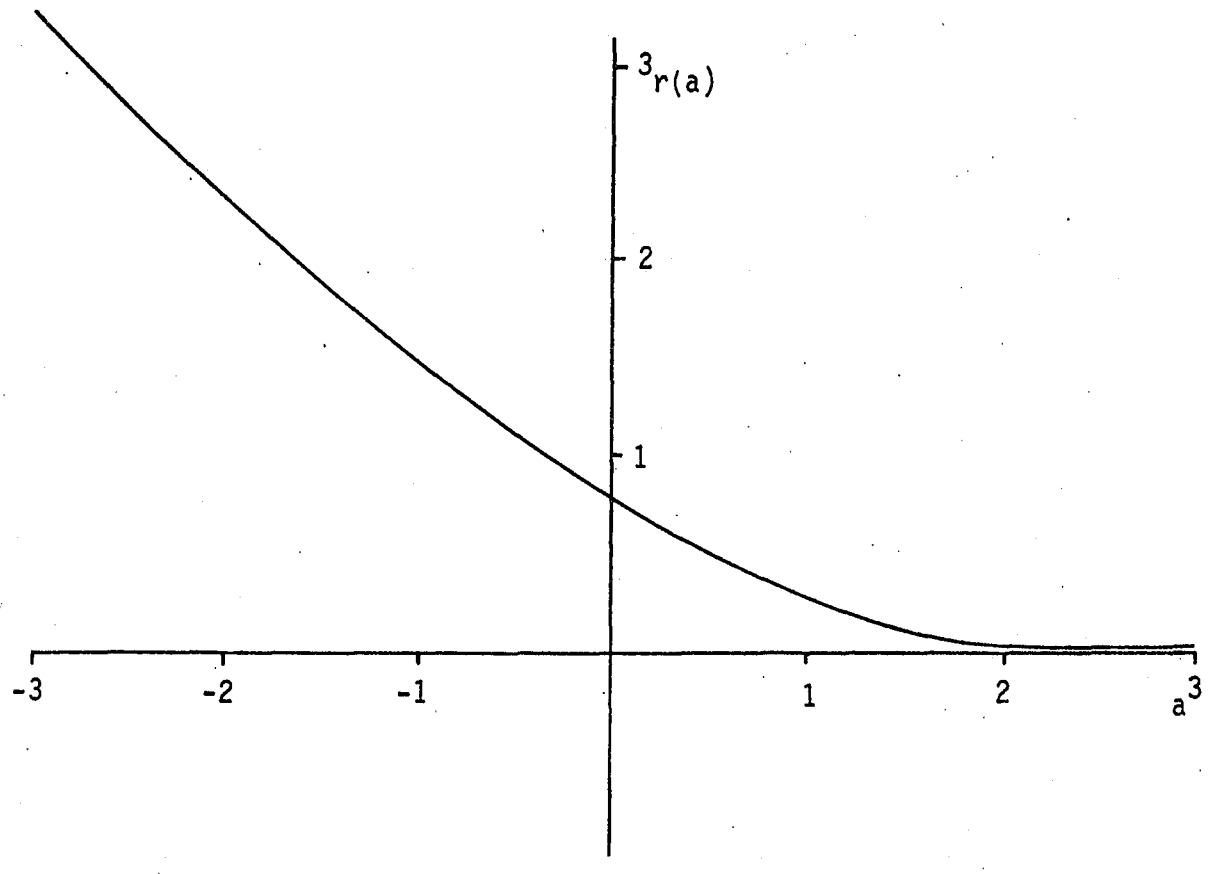

Figure 3

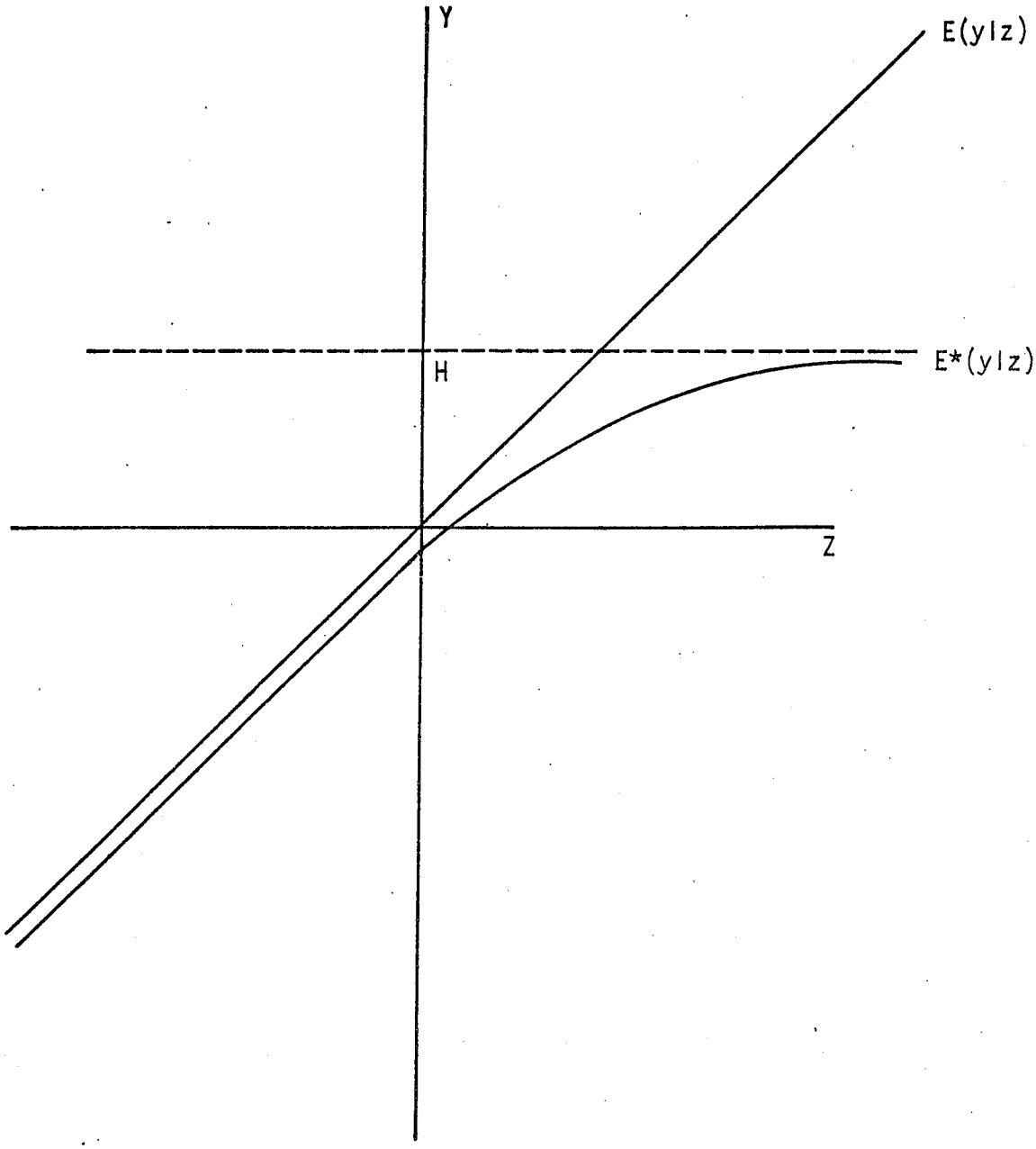

Figure 4

$$C^*(z,y) = C^*[z, z - \sigma r(a)] \tag{25}$$

which can be rewritten as

$$C^*(z,y) = V^*(z) - \sigma C^*[z, r(a)]. \tag{26}$$

Note that r(a) is monotonically decreasing in a and therefore monotonically increasing in z. The sign of $C^*[z,r(a)]$ can be determined using a theorem due to Gurland (1967) regarding the covariance of functions of random variables. Gurland's theorem says that two monotonically increasing (or decreasing) functions of the same random variable will have a positive covariance. Since z and r(a) are two monotonically increasing functions of z, $C^*[z, r(a)]$ is positive and

$$\beta^* = \frac{C^*(z,y)}{V^*(z)} \leq \frac{C(z,y)}{V(z)} = 1 \ . \tag{27}$$

Once again we see that a linear regression coefficient estimate from a sample drawn from a truncated subpopulation will be biased for the slope of $L(y|z)$ because $L^*(y|z)$ has a different slope. Next, we generalize this result to the case of multiple regressors.

Consider the model

$$y = \underline{x}'\underline{\beta} + \varepsilon$$
$$\varepsilon \sim N(0,\sigma^2) \ ,$$

where $\underline{x}$ is a k x 1 vector of exogenous variables. The results contained in our last example can in large part be generalized to cover this case by defining

$$z \equiv \underline{x}'\underline{\beta} \ . \tag{29}$$

The slope vector of $L^*(y|\underline{x})$ is

$$\underline{\beta}^* = [V^*(\underline{x})]^{-1}C^*(\underline{x},y) = [V^*(\underline{x})]^{-1}\underline{C}^*[\underline{x},E^*(y|\underline{x})] \tag{30}$$

where $V^*(\underline{x})$ is the variance matrix of $\underline{x}$ and $\underline{C}^*(\underline{x},y)$ is a column vector of the covariances of y with each x_i. Using the fact that

