ANNALES DE L'INSEE – N° 30-31/1978

Missing data and self-selection in large panels

Zvi GRILICHES
Bronwyn H. HALL
Jerry A. HAUSMAN *

Two problems which occur in analyzing large panels of cross section data are considered : missing data and self-selection. In the case of randomly missing data using only the complete data subsample results in unbiased but inefficient estimates. We demonstrate that in large panels the efficiency gains from using efficient methods are likely to be quite small. For non-random missing data we present a methodology which corrects for the bias which occurs if only the complete data subsample is used. Lastly, we formulate and estimate a model where the missing data arises from self-selection in the decision to remain in school. Using the National Survey of Young Men, we find that accounting for self-selection increases the estimated returns to schooling by 50 %.

* Zvi Griliches : Professor of Economics, Harvard University; Bronwyn H. Hall : Senior Programmer in the Department of Economics, Harvard University; Jerry A. Hausman : Associate Professor of Economics Massachusetts Institute of Technology. We are indebted to the NSF and NIE for financial support and to Gary Chamberlain for very valuable comments.

The discussion of missing data has a long history in statistics and a somewhat more limited history in econometrics. Most of it consists of the discussion of the randomly missing case with suggestions for ad hoc or more elaborate maximum likelihood computational "solutions."[1] There are also some scattered Monte Carlo results suggesting that the performance of most such methods is relatively poor (relative to the alternative of concentrating on the smaller "complete" data subsamples).[2]

The purpose of this paper is somewhat different. First we point out that in the "randomly" missing data model, the only gain from "solving" the problem is the increased efficiency of the resulting parameter estimates. We explore briefly the source and possible size of such efficiency gains and conclude that in large samples, such as the currently available microsurveys and panels, the game may not be worth the candles.

The major problem in econometrics is not just missing data but the possibility (or more accurately, probability) that they are missing for a variety of self-selection reasons. Such "behavioral missing" implies not only a loss of efficiency but also the possibility of serious bias in the estimated coefficients of models that do not take this into account. The recent revival of interest in econometrics in limited dependent variable models (Tobit), sample-selection, and sample self-selection problems has provided both the theory and computational techniques for attacking this problem.[3]

The substantive examples in this paper are taken from our research on the economic returns to schooling using the National Longitudinal Survey of Young Men.[4] Three problems will be addressed: (1) The "filling-in" of missing data (IQ data were missing for about one-third of the sample);
(2) Changing sample size over time, due both to sample attrition due to the inability to find respondents and sample accretion due to the dropping into the working-with-wages sample of those youths who drop out from or finish their schooling;
(3) The self-selection "to work" of the out-of-school subsample, the subsample for which an earnings function could be computed.

Consider the simple model:

$$y = \beta_1 x_1 + \beta_2 x_2 + e$$

which satisfies the usual OLS assumptions and where we have suppressed the constant for notational ease. For some fraction of our sample we are missing observations on x_1. Let us rearrange the data and call the complete sample A and the incomplete sample B. Now, given the assumption that the x_1 values are missing at random, i.e., that the conditional expectation of x_1, $E(x_1|\text{given that it is observed}) = Ex_1$, equals its unconditional expectation, and in particular $E(e|x_1 \text{ observed}) = Ee = 0$, the equation above can be estimated consistently solely from sample A. Obviously, however, there is some more information in sample B. The following questions then arise:

1. How much additional information is there in sample B and about which parameters?

2. How should the missing values of x_1 be estimated?

Options include using only x_2, using x_2 and y, or using x_2 and z, where z is an additional instrumental variable, related to x_2 but not appearing itself in the y equation.

To discuss this, it is helpful to specify an "auxiliary" equation for x_1:

$$x_1 = \delta x_2 + \eta z + v$$

where $E(ve) = 0$. Note that as far as this equation is concerned, the missing data problem is one of missing the dependent variable for sample B. If the probability of being present in the sample were related to the size of v, we would be in the non-random missing case. We shall ignore this possibility for now. It will be taken up in the next section. We also limit ourselves at first to the simplest case, one with no additional instrumental variables present ($\eta = 0$).

We can then rewrite our model as

$$y_a = \beta_1 x_{1a} + \beta_2 x_{2a} + e_a$$

$$x_{1a} = \delta x_{2a} + v_a$$

$$y_b = (\beta_2 + \beta_1 \delta) x_{2b} + e_a + \beta_1 v_b$$

where the a and b subscripts refer to samples A and B, respectively.

Sample A yields estimates of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\delta}$, with variance-covariance matrix Σ_a. Sample B yields an estimate of $\hat{\pi} = (\beta_2 + \beta_1\delta)$ with variance σ_π^2. A maximum likelihood solution would blend the two independent pieces of information optimally, imposing the non-linear constraint $\hat{\pi} = \hat{\beta}_2 + \hat{\beta}_1\hat{\delta}$.

A "first order" procedure, i.e., one that estimates missing x_1's by x_2 alone and does not iterate, would be equivalent to the following: Estimate $\hat{\beta}_{1a}$, $\hat{\beta}_{2a}$, $\hat{\delta}_a$ from sample A, rewrite the y equation as

$$\begin{pmatrix} y_a - \hat{\beta}_{1a}x_{1a} \\ \\ y_b - \hat{\beta}_{1a}\hat{\delta}_a x_{2b} \end{pmatrix} = \beta_2 x_2 + \begin{pmatrix} e_a \\ \\ e_b + \beta_1 v \end{pmatrix} + \varepsilon$$

where ε involves terms which are due to the discrepancy between the estimated $\hat{\beta}_{1a}$ and $\hat{\delta}_a$ and their true population values. Then just estimate β_2 from this "completed" sample by OLS.

It is clear, then, that this procedure results in no gain in the efficiency of β_1, it is just based on sample A.[5] It is also clear that the resulting estimate of β_2 could be improved somewhat using GLS instead of OLS.

