observe intensity and observed purchases are zero. Thus, in Tobin's model the sample selection rule is given by (6), and we may write

$$E(Y_{1i}|X_{1i}, Y_{1i} \geq 0) = X_{1i}\beta_1 + E(U_{1i}|Y_{1i} \geq 0).$$

As noted by Cragg (1971) and Nelson (1975), the rule generating the observed data need not be as closely related to the model of equation (1a) as Tobin assumes. Consider the following decision rule: we obtain data on Y_{1i} if another random variable crosses a threshold, i.e., if

$$(7) \qquad Y_{2i} \geq 0$$

while if the opposite inequality holds we do not obtain data on Y_{1i}. The choice of zero as a threshold is an inessential normalization. Also, note that we could define a dummy random variable $d_i = 1$ with the properties

$$(8) \qquad d_i = 1 \quad \textit{iff } Y_{2i} \geq 0, \qquad d_i = 0 \text{ otherwise}$$

and proceed to analyze the joint distribution of Y_{1i} and d_i, dispensing with Y_{2i} altogether. The advantage in using selection rule representation (7) is that it permits a unified summary of the existing literature.

Using this representation, we may write equation (5) as

$$(9) \qquad E(Y_{1i}|X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + E(U_{1i}|U_{2i} \geq -X_{2i}\beta_2).$$

If U_{1i} is independent of U_{2i}, the conditional mean of U_{1i} is zero, and the sample selection process into the incomplete sample is random. In the general case, the conditional mean of the disturbance in the incomplete sample is a function of X_{2i}. Moreover, the effect of such sample selection is that X_2 variables that do not belong in the population regression function appear to be statistically significant in equations fit on selected samples.[3]

A good example of this phenomenon arises in the Gronau (1974)–Lewis (1974) wage selectivity bias problem. In their analyses Y_{1i} is the wage rate which is only observed for working women, and Y_{2i} is an index of labor force attachment (which in the absence of fixed costs of work may be interpreted as the difference between market wages and reservation wages). If the presence of children affects the work decision but does not affect market wages, regression evidence from selected samples of working women that women with children earn lower wages is not necessarily evidence that there is market discrimination against such women or that women with lower market experience—as proxied by children—earn lower wages. Moreover, regression evidence that such extraneous variables "explain" wage rates may be interpreted as evidence that selection bias is present.

For a final example, I draw on my own work (Heckman, 1974). Letting Y_{1i} be the wage rate for woman i, and Y_{2i} be the difference between market wages and reservation wages, a woman works if $Y_{2i} > 0$. Using results from the theory of labor supply, one can show that under certain simplifying assumptions working hours, h_i, are proportional to Y_{2i}. If this proportionality factor is $1/\gamma(>0)$, we are

[3] If the only regressor in X_{2i} is "1", so that the probability of sample inclusion is the same for all observations, only the intercept is biased.

led to the following model:

$$E(Y_{1i}|X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + E(U_{1i}|U_{2i} \geq -X_{2i}\beta_2) \tag{10a}$$

$$d_i = 1 \quad iff \quad Y_{2i} \geq 0, \qquad d_i = 1 \text{ otherwise} \tag{10b}$$

$$E(h_i|X_{2i}, Y_{2i} \geq 0) = E\left(\frac{Y_{2i}}{\gamma}\middle|X_{2i}, Y_{2i} \geq 0\right). \tag{10c}$$

Equations (10a) and (b) are as before. Equation (10c) exploits the information that we observe Y_{2i} up to a positive factor of proportionality if Y_{2i} is positive.

These examples are not intended as a complete literature survey. Yet they illustrate that the basic statistical models for limited dependent variables, censoring and truncation may be summarized in a simple general model for missing data.

Regression estimates of (1a) fit on a selected sample omit the final term on the right hand side of equation (9). Thus the bias that arises from using least squares to fit models for limited dependent variables or models with censoring or truncation arises solely because the conditional mean of U_{1i} is not included as a regressor. The bias that arises from truncation or selection may be interpreted as arising from an ordinary specification error with the conditional mean deleted as an explanatory variable. In general, one cannot sign the direction of bias that arises from omitting this conditional mean.[4]

A crucial distinction is the one between a truncated sample and a censored sample. In a truncated sample one cannot use the available data to estimate the probability that an observation has complete data. In a censored sample, one can.[5] In the next section, I examine a technique that enables one to use this estimated probability to estimate the missing conditional mean for each observation. The estimated conditional mean may be utilized as a regressor in an ordinary regression analysis so that estimators with desirable large sample properties may be derived from computationally simple methods.

II. Simple Estimators for the Case of Joint Normal Disturbances

Suppose that $h(U_{1i}, U_{2i})$, the joint density of U_{1i} and U_{2i}, is bivariate normal. Using well known results in the literature (see, e.g., Johnson and Kotz (1972), pp. 112–113)

$$E(U_{1i}|Y_{2i} > 0) = E(U_{1i}|U_{2i} > -X_{2i}\beta_2) = \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i$$

$$E(U_{2i}|Y_{2i} > 0) = E(U_{2i}|U_{2i} > -X_{2i}\beta_2) = \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i$$

[4] Goldberger (1975) has shown that if the X_{1i} and U_{1i} are normally distributed, regression estimates of Tobin's model are downward biased in absolute value for the true parameters. Clearly in the case of a two variable model, or in a case of orthogonal regressors, one can unambiguously sign the bias if one has *a priori* information about signs of structural coefficients.