$$E^*(y|\underline{x}) = \underline{x}'\underline{\beta} - \sigma r(a) \tag{31}$$

where

$$a = \frac{H - \underline{x}'\underline{\beta}}{\sigma}, \tag{32}$$

(30) can be rewritten

$$\underline{\beta}^* = [V^*(\underline{x})]^{-1}V^*(\underline{x})\underline{\beta} - \sigma[V^*(\underline{x})]^{-1}\underline{C}^*[\underline{x},r(a)] \tag{33}$$

or

$$\underline{\beta}^* = \underline{\beta} - \sigma[V^*(\underline{x})]^{-1}\underline{C}^*[\underline{x},r(a)] \tag{34}$$

where $\underline{C}^*[\underline{x},r(a)]$ is a vector of covariances of the individual x's with r(a). In this general case, linear regression coefficient estimates based upon truncated samples will in general be biased estimates of $\underline{\beta}$, but the signs of biases are ambiguous. In the case of two x's, for example, (34) reduces to

$$\begin{pmatrix}\beta_1^* \\ \beta_2^*\end{pmatrix} = \begin{pmatrix}\beta_1 - \dfrac{\sigma}{|V^*(\underline{x})|}\left[V^*(x_2)C^*[x_1,r(a)] - C^*(x_1,x_2)C^*[x_2,r(a)]\right] \\ \beta_2 - \dfrac{\sigma}{|V^*(\underline{x})|}\left[V^*(x_1)C^*[x_2,r(a)] - C^*(x_1,x_2)C^*[x_1,r(a)]\right]\end{pmatrix} \tag{35}$$

The signs of bias will obviously depend upon the signs of the three covariances.[5]

In this section, we have considered several examples of the biased coefficient estimates which result when linear regression is applied to truncated samples. We have found it useful to think of truncated samples as random samples from truncated populations. Our conclusions are that linear regression in truncated samples provides biased estimates of the slopes of $L(y|\underline{x})$ and that the directions of such biases are, in general, ambiguous.