How much of a gain is there in estimating β_2 this way? Let the size of sample A be N_1 and of B be N_2. The maximum (unattainable) gain in efficiency would be proportional to $(N_1 + N_2)/N_1$ (when $\sigma_v^2 = 0$). Ignoring the contribution of ε's, which is unimportant in large samples, the variance of $\hat{\beta}_2$ from the sample as a whole would be

$$\mathrm{Var}(\hat{\beta}_{2a+b}) \approx [N_1\sigma^2 + N_2(\sigma^2 + \beta_1^2\sigma_v^2)]/(N_1 + N_2)^2\sigma_{x_2}^2$$

and

$$\mathrm{Eff}(\hat{\beta}_{2a+b}) = \frac{\mathrm{Var}\ \beta_{2a+b}}{\mathrm{Var}(\hat{\beta}_{2a})} \approx \frac{N_1}{N_1 + N_2}\left(1 + \frac{N_2}{N_1 + N_2}\,\frac{\beta_1^2\sigma_v^2}{\sigma^2}\right)$$

where $\sigma^2 = \sigma_e^2$.[6]

Let us look now at a few illustrative calculations. In the sample to be discussed below, y will be the logarithm of the wage rate, x_1 - IQ, and x_2 - schooling. IQ scores are missing for one-third of the sample, hence $N_1/(N_1 + N_2) = 2/3$ and wage rates is relatively small. Its independent contribution $(\beta_1^2\sigma_v^2)$ is small relative to the large unexplained variance in y. Typical numbers are $\beta_1 = .005$, $\sigma_v = 12$, and $\sigma = .4$, implying

$$\text{Eff}(\hat{\beta}_{2a+b}) \approx 2/3\ [1 + \frac{1}{3}\,\frac{.0036}{.16}] = .672,$$

which is about equal to the 2/3's one would have gotten ignoring the term in the brackets. Is this a big gain in efficiency? First, the efficiency (squared) metric may be wrong. A more relevant question is by how much can the standard error of $\hat{\beta}_2$ be reduced by incorporating sample B into the analysis. By about 18% ($\sqrt{.672} = .82$) for these numbers. Is this much? That depends how large the standard error of β_2 was to start out with. For 1973 our "good IQ" subsample consists of 1,540 individuals yielding an estimate of $\hat{\beta}_{2a} = .0641$ with a standard error of .0052. Processing another 700 plus observations we could reduce this standard error to .0043, an impressive but rather pointless exercise, since nothing of substance depends on knowing β_2 within .001.

If IQ (or some other missing variable) were more important, the gain would be even smaller. For example, if the independent contribution of x_1 to y were on the order of σ^2, then with 1/3rd missing, $\text{Eff}(\hat{\beta}_{2a+b}) \approx 8/9$, and the standard deviation of β_2 would be reduced by only 5.7%. There would be no gain at all, if the missing variable was 1½ times as important as the disturbance (or more generally if $\beta_1^2\sigma_v^2/\sigma^2 > (N_1 + N_2)/N_1$).

What do we gain if we have an additional instrumental z variable? As far as β_2 is concerned, this would be reflected in a reduction in σ_v^2, moving the gain closer to $(N_1 + N_2)/N_1$. But now there is also a possibility of some gain in the efficiency of β_1.

Assume that we have good estimates of β_2 and δ from sample A. Then we can rewrite the problem as

$$\begin{pmatrix} y_a - \hat{\beta}_{2a}x_{2a} \\ y_b - \hat{\beta}_{2a}x_{2a} \end{pmatrix} = \begin{pmatrix} x_{1a} \\ \hat{x}_{1b} \end{pmatrix} \beta_1 + \begin{pmatrix} e_a \\ e_a + \beta_1 v \end{pmatrix}$$

where $\hat{x}_{1b} = \hat{\delta}_a x_{2b} + \hat{\eta}_a z_b$ is the "predicted" value of x_1 for sample B using the coefficients of the auxiliary equation (2.2) estimated from the complete A sample.[7] Again, a large sample estimate of the variance of such a combined estimate of β_1 would be

$$\text{Var }[\hat{\beta}_{1(a+bi)}] \simeq \frac{1}{N\sigma^2_{x_1}} \left\{(1 - \lambda)\sigma^2 + \lambda(\sigma^2 + \beta_1^2\sigma_v^2)/R^2\right\}$$

where $N = N_1 + N_2$, $\lambda = N_2/N$, $R^2 = R^2_{x_1 \cdot x_2, z}$, and bi stands for the "instrumented" B sample. The large sample efficiency of OLS in the "complete" A subsample relative to this combined instrumental variable estimator is then

$$\text{Eff}(\hat{\beta}_{1(a+bi)}) \simeq (1 - \lambda)\ [(1 - \lambda) + \frac{\lambda}{R^2}\ (1 + \frac{\beta_1^2\sigma_v^2}{\sigma^2})]$$

In our cross-sectional example $\beta_1^2\sigma_v^2/\sigma^2$ is rather small, 0.02, and hence the whole expression is equal approximately to $(1 - \lambda)^2 + \lambda(1 - \lambda)/R^2$. In cross-sectional micro samples one is unlikely to find a collection of instruments for which $R^2 > .5$. Given $\lambda = 1/3$ and an $R^2 = 1/2$, the upper bound on efficiency is 8/9th. That is, expanding the sample by 50 percent and using a set of instruments for the missing x_1 which correlate with it with an R^2 of about .5, will yield a reduction in the standard error of $\hat{\beta}_1$ of less than 6 percent. And, if x_1 were a more important variable (β_1 higher than assumed) and the equation were better specified (σ^2 lower than assumed), the gain would be even smaller.

Up to now we have talked about the efficiency of "first order" methods, methods that do not re-weight the samples, do not iterate, and do not use the information on y (the dependent variable) to infer the missing values of x_1. The general analysis of such methods is beyond the scope of this paper, but we want to sketch out briefly the potential sources of additional gains in efficiency, if any.