[5] In both truncated and censored samples, Y_1 may be a truncated or limited dependent variable.

where

$$\lambda_i = \frac{f(\phi_i)}{1 - F(\phi_i)} \tag{11}$$

$$-\frac{X_{2i}\beta_2}{(\sigma_{22})^{1/2}} = \phi_i$$

and f and F respectively are the density and distribution function of the standard normal distribution.[6] The Tobin model is a special case with $h(U_{1i}, U_{2i})$ a singular density since $U_{1i} \equiv U_{2i}$.

"λ_i" is the inverse of Mill's ratio and is known as the hazard rate in reliability theory. There are several interesting properties of λ_i:

(1) Its denominator is the probability that observation i has data for Y_1.

(2) The lower the probability that an observation has data on Y_1 the greater the value of λ for that observation.

More precisely, using a result due to Feller (1968) and cited in Haberman's proof of the concavity of the probit likelihood function, (Haberman, 1974, p. 309), it is straightforward to show that

$$\frac{\partial \lambda_i}{\partial \phi_i} > 0,$$

and

$$\lim_{\phi_i \to \infty} \lambda_i = \infty \qquad \lim_{\phi_i \to -\infty} \lambda_i = 0.$$

Thus in samples in which the selectivity problem is unimportant (i.e., the sample selection rule ensures that all potential population observations are sampled), λ_i becomes negligibly small so that least squares estimates of the coefficients of (1a) have optimal properties.

Using these results, we may write

$$E(Y_{1i} | X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i \tag{12a}$$

$$E(Y_{2i} | X_{2i}, Y_{2i} \geq 0) = X_{2i}\beta_2 + \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i. \tag{12b}$$

Thus if we know λ_i, or could estimate it, least squares could be applied to estimate the parameters in equation (12a). Similarly, if we could measure Y_{2i} when $Y_{2i} > 0$, as in Tobin's model, knowledge of Y_{2i} and λ_i would permit direct estimation of β_2 and $(\sigma_{22})^{1/2}$ by least squares without having to resort to optimizing likelihood functions.

We may add disturbances to equations (12a) and (12b) to reach the model

$$Y_{1i} = X_{1i}\beta_1 + \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i + V_{1i} \tag{13a}$$

$$Y_{2i} = X_{2i}\beta_2 + \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i + V_{2i} \tag{13b}$$

where $E(V_{1i}) = E(V_{2i}) = 0$.

[6] Note that $f = h_2(U_{2i}/\sigma_{22}^{1/2})$, the normalized density of U_{2i}.

It is straightforward to demonstrate that the covariance structure is given by

(14a) $$E(V_{2i}^2) = \sigma_{22}(1 + \phi_i\lambda_i - \lambda_i^2)$$

(14b) $$E(V_{1i}V_{2i}) = \sigma_{12}(1 + \phi_i\lambda_i - \lambda_i^2)$$

(14c) $$E(V_{1i}^2) = \sigma_{11}((1 - \rho^2) + \rho^2(1 + \phi_i\lambda_i - \lambda_i^2))$$

where

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

and

(15) $$0 \leq 1 + \phi_i\lambda_i - \lambda_i^2 \leq 1.$$

There are several distinctive features of this covariance structure. Clearly, the error structure is heteroscedastic, if X_{2i} (and hence ϕ_i) contains variables apart from "1". Assuming that we know ϕ_i and hence λ_i, regression estimates of the variances of V_{1i} and V_{2i} based on the least squares residuals from equations (13a) and (13b) respectively are downward biased estimates of the true variances of U_{1i} and U_{2i} respectively. This is a consequence of inequality (15). Similarly, the standard estimator of the covariance of disturbances across equations based on the cross product of the least squares residuals from each equation yields an estimator of the population covariance that is biased towards zero.

The heteroscedasticity present in the disturbances of each equation implies that a generalized least squares procedure (GLS) improves the precision of least squares estimates when they are possible. If data are available on Y_1, Y_2 and λ, GLS should be applied to the system of equations (13a) and (13b). Alternative estimators are possible if the information is utilized that the coefficients of λ_i in equations (13a) and (13b) are functions of the population disturbance covariance structure. However, asymptotic optimality for GLS cannot be claimed even if all available information is exploited because the resulting estimators possess a covariance matrix that does not attain the Cramer Rao lower bound.

The approximate GLS estimators possess the advantage of asymptotic normality. The unweighted estimators are also asymptotically normal but the expression for the residual variance is complex, and standard least squares formulae do not apply.[7]

The GLS estimators have an interesting interpretation. Unlikely observations (i.e., those with a low probability of sample inclusion) receive greater weight than likely observations. This is a consequence of the readily confirmed fact that each element of the covariance matrix for V_{1i} and V_{2i} is a monotonic function of

[7] Thus, the estimated residual variance for (14a) converges to $\sigma_{22}\frac{1}{I_1}\sum_{i=1}^{I_1}(1 + \phi_i\lambda_i - \lambda_i^2)$. Of course, the summation term can be estimated so that the standard OLS variance formulae may be modified. But the GLS sampling variances for the parameters are lower and hence preferable.