3. Consistent Estimation Techniques for Truncated Samples

In this section, two methods for obtaining consistent estimates from a truncated sample are considered. The model is that of (28), which we rewrite as

$$y_t = \underline{x}_t'\underline{\beta} + \varepsilon_t \qquad t = 1, \ldots, T \tag{36}$$

$$\varepsilon_t \sim N(0,\sigma^2) .$$

The sample consists of all observations for which

$$y_t < H_t \qquad t = 1, \ldots, T . \tag{37}$$

The limit value H_t may vary across observations, but it is known at each observation.

a. Instrumental Variable Approach

Amemiya (1973) studies the model:

$$y_t^o = \begin{cases} \underline{\gamma}'\underline{x}_t^o + \varepsilon_t^o & \text{if } \underline{\gamma}'\underline{x}_t^o + \varepsilon_t^o > 0 \\ 0 & \text{otherwise} \end{cases} \qquad t = 1, \ldots, T . \tag{38}$$

$$\varepsilon_t^o \sim N(0,\omega^2)$$

where y_t^o is the observed value of the dependent variable at observation t, $\underline{x}_t^o$ is a vector of exogenous variables, and ε_t^o is a normal disturbance which is independent of $\underline{x}_t^o$. This is the specific case of the Tobit model (Tobin, 1958) in which the lower bound is equal to zero at each observation.

Amemiya obtains consistent estimates of this model using only the non-limit observations. He shows that

$$E^*(y_t^o) = \underline{\gamma}'\underline{x}_t^o + \omega r\left(\frac{\underline{\gamma}'\underline{x}_t^o}{\omega}\right) \tag{39}$$

and that

$$E^*[(y_t^o)^2] = (\underline{\gamma}'\underline{x}_t^o)^2 + \omega\underline{\gamma}'\underline{x}_t^o r\left(\frac{\underline{\gamma}'\underline{x}_t^o}{\omega}\right) + \omega^2$$

$$= \underline{\gamma}'\underline{x}_t^o E^*(y_t^o) + \omega^2 \tag{40}$$

where r is defined in (22). Defining

$$u_t \equiv y_t^o - E^*(y_t^o) \tag{41}$$

and

$$v_t \equiv (y_t^o)^2 - E^*[(y_t^o)^2] \tag{42}$$

(40) implies

$$(y_t^o)^2 = \underline{\gamma}'\underline{x}_t^o(y_t^o - u_t) + \omega^2 + v_t \qquad y_t^o > 0$$

$$= \underline{\gamma}'\underline{x}_t^o y_t^o + \omega^2 + \eta_t \qquad y_t^o > 0 \qquad (43)$$

where $\eta_t = v_t - \underline{\gamma}'\underline{x}_t^o u_t$. (44)

Amemiya uses (43) to obtain estimates of $\underline{\gamma}$ and ω^2. The correlation between η_t and y_t precludes linear regression, so Amemiya proposes instrumental variable estimation of (43) with $(\underline{x}_t^o \hat{y}_t^o, 1)$ as instruments

$$\begin{bmatrix} \hat{\underline{\gamma}} \\ \hat{\omega}^2 \end{bmatrix} = \left[\Sigma^+ \begin{pmatrix} \underline{x}_t^o \hat{y}_t^o \\ 1 \end{pmatrix} (\underline{x}_t^o y_t^o,\ 1) \right]^{-1} \Sigma^+ \begin{pmatrix} \underline{x}_t^o \hat{y}_t^o \\ 1 \end{pmatrix} (y_t^o)^2 \qquad (45)$$

where

$$\hat{y}_t^o = \underline{x}_t^{o\prime} \left(\Sigma^+ \underline{x}_t^o \underline{x}_t^{o\prime}\right)^{-1} \Sigma^+ \underline{x}_t^o y_t^o \qquad (46)$$

and where Σ^+ indicates the sum over all observations at which y_t^o is greater than zero. He shows that $\hat{\underline{\gamma}}$ and $\hat{\omega}^2$ are consistent for $\underline{\gamma}$ and ω^2 under general conditions.

The model displayed in (44) can be written in the form of (46) by making the substitutions

$$y_t^o = L_t - y_t \ , \qquad (47)$$

$$\underline{x}_t^o = \begin{pmatrix} L_t \\ -\underline{x}_t \end{pmatrix} , \qquad (48)$$

$$\underline{\gamma} = \begin{pmatrix} 1 \\ \underline{\beta} \end{pmatrix} , \qquad (49)$$

$$\varepsilon_t^o = -\varepsilon_t \quad , \tag{50}$$

$$\omega^2 = \sigma^2 \quad . \tag{51}$$

So, consistent instrumental variable estimates of our model can be obtained using Amemiya's instrumental variable technique. First, the y_t's are regressed on the $\underline{x}_t$'s, and the predicted values from the regression, $\hat{y}_t$'s, are used to compute $(H_t - \hat{y}_t)$ for each t. Next the vectors $\underline{c}_t$ and $\underline{d}_t$ are constructed where

$$\underline{c}_t = (H_t - y_t) \begin{pmatrix} L_t \\ \underline{x}_t \end{pmatrix} \tag{52}$$

and

$$\underline{d}_t = (H_t - \hat{y}_t) \begin{pmatrix} L_t \\ \underline{x}_t \end{pmatrix} \quad . \tag{53}$$

Each element of $\underline{c}_t$ is regressed on $\underline{d}_t$ and a constant to obtain a vector of predicted values $\hat{\underline{c}}_t$. Finally, we regress $(y_t)^2 - \hat{c}_1$ on $(\hat{c}_2, \ldots, \hat{c}_{k+1})$ and a constant; the coefficients obtained for $(\hat{c}_2, \ldots, \hat{c}_{k+1})$ will be $-\hat{\underline{\beta}}$ and the constant coefficient will be $\hat{\sigma}^2$. This procedure takes advantage of the familiar computational technique for obtaining instrumental variable estimates using a two-stage linear regression algorithm.