The simple model that we started out with at the beginning of this section (with $\eta = 0$), can be rewritten in "reduced form" as:

$$x_{1a} = \delta x_{2a} \qquad\qquad + v_a$$

$$y_{1a} = (\beta_2 + \beta_1\delta)x_{2a} + e_a + \beta_1 v_a$$

$$y_{2b} = (\beta_2 + \beta_1\delta)x_{2b} + e_a + \beta_1 v_b$$

where x_1 has been solved out of the y equation. Sample A provides direct estimates of δ, $(\beta_2 + \beta_1\delta)$ and an estimate of β_1

from the covariance of the residuals from the first two equations. Thus, it also yields an estimate of β_2. Sample B provides an additional estimate of $\beta_2 + \beta_1\delta$, but no additional information on δ and very little on β_1. Most of the extra information derivable from sample B is translated into more information on β_2, the coefficient of the non-missing variable. The additional information about β_1 in a full information context arises out of improving the estimates of reduced form residuals from the y_{1a} equation by imposing the constraint that the coefficients of x_2, $\beta_2 + \beta_1\delta$, should be the same in both parts of the sample. The "first order" method uses the residuals estimated from sample A alone to derive δ and β_1 and uses sample B only to improve β_2. A full maximum likelihood procedure would improve the estimate of β_1 slightly by imposing the restriction that the reduced form residuals in samples A and B be based on the same coefficients and by reweighting the contribution of samples A and B to the estimated β_2. But there are no more observations with additional direct information on β_1, only an improvement in the estimates of the residuals in sample A. Since the first stage (sample A alone) residuals are consistent and since we are talking in the context of large samples, the potential gains from such full-information procedures are unlikely to be large.[8]

III. NON-RANDOMLY MISSING DATA

In this section we shall consider the possibility that the missing x_1's are not missing at random, the probability of missing being somehow related to v, the residual in the x_1 equation. We shall still maintain the assumption that v and e are independent, and concentrate only on improving our estimates of $\hat{x}_{1b}$, without regard for the possible additional information contained in the y equation. The reason for such a single-equation approach is that we are trying to prepare a large body of data for subsequent analysis. There will be many y's we shall want to consider (different variables, different subsamples) and it is both computationally and conceptually inefficient to try and solve the massing data problem anew for each particular subproblem that will arise in the future.

In our sample of NLS Young Men we have, as of 1966 (the first year of the survey), 5,094 young men with complete data on such major variables as schooling and scores on a test of the "knowledge of the world of work" but only 3,324 of these (65%) have IQ scores in their records.[9] It was decided to "fill-in" the missing IQ values using the available data as of 1966 on schooling, parental background, race, region, and scores on a test of the "knowledge of the world of work." The suspicion was raised, however, that IQ scores may be missing non-randomly, in the sense that those with low IQ scores might be less willing to give permission to collect the scores from their original high schools, and that schools frequented by such students may not have their records in good order and might have greater difficulty in retrieving such scores. The following model formalizes such considerations.[10]

Let D = 1 when an IQ score is observed and D = 0 when it is unobserved. We assume that D = 1 when a set of observable variables Z and random component u_1 together exceed some threshold value, which can be set at zero, without any loss of generality. That is

$$D = 1 \text{ if } Z\delta + u_1 \geq 0$$

$$D = 0 \text{ if } Z\delta + u_1 < 0$$

In addition, we have the IQ equation:

$$IQ = X\beta + u_2$$

where X is another set of independent variables which may contain or be contained in Z. We only observe IQ, however, when D = 1. Our problem now is one of a missing dependent variable. While $Eu_2 = 0$ is assumed to be true for the population at large, this need not be the case for the observed subsample (sample A). In particular, for the observed data

$$\begin{aligned} E(IQ|X,Z, \text{ and } D = 1) &= X\beta + E(u_2|X,Z, \text{ and } D = 1) \\ &= X\beta + b_{21}E(u_1|Z, \text{ and } D = 1) \neq X\beta \end{aligned}$$

where we have assumed that u_1 and u_2 have a bivariate normal distribution with means zero, variances σ_1^2 and σ_2^2, correlation ρ_{12}, and $b_{21} = \text{Cov}(u_1u_2)\sigma_1^2 = \rho_{12}\sigma_2/\sigma_1$. As long as there is a correlation between the random term in the selection-into-the-sample equation, and the random term in the equation that we are interested in estimating ($\rho_{12} \neq 0$) and $E(u_1|D = 1) \neq 0$, limiting our attention to the "complete data" subsample will not do. First-order methods will not provide consistent estimates of the missing values because the OLS coefficients derived from sample A are not consistent estimators of β. To show the source and magnitude of this inconsistency, we have to evaluate $E(u_1|D = 1)$ which is the mean of the normal random variate u_1 truncated from below at the point $-Z\delta$. For notational simplicity let us divide the selection equation by σ_1 and rename $Z\delta/\sigma_1 = c$ and $u_1/\sigma_1 = \varepsilon$, where c is now a scalar (differing for each individual observation) and ε is the unit normal variate. Then,

$$E(u_1|u_1 > -Z\delta) = \sigma_1 E(\varepsilon|\varepsilon > -c), \text{ and}$$

$$E(\varepsilon|\varepsilon > -c) = \frac{1}{1 - F(-c)} \int_{-c}^{\infty} \varepsilon f(\varepsilon) d\varepsilon$$

where $f(\varepsilon)$ and $F(\varepsilon)$ are the normal probability density and cumulative normal functions, respectively. Since the derivative of the normal density function is $-\varepsilon f(\varepsilon)$, and $1 - F(-c) = F(c)$, the above expression simplifies to

$$E(\varepsilon|\varepsilon > -c) = \frac{1}{F(c)} \cdot -f(\varepsilon)\Big]_{-c}^{\infty} = \frac{f(c)}{F(c)}$$

since $f(\infty) = 0$, and $f(-c) = f(c)$. Calling the ratio $f(c)/F(c) = M(c)$ (its reciprocal is known in the statistical literature as the "Mills ratio") and collecting terms, we have

$$E(IQ|X, Z\delta/\sigma_1, D = 1) = X\beta + \rho_{12}\sigma_2 M(Z\delta/\sigma_1)$$