ϕ_i, and as $\phi_i \to \infty$, the probability of sample inclusion goes to zero, and

$$\lim_{\phi_i \to \infty} E(V_{2i}^2) = 0$$

$$\lim_{\phi_i \to \infty} E(V_{2i} V_{1i}) = 0$$

$$\lim_{\phi_i \to \infty} E(V_{1i}^2) = (1 - \rho^2)\sigma_{11}.$$

The weighting implicit in GLS underscores the crucial nature of the assumption that all observations are drawn from the same population distribution.

As a practical matter, we do not know ϕ_i and λ_i and hence we cannot estimate equations (13a) and (b) unless there is prior information on λ_i. In the case of a censored sample, it is possible to compute the probability that an observation has data missing on Y_1 and hence it is possible to use probit analysis to estimate ϕ_i and λ_i. Thus, denoting d_i as a random variable with the value of one when Y_i is observed, the sample likelihood for the probit analysis is

$$\mathcal{L} = \prod_{i=1}^{I} [F(\phi_i)]^{1-d_i}[1 - F(\phi_i)]d_i.$$

Subject to the standard identification conditions in probit analysis, it is possible to maximize $\mathcal{L}$ to obtain consistent estimates of β_2, ϕ_i and hence λ_i. These estimates of λ_i may be used in place of the true λ_i as regressors in equations (13a) and (b). When regression estimates of the coefficients in equations (13a) and (13b) are possible, they yield consistent estimates of the true parameters since λ_i estimated from probit analysis is a consistent estimator of the true λ_i, and Slutsky's theorem applies. More efficient estimates may be obtained from the approximate GLS estimates which converge in distribution to the true GLS estimates by the Cramer convergence theorem (Cramer, 1946). Other estimates may be obtained from utilizing the information that the coefficients of λ_i are functions of the population covariance structure. Each set of estimates may be used as initial consistent estimates for estimation of the likelihood function. As Rothenberg and Leenders (1964) have shown, one Newton step toward optimizing the likelihood function produces estimates that are asymptotically efficient in the sense that they attain the Cramer–Rao lower bound.[8]

Consideration of three special cases will help to focus ideas. First consider Tobin's model which is presented in equation (13b) in the notation of this section. In Tobin's original model, we observe Y_2 only if it is positive but for all observations in a random sample we know whether or not Y_2 is positive. In the two stage procedure proposed here, first estimate the probit model determining

[8] The likelihood function is straightforward. Using the notation in the text for the case of Y_1 observed when $Y_2 > 0$, Y_1 not observed otherwise, and Y_2 not continuously measured, the likelihood becomes

$$\mathcal{L} = \prod_{i=1}^{I} \left[\frac{\int_{-X_{2i}\beta_2}^{\infty} h(Y_{1i} - X_{1i}\beta_1, U_{2i})\, dU_{2i}}{\int_{-X_{2i}\beta_2}^{\infty} h_2(U_{2i})\, dU_{2i}} \right]^{d_i} \left[\int_{-X_{2i}\beta_2}^{\infty} h_2(U_{2i})\, dU_{2i}\right]^{d_i} \left[\int_{-\infty}^{-X_{2i}\beta_2} h_2(U_{2i})\, dU_{2i}\right]^{1-d_i}.$$

The likelihood function for the other cases is straightforward.

the probability that Y_2 is positive. This gives an estimate of $-(\beta_2 X_{2i}/(\sigma_{22})^{1/2}) = \phi_i$ and hence λ_i for each observation. This estimate of λ_i may be inserted in equation (13b) and least squares estimates of the coefficients in (13b) may be obtained. Note that a weighted version of (13b) can be estimated to eliminate the heteroscedasticity that arises from sample selection.

For at least two reasons this procedure does not utilize all of the available information. The first reason is that the procedure ignores the information that the probit function estimates β_2 up to a factor of proportionality. One could utilize this information to write (13b) as

$$(13b') \qquad Y_{2i} = (\sigma_{22})^{1/2}(-\phi_i + \lambda_i) + V_{2i}.$$

Estimates of (13b') are guaranteed to produce a *positive* estimate of $(\sigma_{22})^{1/2}$, a feature not guaranteed in direct estimation of (13b) with the λ_i estimated from probit as a regressor.[9] One can estimate both weighted and unweighted versions of (13b') with the weights estimated from (14a).

Still, these estimates are not fully efficient. Consider the weighted estimator of (13b'). The residual variance and the regression coefficient each provide an estimator of $(\sigma_{22})^{1/2}$. One can use this information to constrain the squared regression coefficient in the weighted regression to equal the residual variance. Thus one can solve the following quadratic equation for $(\sigma_{22})^{1/2}$,

$$\frac{1}{I_1}\sum_{i=1}^{I_1} \frac{(Y_{2i} - (\sigma_{22})^{1/2}(\lambda_i - \phi_i))^2}{(1 + \phi_i\lambda_i - \lambda_i^2)} = \sigma_{22}$$

where, as before, the first I_1 observations are assumed to have $Y_{2i} > 0$, and estimated values of λ_i and ϕ_i are used in place of actual values. The left hand side of the equation is the error sum of squares from the weighted regression. This estimate of $(\sigma_{22})^{1/2}$ is consistent and is guaranteed to produce a positive estimate of $(\sigma_{22})^{1/2}$ if the quadratic equation possesses a real root, but is not necessarily more efficient than the previous estimator.[10]

[9] Note that the estimated $\lambda_i - \phi_i$ and the actual Y_{2i} are positive numbers. Hence the least squares estimate of $(\sigma_{22})^{1/2}$ is positive.