b. Maximum Likelihood Approach

Amemiya (1973, p. 1000-1001) also spells out the maximum likelihood method for estimation of the model in (38) which can be adapted as follows. The model displayed in (36) implies that y_t is normally

distributed in the full population with mean $\underline{x}_t'\underline{\beta}$ and variance σ^2. It follows then that the density function of y_t conditional on y_t being less than H_t is

$$g^*(y_t) = \begin{cases} g_t(y_t)/G_t(H_t) & \text{for } y_t < H_t \\ 0 & \text{otherwise} \end{cases} \tag{54}$$

where g_t is the normal density function with mean $\underline{x}_t'\underline{\beta}$ and variance σ^2, and G_t is the corresponding cumulative distribution function. Since the truncated sample is a random sample from the truncated population characterized by (54) the likelihood function of the sample can be written as

$$\Phi = \prod_{t=1}^{T} g^*(y_t) = \prod_{t=1}^{T} \left(\frac{g_t(y_t)}{G_t(H_t)} \right) . \tag{55}$$

We wish to choose those estimates of $\underline{\beta}$ and σ^2 which maximize (55) or, equivalently, which maximize

$$\ln\Phi = -\Sigma \ln G_t - \frac{T}{2} \ln 2\pi - \frac{T}{2} \ln\sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \underline{x}_t'\underline{\beta})^2 \tag{56}$$

where

$$G_t = G_t(H_t) . \tag{57}$$

The values which maximize (56) will be a solution to the first order conditions

$$\frac{\partial \ln\Phi}{\partial \underline{\beta}} = \sum_{t=1}^{T} \left[\frac{g_t}{G_t} + \frac{1}{\sigma^2} (y_t - \underline{x}_t'\underline{\beta}) \right] \underline{x}_t = 0 \tag{58}$$

$$\frac{\partial \ln \Phi}{\partial \sigma^2} = \frac{1}{2\sigma^2} \sum_{t=1}^{T} \frac{(H_t - \underline{x}_t'\underline{\beta})g_t}{G_t} - \frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^{T} (y_t - \underline{x}_t'\underline{\beta})^2 = 0$$

where

$$g_t = g_t(H_t) \ . \tag{59}$$

The second derivatives of (58) are displayed in the appendix.

For a given set of data, we wish to find a solution to equations (58) at which the likelihood function takes on its maximum value. To do so, an algorithm for the maximization of a function of several variables is required. The method used here was developed by Davidon, described by Fletcher and Powell (1963), and made operational by Gruvaeus and Jöreskog (1970). The method searches for a maximum of the likelihood function using analytical first derivatives of the function and an approximation to the matrix of second derivatives, beginning with user supplied starting values of the parameters. We will use the instrumental variable estimates as these starting values. Having found the maximum likelihood estimates of $\underline{\beta}$ and σ^2, we estimate the asymptotic covariance matrix of these estimates as the negative of the inverse of the matrix of second derivatives of $\ln\Phi$. Under certain regularity conditions, the maximum likelihood estimates of $\underline{\beta}$ and σ^2 are asymptotically efficient and asymptotically normally distributed with means $\underline{\beta}$ and σ^2 and variances estimated as noted above.

In this section two techniques for the estimation of simple linear models from truncated samples have been described. The maximum likelihood estimates have more desirable asymptotic properties, but the choice of techniques in finite samples remains open. The instrumental

variable estimates have the practical advantage of being relatively easy to compute. Both techniques are used in the following sections.

4. A Model of the Determination of Earned Income

In order to get an empirical picture of the relation between full-sample and truncated-sample results we estimate an earnings function using a full random sample and a truncated subsample. The model to be estimated determines earned income of male heads of households. The earnings function is not a structural equation but, rather, represents a reduced form equation associated with an unspecified structural system of the demand for and supply of labor. The arguments of the earnings function are exogenous variables which affect demand and/or supply. Among these variables are factors which determine the individual's productive potential such as his age, his education, and his intelligence. Other arguments of the earnings function are family size which should affect labor supply, race which captures earnings differences due to discrimination, and region which should measure earnings differences attributable to regional variations in the productivity of labor and the cost of living.

Our model of earnings is

$$E = \beta_o + \beta_1 \text{ GRSCH} + \beta_2 \text{ HSCH} + \beta_3 \text{ SOME} + \beta_4 \text{ GRAD} + \beta_5 \text{ TEST} + \beta_6 \text{ AGE} + \beta_7 \text{ AGESQ} + \beta_3 \text{ RACE} + \beta_9 \text{ SOUTH} + \beta_{10} \text{ FSZ} + \varepsilon \quad (60)$$

where

E = individual's labor income in 1968 (includes imputed earnings of self-employed)

GRSCH = $\begin{cases} 1 \text{ if individual did not enter high school} \\ 0 \text{ otherwise (base category is high school graduates with no further formal education)} \end{cases}$

HSCH = $\begin{cases} 1 \text{ if individual entered high school but did not graduate} \\ 0 \text{ otherwise} \end{cases}$

SOME = $\begin{cases} 1 \text{ if individual graduated from high school, had further formal training, but did not graduate from college} \\ 0 \text{ otherwise} \end{cases}$

GRAD = $\begin{cases} 1 \text{ if individual graduated from college} \\ 0 \text{ otherwise} \end{cases}$

TEST = individual's score on a sentence completion (IQ) test[6] (scores range from 0 to 13)

AGE = individual's age in years

AGESQ = $(\text{AGE})^2$

RACE = $\begin{cases} 1 \text{ if individual is nonwhite} \\ 0 \text{ if individual is white} \end{cases}$

SOUTH = $\begin{cases} 1 \text{ if individual lives in the Southern Census Region} \\ 0 \text{ otherwise} \end{cases}$

FSZ = number of persons in the individual's household

and ε is a normal disturbance with mean zero and variance σ^2. We assume that ε is independent of the exogenous variables in the model.