As long as there is a correlation between the Z's and the X's and $\rho_{12} \neq 0$, OLS estimates of β which ignore the M term will be biased.[11] There are two ways out of this: (1) As suggested by Heckman (1976), one can estimate δ/σ from a probit analysis of the qualitative variable D (IQ present) as a function of the observed Z's, evaluate the M term explicitly and separately for each observation using the estimated $Z_i\hat{\delta}/\sigma_1$ and add this term to the regression (in sample A) of IQ on X, converting it into a regression of IQ on X and M. Heckman shows that this will yield consistent estimates of β and ρ_{12}. While this procedure is not fully efficient since it does not allow for the rather complicated heteroscedasticity introduced by substituting $Z\hat{\delta}/\sigma_1$ for the "true" $Z\delta/\sigma_1$, and does not utilize information contained in sample B, it is relatively simple and rather convenient for exploratory data analysis.

(2) Alternatively, one can compute the joint maximum-likelihood estimates of β, δ, ρ_{12}, and σ_2 (σ_1 is not, in general, identified in such models and is set routinely to 1). Assume that the first s observations are those with D = 1 (IQ present), out of a total of N, and express the expected value of u_1 in terms of u_2 as $E(u_1 | u_2) = (\rho/\sigma_2)u_2$, where we have dropped the subscripts on ρ for notational ease. Then the log likelihood of an observation with data on IQ is

$$LLF_i(D = 1) = -\ln\sigma_2 - \frac{1}{2}\left(\frac{u_{2i}}{\sigma_2}\right)^2 + \ln F\left(\frac{Z_i\delta + (\rho/\sigma_2)u_{2i}}{(1 - \rho^2)^{\frac{1}{2}}}\right)$$

while the log-likelihood of a missing data observation is

$$LLF_i(D = 0) = \ln F(-Z_i\delta),$$

where

$$F(u) = \int_{-\infty}^{u} e^{\frac{-u^2}{2}} \Big/ \sqrt{2\pi},$$

and we have set $\sigma_1 = 1$. Then the summed log-likelihood function for the combined sample (A + B) is

$$\ln L = -s \ln\sqrt{2\pi}\sigma_2 - \frac{1}{2}\sum_{i=1}^{s}(u_{2i}/\sigma_2)^2 + \sum_{i=1}^{s}\ln F\,\frac{(Z_i\delta + (\rho/\sigma_2)u_{2i})}{(1 - \rho^2)^{\frac{1}{2}}}$$

$$+ \sum_{i=s+1}^{n} \ln\,[1 - F(Z_i\delta)]$$

which can be evaluated directly by non-linear optimization methods. The advantage of the MLE procedure is that it yields the correct asymptotic standard errors for the estimated coefficients, which is not the case for the estimation procedure suggested by Heckman. One of the goals of this paper is to compare the relative cost and performance of such procedures in a substantive research context.

To implement the Maximum Likelihood procedure, we use an algorithm proposed by Berndt, Hall, Hall, and Hausman (1974). It is similar to the method of scoring, but instead of requiring large sample expectations to be taken to evaluate R. A. Fisher's information matrix, it relies on the result that in large samples the covariance of the gradient is a consistent estimate of the information matrix in the neighborhood of the optimum. Letting θ be the parameter vector and the gradient $g(\theta) = \frac{\partial \ell}{\partial \theta}\Big|_{\hat{\theta}}$, its asymptotic covariance in the neighborhood of the optimum is approximated by

$Q(\hat{\theta}) = \sum_{i=1}^{T} \frac{\partial f_i}{\partial \theta}\Big|_{\hat{\theta}} \frac{\partial f_i}{\partial \theta}\Big|_{\hat{\theta}}'$ where f_i is the log likelihood of the ith observation. Thus, the method requires only computation of first derivatives. The updating formula at the jth iteration for θ^{j+1} is then

$$\hat{\theta}^{j+1} = \hat{\theta}^{j} + \lambda^{j} Q(\hat{\theta}^{j})^{-1} g(\hat{\theta}^{j})$$

where $\lambda^j > 0$ is the stepsize chosen in the direction $Q(\hat{\theta}^j)^{-1} g(\hat{\theta}^j)$. The stepsize λ^j is chosen according to the criterion in Berndt, et al., page 656, and convergence to a local maximum is assured. The algorithm has the desirable "uphill" property: an improvement in the likelihood function occurs at each step. Experience with this algorithm has been very satisfactory as long as the derivatives are calculated correctly. The algorithm can be considered a generalization of the Gauss-Newton algorithm. Estimates of the asymptotic covariance matrix of the estimates follow from $Q(\hat{\theta}_{ML})^{-1}$, where $\hat{\theta}_{ML}$ is the value of which maximizes the likelihood function. Care must be taken to insure that the global maximum has been found although in practice no difficulties arose.