[10] The equation for $(\sigma_{22})^{1/2}$ is given by

$$(\sigma_{22})^{1/2} = \frac{-\frac{1}{I_1}\sum \frac{(Y_{2i})(\lambda_i - \phi_i)}{1+\phi_i\lambda_i - \lambda_i^2} + \sqrt{\left[\frac{1}{I_1}\sum \frac{(Y_{2i})(\lambda_i - \phi_i)}{1+\phi_i\lambda_i - \lambda_i^2}\right]^2 + \left(\frac{1}{I_1}\sum \frac{Y_{2i}^2}{1+\phi_i\lambda_i - \lambda_i^2}\right)\left(1 - \frac{1}{I_1}\sum \frac{(\lambda_i - \phi_i)^2}{1+\phi_i\lambda_i - \lambda_i^2}\right)}}{1 - \frac{1}{I_1}\sum \frac{(\lambda_i - \phi_i)^2}{(1+\phi_i\lambda_i - \lambda_i^2)}}$$

for $1 - \frac{1}{I_1}\sum \frac{(\lambda_i - \phi_i)^2}{1+\phi_i\lambda_i - \lambda_i^2} = d \neq 0.$

When the last condition does not hold, it is straightforward to develop the appropriate expression for $(\sigma_{22})^{1/2}$. In either case consistency is readily verified. Nothing guarantees that d is positive. For example, if all observations have a probability of sample inclusion that exceeds 85 percent, $d < 0$, and no real root need exist in a small sample although in a large sample, one must exist. It is interesting to note that nonexistence is most likely in samples with observations for which the probability of sample inclusion is high, i.e., precisely in those circumstances when least squares is an appropriate estimator, the range of variation in λ_i is small, and we would place little weight on the regression estimate of $(\sigma_{22})^{1/2}$.

None of these two step estimators of $(\sigma_{22})^{1/2}$ attains the Cramer-Rao lower bound so that use of the Rothenberg one step estimator is recommended when possible. An advantage of the multiplicity of estimators for β_2 and $(\sigma_{22})^{1/2}$ is that they allow a check on the appropriateness of the model. For example, if the probability of the event $Y_{2i} > 0$ is not as closely linked to the equation for Y_{2i} as Tobin assumed, the β_2 estimated from (13b) will not be proportional to the $\beta_2/(\sigma_{22})^{1/2}$ estimated from probit analysis.

Finally, note that unconstrained estimates of equation (13b) are likely to be imprecise because λ and its estimate are nonlinear functions of the X_2 regressors that appear in that equation. Since ϕ and λ are positively correlated (often strongly so) multicollinearity may be a problem and for that reason constrained estimators can produce more reasonable results.

The procedure for more general models is similar to that outlined for Tobin's model. In our second example, suppose that we observe Y_1 only when $Y_2 > 0$, that we do not observe actual values of Y_2, but we know whether or not $Y_2 > 0$ for all observations from a random sample. This is the model of Gronau and Lewis.

As before, we may estimate ϕ and λ from probit analysis. The estimated λ is then used as a regressor in equation (13a). Regression estimates of the parameters are consistent estimators. To estimate the approximate generalized least squares version of (13a), we may use the residuals from this regression to estimate the weights given in equation (14c).[11]

An alternative procedure uses the information from (14c) in conjunction with (13a) to simultaneously estimate β_1, ρ, and σ_{11}. From the definition of ρ given below equation (14c) note that equation (13a) may be written as

(13a′) $$Y_{1i} = X_{1i}\beta_1 + \rho(\sigma_{11})^{1/2}\lambda_i + V_{1i}.$$

The weighted estimator that utilizes the information that the coefficient of λ_i is a parameter of the population variance, chooses β_1, ρ, and σ_{11} to minimize

$$\sigma_{11} = \frac{1}{I_1}\sum \frac{(Y_{1i} - X_{1i}\beta_1 - \rho(\sigma_{11})^{1/2}\lambda_i)^2}{1 + \rho^2(\phi_i\lambda_i - \lambda_i^2)}$$

with the λ_i estimated from probit analysis used in place of the true λ_i. As before, we cannot be sure that in small samples this estimator exists although in large samples it must exist. Asymptotically, this procedure yields estimators that are consistent but are inefficient compared to maximum likelihood estimators.

As a final example that is the topic of the empirical work reported below, consider the model of equations (10a)–(10c) with normality assumed for U_{1i} and U_{2i}, and censored sampling assumed. Y_2 is observed up to an unknown factor of proportionality when $Y_2 > 0$, and Y_1 is observed only when $Y_2 > 0$. This example combines aspects of the two previous examples.

[11] Simply regress each squared residual from the unweighted regression V_{1i}^2 on "$\phi_i\lambda_i - \lambda_i^2$" and an intercept. The intercept estimates σ_{11} while the slope estimates $\rho^2\sigma_{11}$. Under general conditions, these estimates are consistent for the true parameters and permit estimation of the weights required in the weighted regression. Nothing in the procedure ensures that the estimated variance is positive or that the estimate of ρ^2 lies in the unit interval.