Two characteristics of the functional form are worthy of note. First, the model contains four dichotomous variables which measure the individual's education. The education variable was originally available in nine categories. For the sake of parsimony, we have collapsed these nine categories into five categories. Second, we have specified a

quadratic form for the partial impact of age upon earnings because most researchers have found age to have a first positive and then negative effect upon earnings, a result which has theoretical appeal.

5. Data

The data used in this analysis were taken from the five-year panel collected by the Survey Research Center (SRC) of the Institute for Social Research at the University of Michigan. This data set consists of five annual interviews with approximately 5000 family units, 40 percent of which were selected from members of the Survey of Economic Opportunity (SEO) sample. The remaining 60 percent of the sample units make up a national cross-sectional sample. In the present paper the SEO follow-up sample is excluded. One-half of the national cross-sectional sample has been randomly selected for analysis; the remainder is available for future analysis. Finally, attention is restricted to male-headed families which had the same head over the entire five-year period; a sample of 864 families is available. This sample is assumed to be a random sample from the full population of male heads of households.

A truncated subsample was constructed from the sample of 864 observations by excluding an observation if the family's income was more than 150 percent of the family's poverty threshold ("Orshansky Ratio" greater than 1.5). For a given family size this restriction imposed an upper bound on total family income. A group of 253 observations remained. The sample is truncated on the basis of total family income and the model explains only head's labor income, so we must

translate the truncation on family income into a truncation on head's labor income. To do so, it is assumed that all other components of family income are fixed. The upper bound on family income in the truncated sample is 150 percent of the family's poverty threshold. The upper bound on head's labor income is taken to be the family income bound minus family income from sources other than the head's labor income.

The full and truncated sample means and standard deviations of the variables in our model are displayed in Table 3. Note that men in the truncated sample have, on average, much lower earnings, much less education, and only slightly lower intelligence test scores. The average man in the truncated sample is slightly older than the average man in the full sample. Further, we see that the proportions of nonwhites and of southerners are higher in the truncated sample. Finally, there is the interesting result that the average family size is virtually the same in the two samples. One might have thought that larger families with higher poverty thresholds would be over-represented in the truncated sample, but such is not the case. Apparently, larger families have sufficiently higher incomes that they are no more likely to fall below the bound of 150 percent of their poverty thresholds.

6. Estimation of the Model

Four sets of estimates of the parameters of the earnings model shown in (60) are described in this section. The first set, hereafter denoted LR-F, was obtained with the full sample using linear regression. The remaining three sets of estimates were obtained with the truncated

Table 3

Sample Means and Standard Deviations from the Full Sample and the Truncated Sample

	Full Sample Mean	Full Sample Standard Deviation	Truncated Sample Mean	Truncated Sample Standard Deviation
E	7100.	5920.	2340.	2410.
GRSCH	.264	.441	.482	.501
HSCH	.153	.360	.150	.358
SOME	.244	.430	.194	.396
GRAD	.152	.359	.059	.237
TEST	9.72	2.16	8.89	2.51
AGE	45.2	15.7	50.6	19.3
RACE	.101	.301	.174	.380
SOUTH	.338	.473	.466	.500
FSZ	3.47	1.73	3.42	2.17
Sample Size	864		253	

sample using linear regression (LR-T), the instrumental variable technique (IV-T), and the maximum likelihood technique (ML-T).[7] These four sets of estimates are displayed in Table 4; estimated standard errors are reported in parentheses.

First, consider the LR-F estimates. The education variables appear to have a strong impact upon earnings. It is somewhat surprising that the coefficients of HSCH and SOME are not significantly different from zero. This result implies that the earnings of high school graduates are not significantly different from the earnings of those who completed only some high school or from those who graduated from high school, had some formal training beyond high school, but did not graduate from college. The hypothesis that the coefficients of the four education variables are jointly equal to zero can be tested with an F-test. The F-statistic with 4 and 853 degrees of freedom is 32.9 which allows the hypothesis to be rejected at the .001 level of significance. The age-earnings profile implied by the model exhibits a positive effect of increasing age until age 46. The variable RACE has a negative but insignificant impact upon earnings. The lack of significance may be due to the heterogeneity of the nonwhite group which includes Orientals as well as Blacks and people of Spanish heritage. The variable SOUTH has a strong negative impact upon earnings. The family size variable has a strongly significant and positive impact upon earnings, which might suggest that heads with responsibilities for larger families are likely to be more highly motivated to work or that richer people have more children, in which case the model is mis-specified. The

Table 4

Estimates of the Earnings Model
(Standard errors in parentheses)

	Full Sample	Truncated Sample		
	Linear Regression	Linear Regression	Instrumental Variable	Maximum Likelihood
CONSTANT	-10,100. (1,600.)	47.9 (957.)	4,220. (4,420.)	821. (1,640.)
GRSCH	-2,410. (524.)	-879. (342.)	-2,850. (2,020.)	-1,460. (588.)
HSCH	-563. (558.)	-172. (401.)	-1,010. (2,000.)	-840. (692.)
SOME	487. (494.)	213. (382.)	178. (2,220.)	702. (711.)
GRAD	4,140. (561.)	152. (521.)	3,680. (2,670.)	605. (976.)
TEST	275. (85.5)	-27.9 (47.1)	44.2 (180.)	-22.0 (72.1)
AGE	680. (62.4)	110. (37.2)	231. (170.)	172. (62.4)
AGESQ	-7.45 (.656)	-1.50 (.373)	-2.84 (1.67)	-2.27 (.605)
RACE	-764. (560.)	-273. (292.)	-864. (926.)	-418. (448.)
SOUTH	-909. (350.)	-25.9 (215.)	-239. (794.)	-52.8 (353.)
FSZ	344. (104.)	545. (56.2)	-53.4 (258.)	373. (89.8)
$\hat{\sigma}$	4,710.	1,580.	3,140.	1,950. (136.)
R^2	.367	.569	-	-

intelligence test variable has a positive coefficient which is different from zero at the .01 level of significance. Finally, it should be noted that this model fits the data remarkably well for a cross-sectional individual model as indicated by the R^2 of .367.