Table 1 lists our estimates for the "being in sample A" (IQ present) equation. Table 2 lists the major coefficients in the estimated IQ equation based on (a) OLS, (b) OLS with the added M term (the Heckman procedure), and (c) the MLE results.[12] For this example, there is little difference in the estimates of the first equation (IQ present) listed in Table 1. The coefficients of the independent probit equation and the MLE coefficients are almost equal to two decimal places. There is a bit more happening in Table 2. The addition of the M variable shifts some of the coefficients noticeably, by 10 to 20 percent, even though its own coefficient is not "statistically significant" at the conventional significance levels. This is a reflection of the rather high multicollinearity between the M term, which is a non-linear function of the variables listed in Table 1, and some of the variables included in the IQ equation. The MLE estimates turn out to be between the OLS and the "Heckman" estimates, indicating that the latter may overshoot somewhat in their adjustment for selectivity bias, but the changes are not very large. In the OLS with M version, the M term is not statistically significant. In the MLE equations, the

Table 1. Estimates of the "Having-Good-Data" (IQ present) Equation. NLS Young Men, N = 5094

Variables	Coefficients (standard errors) Probit	 MLE
Constant	.635 (.183)	.635 (.180)
Black	- .607 (.058)	- .608 (.057)
FOMY 14	- .036 (.013)	- .035 (.013)
Culture	.140 (.028)	.141 (.028)
Together	.123 (.060)	.122 (.060)
Age 66	- .073 (.009)	- .074 (.010)
KWW	.0066 (.0036)	.0065 (.0035)
S66	.110 (.017)	.110 (.017)
SLT9	-2.69 (.120)	-2.70 (.118)
Log L	-2038	-2010

FOMY 14 = Occupation of father (or heard of household) when respondent was 14 scaled by the median earnings of all U.S. males in this occupation in 1959, in thousand dollars.

Culture = Index based on the availability of newspapers, magazines, and library cards in the respondent's home.

Together = Both parents in household when respondent was 14.

KWW = Score on the "Knowledge of the World of Work" test administered in 1966.

S66 = Schooling completed in 1966.

SLT9 = Dummy variable, equals 1 when S66 less than 9.

IQ = Score on IQ type tests collected from the high school last attended by the respondent.

Table 2. Alternative Estimates of the IQ Equation: NLS Young Men

Variables	Coefficients (standard errors) of Major Variables		
	OLS	OLS including estimated M variable from probit	MLE
	N = 3324	N = 3324	N = 5094
Black	-9.94 (.64)	-11.04 (.80)	-10.69 (.72)
FOMY 14	.248 (.129)	.196 (.131)	.213 (.133)
Culture	.89 (.31)	1.11 (.32)	1.04 (.31)
KWW	.608 (.036)	.619 (.036)	.615 (.036)
S66	1.98 (.17)	2.17 (.19)	2.11 (.19)
MED	.366 (.097)	.371 (.097)	.369 (.095)
M		3.71 (2.29)	
ρ	0	.304	.202 (.100)
σ_2	12.21	12.20	12.31
Log L	-13035	-13032	-13060

Other variables in the equation: Siblings, RNS 66, FED, FAMY 66, Age 66.

MED = Mother's education.

FED = Father's education.

M = "Inverse Mills Ratio," $f(Z\hat{\delta})/F(Z\hat{\delta})$ computed from the probit equation given in Table 1.

Siblings = Number of siblings.

RNS 66 = Region South in 1966.

FAMY 66 = Total Family Income in 1966.

See notes to Table 1 for additional definitions.

The standard errors for the OLS with M equation are only approximate, since they ignore the heteroscedasticity introduced by the estimated M term.

estimated ρ is statistically significantly different from zero, but not very high. A more general test for selection bias is given by comparing the summed likelihood of the separately estimated in-sample probit and OLS IQ equation and the jointly estimated maximum likelihood: twice the difference in the log-likelihoods is 3.74 which is to be compared to a critical $\chi^2(1) = 3.84$ at the 5 percent significance level. Thus, one could maintain the hypothesis that $\rho = 0$ and that no bias is introduced by treating the missing values as random.[13] But, because the estimated bias is on the borderline of statistical significance and because we are interested in how much difference the procedure makes substantively, we shall proceed to estimate the missing IQ values using the MLE results.

To estimate the missing IQ values we use not only the estimated coefficients from the IQ equation but also the knowledge that the values are missing. That is, by a similar argument as before

$$\begin{aligned} E(IQ|D = 0) &= X\beta + \rho\sigma_2 \; E(u_1|-Z\delta > u_1) \\ &= X\beta - \rho\sigma_2 \; f(-Z\delta)/F(-Z\delta) \\ &= X \;\; - \rho\sigma_2 \; f(Z\delta)/[1 - F(Z\delta)] \end{aligned}$$

and hence the missing IQ values are estimated by

$$\hat{IQ} = X\hat{\beta} - \hat{\rho}\hat{\sigma}_2 \; f(Z\hat{\delta})/[1 - F(Z\hat{\delta})]$$

where we have normalized $\sigma_1 = 1$.

We now apply this machinery to the estimation of an earnings function using the 1973 data base for these same young men. In 1973, there were 2,246 young men who were not enrolled in school, interviewed, and had complete data on schooling, wage rates and work experience.[14] However, only 1,540 of the 2,246 had IQ scores. Table 3 lists the means, standard deviations, and correlation coefficients of three estimates of the missing IQ scores (N = 706) T_1--based on the OLS regression, T_2--based on the coefficients of the OLS + M regression and the missing values adjustment using the independently estimated probit equation to evaluate the $f(Z\delta)/[1 - F(Z\delta)]$ term, and T_3--based on the Maximum Likelihood estimates, including also the $-\rho\sigma_2 \; f(Z\delta)/[1 - F(Z\delta)]$ term. Table 4 shows the estimated schooling and IQ coefficients in the 1973 earnings function, separately for the "complete" and "incomplete" subsamples and for the combined total sample using all three ways of computing

Table 3. Predicted IQ Scores for Working NLS Young Men in 1973 with IQ Missing (N = 706)

	Mean	Standard Deviation	Correlation Coefficients	
			T_2	T_3
T_1	88.4	14.7	.991	.996
T_2	84.0	14.3		.999
T_3	85.5	14.4		

$T_1 = X\hat{\beta}$; $\hat{\beta}$ from column 1 of Table 2.