For specificity, let Y_{1i} be the market wage that woman i could earn were she to work. Y_{2i} is the difference between market wages Y_{1i} and reservation wages Y_{3i}. Hours of work are proportional to the difference between market wages and reservation wages when this difference is positive with the factor of proportionality denoted by $1/\gamma$. This proportionality factor must be positive if the model is to accord with economic theory.

In this model, the parameters of the functions determining Y_{3i} and Y_{1i} are of direct interest. The equation for reservation wages Y_{3i} may be written as

(15a) $$Y_{3i} = X_{3i}\beta_3 + U_{3i}$$

(15b) $$Y_{1i} = X_{1i}\beta_1 + U_{1i}$$

hence

(15c) $$Y_{2i} = X_{1i}\beta_1 - X_{3i}\beta_3 + U_{1i} - U_{3i} = X_{2i}\beta_2 + U_{2i}$$

where

$$X_{2i} = (X_{1i}X_{3i}) \quad \text{and} \quad U_{2i} = U_{1i} - U_{3i}.$$

The hours of work equation is given by

(15d) $$h_i = \frac{1}{\gamma}(Y_{1i} - X_{3i}\beta_3) - \frac{1}{\gamma}U_{3i}$$

or, in reduced form,

(15e) $$h_i = \frac{1}{\gamma}(X_{1i}\beta_1 - X_{3i}\beta_3) + \frac{1}{\gamma}(U_{1i} - U_{3i}).$$

U_{3i} and U_{1i} are assumed to be joint normal variates with zero mean and the covariance structure is unconstrained.

Note that

$$E(U_{1i}^2) = \sigma_{11}, \qquad E(U_{1i}U_{2i}) = \sigma_{12} = \frac{1}{\gamma}(\sigma_{11} - \sigma_{13})$$

$$E(U_{2i}^2) = \frac{1}{\gamma^2}(\sigma_{11} - 2\sigma_{13} + \sigma_{33})$$

where

$$E(U_{ji}U_{3i}) = \sigma_{j3} \qquad j = 1, 2, 3.$$

In this notation,

(15b') $$E(Y_{1i}|h_i > 0) = X_{1i}\beta_1 + \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i$$

(15e') $$E(h_i|h_i > 0) = \frac{1}{\gamma}(X_{1i}\beta_1 - X_{3i}\beta_3) + \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i$$

or, equivalently,

$$(15d') \qquad E(h_i|h_i>0)=\frac{1}{\gamma}(E(Y_{1i}|h_i>0)-X_{3i}\beta_3)-\frac{1}{\gamma}\frac{(\sigma_{33}-\sigma_{13})\lambda_i}{(\sigma_{22})^{1/2}}.$$

The two stage procedure applicable to this model is (1) to estimate the probit function determining whether or not a woman works. This yields an estimate of λ_i. (2) Use the estimated λ_i as a regressor in (15b'), and (15e'). Alternatively, the hours of work equation may be estimated from (15d') using the predicted value from wage function given by (15b') as a regressor. The advantage of this procedure is that it permits estimation of a unique value of $1/\gamma$ whereas if the model is overidentified, equation (15e') leads to a multiplicity of estimates for γ. Note that the usual rank and order restrictions apply for identifiability of γ. For example, if X_{1i} contains one variable not contained in X_{3i}, and the rank condition applies, γ and hence the vector β_3 are estimable parameters.

As in the other cases, approximate GLS estimators may be developed. The procedure for developing such estimators follows exactly along the lines discussed in the simpler models, and so will not be elaborated here.

The analysis for truncated samples is identical to the previous analysis for censored samples provided that an estimate of λ is available. This estimate may come from other data sets or from subjective notions. Clearly the quality of the resulting estimator depends on the quality of the estimate of λ. Amemiya (1974) has proposed an initial consistent estimator for the Tobin model that is applicable to the case of truncated samples. Moreover, a straightforward extension of his estimator leads to initial consistent estimators for the Gronau–Lewis model, and the expanded model just discussed. The advantage of Amemiya's estimator is that it is based on sample evidence. While Amemiya's estimator is more cumbersome to apply, it is clearly an alternative to the one proposed here, and has the advantage that it can be used in truncated samples.

III. Empirical Performance of the Estimator

In this section, I report the results of an empirical analysis of the joint model of labor force participation, wages and hours of work presented in the last section. Elsewhere (Heckman, 1976) I present a more extensive empirical analysis of this model and demonstrate that the proportionality assumption of equations (10c) and (15d) and (15e) may be inappropriate because of worker and employer fixed costs. Here, I assume this model is correct and report the results of using the computationally simple estimator to estimate the parameters of the sample likelihood function.[12] As we shall see, the initial consistent estimator proposed here locates the optimum rather precisely. This exercise is of more than methodological interest. In an earlier paper (Heckman, 1974) I estimated this model using data on female labor force experience that was erroneously coded by the primary data source. Thus the analysis here permits an examination of the effect of this coding error on the estimates presented in the earlier paper.

[12] A derivation of the sample likelihood function for this model is provided in Heckman (1974).

The data source is the National Longitudinal Survey tape of women 30–44 interviewed in 1967. From an original sample of 5,083 women, a working sample of 2,253 white, married spouse present women with usable data was constructed. The reasons for sample exclusion are given in Appendix A-1. Sample means for the data used in the empirical analysis are reported in Appendix A-2. The measure of labor supply used in this paper is annual hours worked defined as the product of weeks worked with average hours per week. A woman works if she has nonzero hours of work in 1967.