The three sets of estimates of the coefficients of our model based upon the truncated sample differ considerably. The linear regression estimates are inconsistent for $\underline{\beta}$ and σ^2, but the biases cannot be calculated explicitly as the true values of the parameters are unknown. In addition, the calculated standard errors of the linear regression estimates are spurious. It is interesting nevertheless to compare the three sets of estimates from the truncated sample with the estimates from the full sample.

First, comparing the linear regression estimates from the two samples, note that the sign of each coefficient is preserved except for the TEST coefficient. Second, of the remaining coefficient estimates, the LR-T estimates are closer to zero than the LR-F estimates with the single exception of the coefficient on FSZ. This single exception makes sense if we remember that larger families had higher income cut-offs due to the truncation on the "Orshansky Ratio." Third, the ordering of the estimates of the education coefficients are different in the two samples. In the full sample, a monotonically increasing relationship exists between education and earnings, but in the truncated sample it appears that college graduation reduces earnings. Fourth, the age-earnings profile based on the truncated sample is flatter and has an earlier peak at age 37. Finally, the conventional estimate of the standard deviation of ε is smaller in the truncated sample, and the R^2 is larger. Generally, it seems that

linear regression on the truncated sample yields quite unsatisfactory estimates of the coefficients of the earnings function, judging on the basis of comparison with the LR-F estimates.

The instrumental variable estimates of the model seem to be more satisfactory. The sign of each coefficient is the same as in the full sample regression, except for FSZ, a somewhat puzzling exception. The magnitudes of the IV-T coefficients tend to be closer to the magnitudes of the LR-F estimates than are those of the LR-T estimates. In spite of these similarities the instrumental variable estimates are substantially different from the full sample estimates, but there is no apparent pattern to the differences. Once again, we see a monotonic relationship between earnings and education.

Turning to the maximum likelihood estimates, we see that the estimated TEST coefficient is, once again, negative while the estimated FSZ coefficient is positive. The signs of the other maximum likelihood estimates correspond to the full sample estimates. The ordering of the estimated coefficients of SOME and GRAD are as they were in the set of linear regression estimates from the truncated sample, implying a non-monotonic relationship between earnings and education. Generally, the ML-T estimates are closer to the LR-F estimates than are the LR-T estimates but not so close as are the instrumental variable estimates. To test the hypothesis that the coefficients of the education variables are jointly equal to zero, it is convenient to use a likelihood ratio test. Recall that

$$-2 \ln \left(\frac{\Phi_{H0}}{\Phi_{H1}} \right) \sim \chi^2_k \qquad (61)$$

where Φ_{H0} is the value of the likelihood function under the null hypothesis, Φ_{H1} is the value of the likelihood function under the alternative hypothesis, and k is the number of restrictions imposed by the null hypothesis, in this case four. The χ^2 statistic is 220, which allows the hypothesis that education has no effect on earnings to be rejected at the .005 level of significance.

Our maximum likelihood estimates are based upon two assumptions: that there is no systematic difference between the truncated and full populations apart from that due to truncation and that the stochastic term in that function has a normal distribution conditional on the values of the exogenous variables. We can test the first assumption under the maintained hypothesis that the second is correct by means of a likelihood ratio test. In order to perform this test we evaluate the likelihood for the truncated sample using the LR-F and ML-T estimates of $\underline{\beta}$ and σ^2. When we do so we obtain a very large χ^2 statistic with twelve degrees of freedom of 1642 which forces the rejection of the first assumption (conditional on the normality of ε) at a level of significance very close to zero.

If the second assumption is incorrect, however, this test is difficult to interpret. What is required is a test of the conditional normality of the disturbance term in the full sample. The predicted values ($\hat{E}$'s) and residuals associated with the LR-F estimates were computed. The observations were then grouped into deciles on the basis of the predicted values. Within each decile a Pearson χ^2 test of the hypothesis that the residuals were distributed normally was performed (Blum and Rosenblatt, 1972, pp. 408-409). The residuals in each decile were sorted into eight cells for this test, so the χ^2

statistic has seven degrees of freedom. The cell boundaries were set such that the probability of a residual falling into any cell, under the null hypothesis, was .125. Table 5 reports the cell frequencies of the residuals and the χ^2 statistic for each decile. The 5 percent critical value for a χ^2 with seven degrees of freedom is 14.1, and the 1 percent critical value is 18.5. The assumption of normality cannot be rejected for five deciles at the 1 percent level of significance and for two deciles at the 5 percent level. One would have to agree that the assumption of normality is somewhat tenuous.

7. Conclusions and Extensions

In this paper the problems encountered in using linear regression to estimate simple linear models from truncated samples have been examined. Two techniques for obtaining consistent estimates in such situations were described. Finally, these techniques were used to estimate an earnings function with an artificially truncated sample. There are two lessons to be learned from this exercise.

First, it is clear that linear regression will not, in general, provide consistent estimates of linear models when samples are truncated on the basis of the dependent variable. Empirically, it appears that this problem may be of substantial magnitude as it seemed to be in the example.