$T_2 = X\tilde{\beta} - \tilde{\rho}\tilde{\sigma}_2 f(Z\tilde{\delta})/[1 - F(Z\tilde{\delta})]$; $\tilde{\beta}$ and $\tilde{\rho}\tilde{\sigma}$ from column 2 of Table 2, $\tilde{\delta}$ from column 1 of Table 1.

$T_3 = X\beta^* - \rho\sigma_2 f(Z\delta^*)/[1 - F(Z\delta^*)]$, β^* and $\rho\sigma_2$ from column 3 of Table 2, δ^* from column 2, Table 1.

Table 4. Working NLS Young Men in 1973. Estimates of the Coefficients of Schooling and IQ for Different Subsamples and Alternative Treatment of Missing IQ Values

Sample and Type of IQ and Variables	Coefficient (standard error)		Residual Standard Error
	Schooling	IQ	
A: "Good data" Subsample, N = 1540			
1.	.064 (.005)		.387
2. IQ	.055 (.006)	.0031 (.0008)	.385
B: Missing IQ Subsample, N = 706	.058 (.006)		.394
A + B: Combined			
N = 2246			
1.	.064 (.004)		.391
2. IQ + T_1	.051 (.004)	.0043 (.0008)	.388
3. IQ + T_2	.052 (.004)	.0042 (.008)	.388
4. IQ + T_3	.052 (.004)	.0042 (.0008)	.388

Dependent variable LW73. Other variables in the equation: AFEXP, XBT, SMSA, RNS 66, Black.

AFEXP = Length of military service, in years.

XBT = $e^{-0.1 \text{ EXP } 73}$, EXP73 = postschool work experience estimated on the basis of the work record since 1966 and the date of first job after school or the date stopped school (in years; truncated at age 14) if respondent reported starting work earlier.

See Table 3 for definition of T_1, T_2, and T_3.

the missing IQ values.

Let us first look at the biased schooling coefficients in samples A and B, excluding the IQ variable whose values are partially missing. We called them $\pi = (\alpha + \beta\delta)$ in Section II. They are given in lines A1 and B1 of Table 4. Though sample B indicates a somewhat lower return to schooling (.058 v. .064 for A), the difference is not statistically significant and we can maintain the hypothesis that sample B comes from the same population as A. The second part of the table indicates, as might have been expected from the earlier results, almost no difference in the performance of the different ways of filling-in the missing IQ scores. Our estimate of the schooling coefficient holding IQ constant is .055 (.006) in the "all good data" subsample and .052 (.004) in the combined sample with extrapolated IQ measures. The difference is hardly significant.[15] Moreover, the apparent gain in precision is exaggerated, since the computed standard errors do not take into account either the heteroscedasticity introduced by the extrapolated IQ measures or the fact that these measures are based on estimated rather than population coefficients.

The procedure discussed in this section was a partial one: we treated the problem of filling-in missing IQ values separately from the problem of estimating the earnings equation and we did not allow for the possibility that the disturbance in the variable (IQ)-present equation. In this particular substantive context, there seemed no need to move in that direction. The next sections will take up, however, examples where such a dependence could be quite important.

IV. SAMPLE SELECTION BIAS OVER TIME

In this section we come closer to utilizing the panel nature of our data. The conventional use of such data has been to pick a particular year, either latest available (1973) or one that has the more interesting data (e.g., answers to the union membership question in 1971) and analyze it separately, as if it were a single cross-section. Or at best, as we shall do in the following section, compare it to the results for another year, treating them as an independent replication. Alternatively, in trying to utilize fully the recently developed methods of error-components analysis, one tends to focus on the subsample of those observations that are available and are "good" for all years in the sample. The first approach results in varying sample size and serious questions about the comparability of the results over time. The second approach tends to focus on a rather small subset of the original data, raising serious questions about the representativeness and generalizability of such results.

It is our intent in this section to explore the effects of such sample selection rules on the estimates of the standard semi-log wage equation, focusing in particular on their effects on the estimated coefficient of years of school completed. These are only the first steps in this direction. We are trying out only rather simple models and not allowing yet fully for the dynamic aspects of the problem. What follows is a report on the exploratory stage of our analysis.

In trying to estimate the wage equation over time we face several sources of sample change and attrition. (1) Non-interview in a particular year, which can be divided roughly into (a) a permanent loss of the respondent from the sample and (b) temporary loss, primarily because of service in the armed forces. (2) The fraction of respondents still enrolled in school full-time is declining over time. If one insists on having only those who are out-of-school and working early in the history of the panel (1966-1968), one tends to lose very large fractions of the potential population. (3) "Bad data." Having "fixed-up" the missing IQ measures in Section III, the major other source of missing data is the experience variable. It is based on the job history of the individual and that is often incomplete in the record. Table 5 presents a distribution of the fraction not-in-the-sample by major reason by year. It illustrates the two major and contradictory forces at work over time: Attrition versus completion of schooling.

While each of the reasons for being out-of-the-sample may be caused by somewhat different forces and should perhaps be analyzed separately, we focus instead on specifying a single probit equation for being-in-the-sample, as a function of the major socio-economic variables as observed in 1966.[16] It can be thought of as a kind of reduced-form equation summarizing

Table 5. Sources of Change in Sample Composition. NLS Young Men, 1966-1973. N = 5094[a]

	Percentage by Year							
Category	1966	1967	1968	1969	1970	1971	1973	1966-1973 combined[b]
In sample[c]	34.3	37.7	39.9	41.7	41.9	41.7	45.6	14.6
Out-of-sample: Non-interview	0.0	3.2	6.5	9.4	10.8	13.6	18.0	19.9
In Armed Forces[d]	0.0	5.0	10.8	13.4	12.6	9.7	5.0	24.3
Enrolled in School (full-time)	62.0	49.9	37.5	26.9	19.9	15.8	9.9	35.7
"Bad" data	3.6	4.2	5.3	8.5	14.7	19.2	21.6	5.5

[a] Based on "good" data for KWW and S66 in 1966. Original panel size was 5,225. The discrepancy of 2.5% could be added to the "Bad" data row.