Estimates for the probit model predicting the probability that a woman works are given in Column 1 of Table 1. The variables used in the analysis are self explanatory. These estimates are to be compared with the estimates of this probability derived from the sample likelihood function. The agreement is rather close.

TABLE 1
PROBIT ESTIMATES

	Original Probit Estimates of $\beta_2/(\sigma_{22})^{1/2}$ ("t" stat. in parentheses)		Estimates from the Likelihood Function
Number of children less than 6	-0.44968	(-9.987)	-0.4410
Assets	-0.6880×10^{-5}	(-3.01)	-0.7157×10^{-5}
Husband's hourly wage rate	-0.01689	(-1.16)	-0.0366
Wife's Labor market experience*	0.07947	(16.67)	0.0774
Wife's education	0.0302	(2.306)	0.0406
Constant	-1.1553	(7.569)	-1.1331

Log likelihood: -1186.8

Probability that a woman works is: $\int_{-\beta_2 X_{2i}/(\sigma_{22})^{1/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$

701 women work of a sample of 2,253 women.

* The number of years the woman worked full time since marriage.

Next, I report estimates of the parameters of equations (15b) and (15c) obtained from regressions with and without "λ" variables to correct for sample censoring bias. The natural logarithm of the hourly wage rate is used for Y_{1i}. Annual hours worked are used as the measure of labor supply, h_i.

The top portion of Table 2 records the empirical results for the coefficients of the hourly wage function. Column 1 records the results of estimating the wage function by least squares on the subsample of working women without correcting for censoring bias. Column 2 records the result of estimating the same function entering the estimated λ_i as a regressor. Note that in this sample we cannot reject the null hypothesis that sample censoring for wage functions is an unimportant phenomenon. This result stands in marked contrast to the empirical results in Gronau (1974) who found significant selectivity bias.

Generalized least squares estimates of the wage equation are given in Column 3.[13] The weights for the GLS estimator were derived from regressions on

[13] These estimates only use the information available on the wage equation, and do not exploit the interequation covariance structure.

TABLE 2

LEAST SQUARES SYSTEM FOR ANNUAL HOURS WORKED
("t" STATISTICS IN PARENTHESES)

Variable	(1) Traditional Regression	(2) Traditional Regression Corrected for Censoring	(3) Single Equation Generalized Least Squares Estimates	(4) Maximum Likelihood Estimates	
		ln Hourly Wage Rate			
Wife's labor market exp.	0.01509(8.24)	0.029(5.2)	0.0203(5.2)	0.020(8.23)	
Wife's education	0.0678(13.3)	0.0687(13.0)	0.0686(13)	0.0679(13.3)	
λ_i	—	0.0992(1.52)	0.1002(1.52)	—	
Intercept	−0.2718	−0.437(1.45)	−0.435(3.4)	−0.412(5.3)	
R^2	0.254	0.257	0.83		
		Annual Hours Worked			
	(1) Traditional Regression	(2) Regression Corrected for Censoring	(3) Implied Marginal Effect from Estimates of (2) Evaluated at Sample Mean	(4) Single Equation GLS Estimates for Heckman Model	(5) Maximum Likelihood Estimates
Nbr. of children less than 6	−140.78(−2.92)	5 11.18(2.3)	−96.81	608(2.5)	−290
Assets	$-6.61 \times 10^{-4}(-2.81)$	$8.463 \times 10^{-3}(2.25)$	-8.39×10^{-4}	$1.08 \times 10^{-4}(2.6)$	-4.71×10^{-4}
Husband's weekly wage rate	−54.34(−3.64)	−32.31(−195)	−55.14	−17.96(−2.94)	−240.78
Wife's labor market exp.	35.55(8.76)	−65.077(−1.98)	42.37	−79(−2)	509.2
Wife's education	37.66(3.12)	−6.794(0.3)	34.17	−14.6(0.7)	267.1
λ_i	—	−2020.7(3)	—	−2311(3.2)	
Intercept	899.45	4009.5		2769.2(4.33)	7.46×10^3
R^2	0.1354	0.1444		0.7862	

the first stage residuals using the procedure reported in footnote 11 in the previous section. Note that the GLS equation is very similar to the least squares equation reported in Column 2. The R^2 shown in Column 3 is the R^2 for the regression using weighted variables.

Finally, estimates of the wage function obtained from optimizing the likelihood function are reported in Column 4. Note that the GLS estimates closely approximate the maximum likelihood estimates.

The results in the bottom row of Table 2 are less reassuring. Column 1 records estimates of reduced form equation (15e) that ignore the possibility of selection bias.[14] Column 2 records estimates of this equation with λ_i included as a regressor. Note that the coefficient on λ_i is statistically significant and negative, and that this result remains in the GLS estimates, so that there is considerable evidence that there is pronounced selection bias in estimating hours of work functions on sub-samples of working women.