Second, we cannot be too optimistic about the possibility of using the Rural Income Maintenance Experiment data to estimate the earnings model as it now stands. We can obtain consistent estimates from truncated samples only if the model is correctly specified, and

Table 5

Goodness of Fit Tests of the Normality of the Disturbance in the Earnings Equation

Decile of $\hat{E}$ (Low to High)	Cell Frequencies (Low to High)								χ^2
	1	2	3	4	5	6	7	8	
1	.00	.00	.19	.24	.22	.10	.19	.06	46.1
2	.00	.13	.21	.21	.19	.15	.10	.01	32.7
3	.00	.17	.16	.23	.21	.14	.05	.03	36.2
4	.05	.19	.19	.23	.13	.16	.03	.02	31.0
5	.03	.13	.23	.19	.20	.16	.05	.01	33.8
6	.06	.13	.14	.22	.16	.15	.08	.06	15.4
7	.06	.13	.15	.21	.21	.09	.07	.08	17.4
8	.13	.09	.16	.23	.17	.05	.09	.07	18.4
9	.09	.09	.16	.17	.13	.19	.06	.10	10.0
10	.10	.17	.19	.09	.09	.11	.08	.18	10.3

this model appears to be misspecified. Nevertheless, techniques have been proposed which will provide consistent and possibly asymptotically efficient estimates of correctly specified models using truncated samples. Obviously, it is best to use full samples when they are available, but truncated samples contain information which economists cannot afford to waste.

The maximum likelihood technique developed in this paper can be easily extended to the estimation of multiple equation models. Jöreskog (1973) has developed a maximum likelihood technique for the estimation of a general linear model. Most types of linear models can be expressed in the form of Jöreskog's model. The model implies a joint multinormal distribution of observed exogenous ($\underline{x}$) and endogenous ($\underline{y}$) variables. Each specific model implies a set of restrictions on the covariance matrix of the variables, as the elements of the matrix are functions of the parameters of the model. These parameters can be estimated from a random sample via the maximization of the likelihood function

$$\Phi = \prod_{t=1}^{T} m(\underline{x}_t, \underline{y}_t) \tag{62}$$

where m is the multinormal density function. Consider a sample that has been truncated when y_{1_t} exceeded some predetermined value, H_t. The likelihood function for the Jöreskog model would then be

$$\Phi^* = \prod_{t=1}^{T^*} \frac{m(\underline{x}_t, \underline{y}_t)}{M_1(H_t)} \tag{63}$$

where M_1 is the marginal cumulative distribution of y_1. Estimates of this model can be obtained from truncated samples by finding parameter values which maximize (63).

Having developed techniques for the estimation of multiple equation models from truncated samples, we can formulate a more plausible structural model of the determination of earnings. One such model might be

$$\ln S = \alpha \ln W + \underline{\beta}'\underline{Z}_1 + \varepsilon \tag{64}$$

$$\ln W = \underline{\gamma}'\underline{Z}_2 + \omega \tag{65}$$

$$\ln E = \ln S + \ln W \tag{66}$$

where S is hours worked, W is the individual's wage rate, E is the individual's earnings, and $\underline{Z}_1$ and $\underline{Z}_2$ are vectors of exogenous variables. Equation (64) is the demand function facing the individual and equation (65) is the supply function of the individual. The identity (66) permits the possibility of an earnings truncation in this model.

The multiple equation technique can also be used for the estimation of models which more fully exploit panel data. One such model is

$$Y^* = \underline{\beta}'\underline{Z} + \varepsilon$$

$$Y_j = Y^* + v_j \qquad j = 1, \ldots, J \tag{67}$$

where $\underline{Z}$ is a vector of exogenous variables, Y* is permanent income, and Y_j is measured income in period j. If such a model is to be estimated with income maintenance experiment data, then the truncation on first period income should not be ignored. The technique suggested above will allow us to account for the truncation.

NOTES

[1]After completing an earlier draft of this paper, I became aware of a paper by Hausman and Wise (1975) which deals with this same subject in a similar fashion.

[2]This density function may remind the reader of the Tobit model developed by Tobin (1958) and analyzed by Amemiya (1973). The difference between the two models is that the Tobit model implies a nonzero probability of limit observations while our model implies a zero probability of such observations.

[3]Johnson and Kotz (1970, Vol. 1, p. 81) display the expected value of the doubly truncated normal distribution, which can be simplified to (19) when the lower truncation point is minus infinity.

[4]In Figure 3, we see that r(a) is a decreasing function of a, but we can obtain this result analytically. Differentiating r(a) we obtain

$$\frac{d[r(a)]}{da} = \frac{1}{[F(a)]^2}\left(-aF(a)f(a) - [f(a)]^2\right)$$

$$= -\frac{f(a)}{F(a)}\left[a + \frac{f(a)}{F(a)}\right]$$

$$= -r(a)[a + r(a)] \; . \qquad \text{(F1)}$$

Since r(a) is strictly positive everywhere, the sign of (F1) will be determined by the sign of [a + r(a)]. When a is greater than or equal to zero [a + r(a)] is obviously positive. In the case where t is less than zero, we define

$$u = -a \tag{F2}$$

so that

$$a + r(a) = -u + r(-u) \quad , \quad u > 0 \; . \tag{F3}$$

Using the symmetry of the normal distribution we conclude that

$$r(-u) = \frac{f(-u)}{F(-u)} = \frac{f(u)}{1-F(u)} = \frac{1}{R(u)} \tag{F4}$$

where R(u) is Mills' Ratio (See Johnson and Kotz, Vol. 2, p. 278). So for a less than zero

$$a + r(a) = -u + \frac{1}{R(u)} \; , \qquad u > 0 \; . \tag{F5}$$

Since Mills' Ratio satisfies the inequality (Johnson and Katz, Vol. 2, p. 279)

$$0 < R(u) < \frac{1}{u} \qquad u > 0 \tag{F6}$$

we know that

$$\frac{1}{R(u)} > u \qquad u > 0$$

$$-u + \frac{1}{R(u)} > 0 \qquad u > 0$$

$$a + r(a) > 0 \qquad u > 0 \; . \tag{F7}$$

Having established that a + r(a) is strictly positive everywhere, we conclude that d[r(a)]/da is negative for all values of a.