[b] Not in school and good data in all years. Obviously an observation may be missing from the sample for different reasons in different years. The hierarchy used in this column was first, in Armed Forces any year, second, not interviewed for another reason any year, third, enrolled full-time in school any year, and finally, missing other data.

[c] Being not-enrolled in school (full-time) and having data on wage rates and work experience in the particular year.

[d] To the extent known and stated in the record.

the net effect of the various variables on the sample participation probability. Table 6 presents estimates of the major coefficients in such probit equations. These are to be used in turn, in the Heckman (1976) mode, to construct the inverse Mills ratio variables, which are then added to the estimated wage equations. Their coefficients are also used as initial values for the MLE algorithm. The changing character of the selection process can be seen clearly in this table. Early on, those in the sample are distinctly older than the group as a whole, with less schooling, and a lower estimated IQ. As time goes by, these differences attenuate and by 1973 the major reasons for not being in the sample are non-interview and bad data rather than further full-time study. The first two tend to be related negatively to personal and family socio-economic characteristics while the third is related positively. By 1973 their effects come close to cancelling each other out.

This can also be seen in Table 7, which lists the estimated coefficients of schooling in the wage equation by year, before and after allowing for sample selectivity bias. Before 1970 the sample selectivity bias is positive, causing an overestimate of the schooling coefficient. By 1971 it changes sign, implying an underestimation of the true coefficient. There is some indication here, as was also true in the earlier section, that the Heckman-type estimates tend to "overadjust" for selectivity bias, since the MLE estimates are by and large between the OLS and the Heckman-type estimates.

The major difficulty with the estimates presented above is that they lump together the various reasons for not being in the sample into one selectivity equation. Actual sample selectivity occurs because of rather different reasons and on different and somewhat independent margins. To take adequate account of this would require the development of multivariate truncation models, a topic to which we hope to return but which is beyond the scope of this paper. We can, however, narrow the problem by focusing only on a subpopulation where sample selectivity is based on one set of considerations only and hence can be modeled more legitimately by a single selection equation. In the following section we limit ourselves to the subset of interviewees with good data, interviewed in 1971 and 1973, and focus solely on their dropping out from or finishing school and going to work decision as the source of sample selection bias.

Table 6. Major Coefficients of the Probit Equation for Being in the Working--Out of School--Good Data Samples. NLS Young Men, 1966-1973. N = 5094

Variables	Year							1968- 1973 Combined
	1966	1967	1968	1969	1970	1971	1973	
Age 66	.37 (.01)	.25 (.01)	.19 (.01)	.13 (.01)	.10 (.01)	.10 (.01)	.08 (.01)	.24 (.01)
S66	-.12 (.02)	-.03 (.01)	-.02 (.01)	-.01 (.01)	-.01 (.01)	-.00 (.01)	-.03 (.01)	-.03 (.02)
Black	-.21 (.07)	-.14 (.06)	-.14 (.06)	-.20 (.06)	-.30 (.06)	-.31 (.06)	-.29 (.06)	-.27 (.08)
FOMY 14	.06 (.01)	.03 (.01)	.04 (.01)	-.03 (.01)	-.04 (.01)	-.02 (.01)	-.02 (.01)	-.07 (.02)
IQ 3	-.019 (.002)	-.015 (.002)	-.013 (.002)	-.012 (.002)	-.008 (.002)	-.006 (.002)	.001 (.002)	-.006 (.002)
N in sample	1750	1920	2032	2125	2136	2123	2321	742
$\chi^2(13)$	2742	1882	1331	734	474	406	359	1122

Standard errors in parentheses.

Other variables in equation: MED, FED, Culture, SIBS, Together, RNS 66, SMSA 66, KWW. See notes to Tables 1 and 2 for definition of variables.

$\chi^2(13)$--relative to the log likelihood of the simple odds of being in the sample. Critical value at the 5 percent level is 22.

Table 7. Estimated Coefficient of Schooling in Different Years, Allowing for Sample Selectivity Bias, NLS Young Men, 1966-1973

Year	Coefficient of Schooling			Estimated		SEE	
	Without M	With M	MLE	ρ	$\tilde{\rho}$	Without M	With M
1966	.0525	.0413	.0431	-.45*	-.37*	.370	.363
1967	.0545	.0375	.0416	-.71*	-.49*	.337	.332
1968	.0540	.0444	.0451	-.48*	-.41*	.335	.334
1969	.0544	.0478	.0447	-.48*	-.61*	.327	.326
1970	.0662	.0632	.0576	-.27	-.60*	.345	.345
1971	.0638	.0673	.0660	.30	.18	.352	.352
1973	.0481	.0510	.0503	.39	.34*	.402	.402
Combined 1966-1973 subset	.045	.048	n.c.	.17*	n.c.	.321	.320

Dependent variable: LW = Logarithm of the wage rate on current or last job.

Other variables in the equation: Black, IQ 3, XBT, SMSA, RNS 66. In the combined 1966-1973 subset also year dummies for each of the survey years.

M = Inverse Mills ratio, computed from the probit equations described in the previous table.

MLE = Maximum Likelihood Estimate of the schooling coefficient.

ρ = Estimated correlation between the disturbances in the sample selection and wage equation.

$\tilde{\rho}$ = ρ from MLE.

SEE = Estimated residual standard error in the wage equation.

* = Statistically significantly different from zero at the .05 level.

n.c. = Not computed.