The estimated negative coefficient on λ_i is disturbing since if the model of equations (10a)–(10c) is true, this coefficient yields an estimate of the standard deviation $(\sigma_{22})^{1/2}/\gamma$ (see equation (15e′)) and should be positive. Further, the negative coefficient on the wife's labor market experience, taken in conjunction with the positive effect of experience on her wages, implies that the estimated value of γ is negative, contrary to the premises of the model.[15] Finally, inspection of Column 5 shows that the maximum likelihood estimates do not correspond to the estimates of Column 2 or Column 4.

These empirical results have led me to develop a more general model in which the hours of work equation is not as closely related to the participation equation as it is postulated to be above. Such a model arises naturally when there are fixed time and money costs of work and child care, and is reported in another paper. (Heckman, 1976).

An alternative estimator of the model under discussion that avoids an embarrassing confrontation with data is obtained by noting that if the model were true, probit coefficients for the work-no work decision would estimate the coefficients of β_2 in the hours of work equation (15c) up to a factor of proportionality. In the notation of this section, the factor of proportionality is simply given by $\sigma_{22}^{1/2}/\gamma$. In an obvious way, we may adapt the estimator for the Tobin model proposed in equation (13b′) and write

$$(16) \qquad h_i = \frac{\sigma_{22}^{1/2}}{\gamma}(-\phi_i + \lambda_i) + \frac{V_{2i}}{\gamma}, \qquad i = 1, \ldots, I_1,$$

where

$$\phi_i = -\frac{\beta_2 X_{2i}}{(\sigma_{22}^{1/2}/\gamma)}.$$

[14] The labor supply equation is just identified because the only variable that appears in the wage function that does not appear in the reservation wage equation is labor market experience. Hence the choice between estimating equations (15d′) and (15e′) is immaterial.

[15] The estimate of $1/\gamma$ is obtained by dividing the coefficient for experience in the female wage equation (0.0203 in the GLS estimates) into the coefficient for experience in the hours of work equation (−79). The resulting estimate is −3891.6.

Probit analysis yields estimates of ϕ_i and λ_i. Hence we may estimate $(\sigma_{22})^{1/2}/\gamma$ by regressing h_i on $(-\phi_i+\lambda_i)$. This estimate is guaranteed to be positive.[16] Thus, we can estimate equation (15c) and hence we can estimate the effect of experience on hours of work. Using the coefficient on the experience variable from the wage

TABLE 3
MAXIMUM LIKELIHOOD ESTIMATES AND INITIAL CONSISTENT ESTIMATES OF THE HECKMAN (1974) MODEL

Annual Hours
("t" statistics in parentheses)

	Likelihood Optimum	Estimates in Original Paper	Initial Consistent Estimates	First Step Iterate
Natural Logarithm of Market Wage Equation Y_{1i} (Coefficients of β_1)				
Intercept	−0.412 (5.28)	−0.982 (8.93)	−0.435	−0.593 (8.70)
Education	0.0679 (13.58)	0.0761 (10.15)	0.0686	0.0688 (17.20)
Experience	0.0200 (10.00)	0.048 (12.00)	0.0205	0.025 (1.14)
Natural Logarithm of Reservation Wage Y_{3i} (Coefficients of β_3)				
Intercept	−0.1191 (1.77)	−0.623 (32.28)	−0.103	−0.0964 (2.10)
Effect of hours on reservation wage (γ)	0.152×10^{-3} (7.96)	0.63×10^{-3} (12.60)	0.9×10^{-3}	0.148×10^{-3} (1383.18)
Husband's wage	0.00946 (2.49)	0.051 (7.29)	0.00418	0.0238 (4.76)
Wife's education	0.0574 (10.14)	0.0534 (7.63)	0.061	0.0548 (13.70)
Assets	0.185×10^{-5} (3.14)	0.135×10^{-5} (2.45)	0.1702×10^{-6}	0.285×10^{-5} (0.41)
Nbr. children 6	0.114 (6.48)	0.179 (52.63)	0.115	0.116 (7.25)
Std. Deviation in Mkt. Wage Equation $\sigma_{11}^{1/2}$	0.329 (32.90)	0.452 (37.36)	0.320	0.253 (23.00)
Std. Deviation in Reservation Wage Eq. $\sigma_{33}^{1/2}$	0.363 (24.20)	0.532 (28.00)	0.351	0.259 (26.16)
Interequation Correl. $\frac{\sigma_{13}}{\sqrt{\sigma_{11}\sigma_{33}}}$	0.725 (11.69)	0.6541 (14.22)	0.353×10^{-4}	0.317 (4.23)
Log Likelihood	−5,778	—	−6,414	−6,102
Log Likelihood under null hypothesis of no selection bias	−5,783			

[16] Either weighted or unweighted estimators may be used, and as discussed in Section II, a more efficient estimator exploits the information that the regression coefficient is the square root of the population variance.

equation divided into the experience coefficient for the hours of work equation, we may estimate $1/\gamma$, and hence $(\sigma_{22})^{1/2}$. For these data these estimates are positive.