[5] Goldberger (1975) has shown that the signs of the biases are ambiguous.

[6]This test consisted of 13 items selected from the verbal part of the Lorge-Thorndike Intelligence Test.

[7]Linear regression estimates were obtained using the Regan 2 program in the Statjob series. Instrumental variable estimates were calculated using the Time Series Processor available from DACC. Maximum-likelihood estimates were obtained using a minimization program written by Gruvaeus and Jöreskog and adapted by me.

APPENDIX

To derive the first and second order conditions for the maximization of (64) we have made use of several results:

$$\frac{\partial G_t}{\partial \underline{\beta}} = -g_t \underline{x}_t \tag{A1}$$

$$\frac{\partial G_t}{\partial \sigma^2} = -\frac{1}{2\sigma^2}(L_t - \underline{x}_t'\underline{\beta})g_t \tag{A2}$$

$$\frac{\partial g_t}{\partial \underline{\beta}} = \frac{1}{\sigma^2}(L_t - \underline{x}_t'\underline{\beta})g_t\underline{x}_t \tag{A3}$$

$$\frac{\partial g_t}{\partial \sigma^2} = g_t\left[\frac{(L_t - \underline{x}_t'\underline{\beta})^2 - \sigma^2}{2\sigma^4}\right] \tag{A4}$$

The first order conditions are displayed in (66). The second derivatives are as follows:

$$\frac{\partial^2 \ln\Phi}{\partial\underline{\beta}\partial\underline{\beta}'} = \sum_{t=1}^{T}\left[\frac{g_t}{G_t^2}\left(g_t + \frac{G_t}{\sigma^2}\left[L_t - \underline{x}_t'\underline{\beta}\right]\right) - \frac{1}{\sigma^2}\right]\underline{x}_t\underline{x}_t'$$

$$\frac{\partial^2 \ln\Phi}{\partial\sigma^2\partial\underline{\beta}} = \sum_{t=1}^{T}\left[\frac{g_t}{2\sigma^2 G_t^2}\left(\frac{G_t}{\sigma^2}\left[L_t - \underline{x}_t'\underline{\beta}\right]^2 - G_t + g_t\left[L_t - \underline{x}_t'\underline{\beta}\right]\right) - \frac{(y_t - \underline{x}_t'\underline{\beta})}{\sigma^4}\right]\underline{x}_t \tag{A5}$$

$$\frac{\partial^2 \ln\Phi}{\partial\sigma^2\partial\sigma^2} = \sum_{t=1}^{T}\left[\frac{(L_t - \underline{x}_t'\underline{\beta})g_t}{4\sigma^4 G_t^2}\left(\frac{G_t}{2\sigma^2}\left[L_t - \underline{x}_t'\underline{\beta}\right]^2 - G_t + g_t\left[L_t - \underline{x}_t'\underline{\beta}\right]\right) - \frac{(y_t - \underline{X}_t'\underline{\beta})^2}{\sigma^6} + \frac{1}{2\sigma^4}\right] .$$

REFERENCES

Amemiya, Takeshi. "Regression Analysis When the Dependent Variable is Truncated Normal." Econometrica, Vol. 41 (1973), pp. 997-1016.

Blum, Julius R. and Judah I. Rosenblatt. Probability and Statistics, W. B. Saunders Co., Philadelphia, 1972.

Cain, Glen G. and Harold W. Watts. Income Maintenance and Labor Supply--Econometric Studies, Institute for Research on Poverty Monograph Series, Markham Press, Chicago, 1973.

Evans, Lewis. "A Linear Approximation to the Conditional Expectation Function Specified in Tobin's Limited Dependent Variable Model." Mimeo, University of Wisconsin, 1975.

Fletcher, R., and M. J. D. Powell. "A Rapidly Convergent Descent Method for Minimization." The Computer Journal, Vol. 6 (1963): 163-168.

Goldberger, Arthur S. "Linear Regression in Truncated Samples." Manuscript, Social Systems Research Institute, University of Wisconsin, 1975.

Gruvaeus, Gunnar T. and Karl G. Jöreskog. "A Computer Program for Minimizing a Function of Several Variables." Educational Testing Service Research Bulletin RB-70-14, Princeton, 1970.

Gurland, John. "An Equality Satisfied by the Expectation of the Reciprocal of a Random Variable." The American Statistician, 21 (April 1967): 24-25.

Hausman, Jerry A. and David A. Wise. "Social Experimentation, Truncated Distributions, and Efficient Estimation." Manuscript, Mathematica, Inc., 1975.

Johnson, Norman L. and Samuel Kotz. Continuous Univariate Distributions, Vol. 1-2, Houghton Mifflin Co., New York, 1970.

Jöreskog, Karl G. "A General Method for Estimating a Linear Structural Equation System," in Arthur S. Goldberger and Otis Dudley Duncan, eds. Structural Equation Models in the Social Sciences, Seminar Press, Inc., New York, 1973.

Survey Research Center. A Panel Study of Income Dynamics. Volumes I-II, Institute for Social Research, University of Michigan, Ann Arbor, 1972.

Tobin, James. "Estimation of Relationships for Limited Dependent Variables." Econometrica, Vol. 26 (1958), pp. 24-36.