Unweighted estimates of equation (16) are used to develop the initial consistent estimate of the natural logarithm of the reservation wage function that are displayed in Column 3 of Table 3. The estimates of the population wage function are taken from the estimates reported in the second column of the first row of Table 2.[17]

The initial consistent estimates displayed in Table 3 are to be compared with the coefficients displayed in Column 1 obtained from optimizing the likelihood function. For most coefficients, the agreement between the two estimates is rather close. The only exceptions come in the estimate of γ and in the estimate of the intercorrelation between the disturbances of the market wage equation (U_{1i}) and the reservation wage equation (U_{3i}). Note that a comparison of Columns 1 and 2 suggests that the coding error that appeared in the original Parnes tapes introduced considerable error in the estimated coefficients. In particular, the effect of experience on wages was overstated in my previous paper while the effect of wages on labor supply ($1/\gamma$) was understated. Finally, note that the first step iterate of the initial consistent estimator, an asymptotically efficient estimator, is numerically close to the maximum likelihood estimator but for most coefficients is not as close as the initial consistent estimator.

Summary and Conclusions

This paper discusses the bias that arises from sample selection, truncation and limited dependent variables within the familiar specification error framework of Griliches and Theil. A simple estimator for censored samples, due to Gronau and Lewis, is discussed and applied to reestimate a model of female labor supply, wages and labor force participation. The estimates compare quite closely to the estimates obtained from maximum likelihood.

The estimator discussed here is viewed as a complement to Amemiya's estimators (1973, 1974) for related models. No comparison of relative efficiency has been performed. Neither estimator is efficient compared to maximum likelihood, but both are computationally more flexible than maximum likelihood and for that reason both are more useful in exploratory empirical work.[18]

University of Chicago

[17] Estimates of the covariance structure are obtained from the inter-equation residual correlation between the residuals from equation (16) and the wage function (15b'). Note that the estimate of $\sigma_{22}^{1/2}/\gamma$ taken from the regression coefficient of equation (16) is 583.01.

[18] An example of the potential in cost saving may be useful. It cost \$700 to produce estimates of the likelihood function reported in Table 3 and \$15 to produce the initial consistent estimates and the GLS estimates.

Appendix A–1

Sample Selection Criteria

In the original National Longitudinal Survey tape of women 30–44 interviewed in 1967, 5,083 observations were available. The following rejection criteria were employed to reach the working sample of 2,253 total women, 701 of whom are working in the survey year. For a description of the data source, see Shea, *et al.*, 1970

(1) Nonwhite	(1,552)
(2) Nonmarried spouse present	(1,971)
(3) Husband has no income	(194)
(4) Wife has a job, but not working in survey week	(107)
(5) Wife's work experience not available	(357)
(6) Education of wife not available	(7)
(7) Unknown wage rate for working woman	(177)

Note that observations may be rejected for any of the seven reasons listed. Assets were assigned in 176 cases from the equation fit on the subsample of working women.

$$\text{Assets (1967)} = -9{,}205 + 171.80\ (\text{husband's wage rate}) - 53.29\ (\text{wife's experience}) + 2{,}034\ (\text{wife's education})$$

Appendix A–2

DESCRIPTION OF DATA

(1967 National Longitudinal Survey of Women 30–44)
701 working women; 2,253 women

	Mean for Working Women	Mean for all Women
Nbr. of children less than 6	0.252	0.5512
Assets ($)	12,466	13,963
Husband's income ($)	6,531	6.924
Wife's annual hours	1,527	—
Weekly wage ($/wk.)	75.92	—
Weeks worked	41.2	—
Labor force experience (years)	11.5	7.75
Wife's education	11.3	11.33
Husband's wage rate	3.02	3.16
Log of wife's weekly wage	4.12	—
Hours per week of wife	36.3	—
Selection factor (λ_i)	1.033	1.585
Participation rate	1.0	0.373

References

Amemiya, T., "Multivariate Regression and Simultaneous Equation Models When the Dependent Variables are Truncated Normal," *Econometrica* 42 No. 6, November 1974: 999–1011.

———, "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica* 41, No. 6: 997–1017.

Cragg, J., "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica* 39, No. 5, September, 1971: 829–844.

Cramer, H., *Mathematical Methods in Statistics*. Princeton, 1946.

Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 2. Wiley, New York, 1971.

Goldberger, A., "Linear Regression in Truncated Samples," Social Systems Research Institute, University of Wisconsin-Madison, May 23, 1975.

Gronau, Reuben, "The Effect of Children on the Housewife's Value of Time," *Journal of Political Economy*, March/April, 1973.

———, "Wage Comparisons—A Selectivity Bias," *Journal of Political Economy*, November/December, 1974.

Haberman, Shelby, *The Analysis of Frequency Data*. University of Chicago Press, 1974.

Heckman, James J., "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, July 1974.

———, "Sample Selection Bias as a Specification Error," Unpublished Manuscript, Rand Corporation, April 1976.

Johnson, N. and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York, 1972.

Lewis, H. Gregg, "Comments on Selectivity Biases in Wage Comparisons," *Journal of Political Economy*, November/December 1974.

———, *Unionism and Relative Wages*. University of Chicago Press, 1963.

Nelson, Forrest, "Censored Regression Models with Unobserved Stochastic Censoring Thresholds," National Bureau of Economic Research Working Paper No. 63, 1975.

Rothenberg, T. and C. Leenders, "Efficient Estimation of Simultaneous Equation Systems," *Econometrica* 32, No. 1, January 1964: 57–76.

Shea, *et al.*, *Dual Careers: A Longitudinal Study of Labor Market Experience*. Columbus, Ohio: Center for Human Resource Research, Ohio State University, May 1970.

Tobin, J., "Estimation of Relationships for Limited Dependent Variables," *Econometrica* 26, 1958: 24–36.