V. SAMPLE SELECTION BIAS AND ENDOGENEITY

In this section we consider the case of a dependent variable "missing" or not being relevant to the assumptions of the model, for a significant fraction of the sample.[17] For example, a significant fraction of the young men in the NLS were still enrolled in school full time at the time of the latest available survey (1973). For these young men there is either no wage information in the record or the wages recorded are likely to be associated with a part-time job, not measuring adequately their current or future human capital. Also the information on their schooling level may be misleading, since it cannot be taken as "completed." The usual solution to this problem has been to limit estimation to the "good" and relevant data subset, ignoring the possibility that those who have chosen to stay in school longer may not be a random sample with respect to the determinants of the returns to schooling, the major focus of the equations to be estimated.

In tackling this problem we have to extend the model outlined in Section III significantly. It is not enough just to specify a "being-in-sample" (out-of-school) equation and estimate it jointly with the earnings function. The trouble arises because the major variable of interest in the earnings equation--schooling, is itself a product of the operation of the in-sample (dropping-out-of-school) equation in the previous years. Allowing for correlation between the disturbances in the in-sample and earnings equations implies also a correlation between completed schooling and the disturbance in the earnings equations, unless we were to make rather strange no-serial correlation assumptions. After all, completed schooling is just the integral of the in-school equation (the complement of the out-of-school equation), containing all the lagged values of the disturbances in that equation. Hence, there is no escape from treating schooling itself as an endogeneous variable in this context. This leads us then to the problem of estimating a system of simultaneous equations in the context of limited dependent variables, an area in which very little work has been done to date.[18]

Our model can be described as follows: Let S* be "desired" (completed) schooling. Actual schooling S is assumed to equal S* if the person is not in school and working. Thus, being in sample is defined by

$$S \geq S^*$$

If a person is still in school then $S < S^*$. The ultimate desired level of schooling S* is given by the equation

$$S^* = Bb_1 + Xb_2 + e_2$$

where B is a set of family background variables and X is a set of variables such as age, armed forces experience, and IQ, whose impact on desired versus actual schooling may not be the same. Actual schooling equals desired schooling if out-of-school. For in-school people actual schooling is a function of their age and the "speed" with which they are making scholastic progress. This "speed" depends on IQ, on the presence of absence of an interruption such as service in the armed forces, and on a set of other unmeasured random variables u. Thus actual schooling S if still in-school (out-of-sample) is given by

$$S = Xb_3 + u$$

Being-in-the-sample then occurs when

$$Xb_3 + u \geq Bb_1 + Xb_2 + e_2$$

Or, defining $e_1 = u - e_2$, when

$$e_1 \equiv u - e_2 \geq [B\beta_1 + X(\beta_2 - \beta_3)]$$

This points out the expectation that the coefficients of the background variables such as father's occupation or mother's education should be proportional to each other in the sample selection and completed schooling equations, while other variables such as age or IQ need not satisfy this constraint. It also suggests the probability of a negative correlation between the in-sample and schooling equations.

To discuss the statistical aspects of this model we shall rewrite all this more compactly by combining the B and X matrix into one matrix Z of exogeneous variables, relabeling $S = S^*$ for the in-sample people as y_2, and attaching to the model a wage (y_3) equation. All this then leads to a three equation model, consisting of the in-sample (not-in-school) equation

$$pr(D = 1) = pr[Z_1\delta + e_1 \geq 0],$$

the completed schooling equation

$$y_2 = Z_2\gamma_2 + e_2$$

and the wage equation

$$y_3 = \beta y_2 + Z_3\gamma_3 + e_3,$$

where y_2 is schooling (in years), y_3 is the logarithm of the wage rate, $Z_1\delta = -[Bb_1 + X(b_2 - b_3)]$, $Z_2\gamma_2 = Bb_1 + Xb_2$, and Z_3 is another set of exogeneous independent variables which may overlap with Z_1 and Z_2. The identification of β depends, in part, on some of the Z's in the schooling equation not entering the wage equation directly. Note that the system as written down is triangular, simultaneity occurring because we have allowed for possible correlations among the e's. We assume that the e's have a joint multivariate normal distribution, with a mean vector of zero, and variance covariance matrix Σ,

$$\Sigma = \left[\begin{array}{c|cc} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \hline \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{array}\right] = \left[\begin{array}{c|c} 1 & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$$

where we have set $\sigma_{11} = 1$. Then the probability of observing <u>in-sample</u> respondents can be written as the product of the conditional probability of the observation being in the sample given, Z_1, e_2 and e_3, and the marginal probability of observing e_2 and e_3 given Z_2 and Z_3:

$$pr(D_i = 1) = pr(Z_1\delta + e_1 \geq 0|e_2,e_3,Z_1)\ f(e_2,e_3|Z_2,Z_3)$$

Let us call the vector $[e_2,e_3] = e'$. Expressing e_1 in terms of e_2 and e_3, we have for the log likelihood function of the out-of-school observations:

$$LL_a = \log F\left(\frac{Z_1\delta + \Sigma_{12}\Sigma_{22}^{-1}e}{(1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{\frac{1}{2}}}\right) - \frac{1}{2}\log|\Sigma_{22}| - \frac{1}{2}e'\Sigma_{22}^{-1}e$$

where the first term corresponds to the probability of being in-the-sample (out of school) as a function of e_2 and e_3 and Z_1, with $\Sigma_{12}\Sigma_{22}^{-1}$ playing the role of b_{12} in the earlier formulae, the denominator being the conditional standard deviation of e_1, and where the second and third terms correspond to the probability of observing the particular e_2 and e_3.

The log likelihood of the in-school, out-of-sample, observation is given, as before by

$$LL_b = [1 - F(Z_1\delta)]$$

We need now to transform these equations from the unobserved e's into the observable vector $[y_2 y_3]$. Using the transformation

$$u = \begin{bmatrix} u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} y_2 - Z_2\gamma_2 \\ y_3 - \beta y_2 - Z_3\gamma_3 \end{bmatrix}$$

and denoting the Jacobian of the transformation by J, we have

$$LL_a = \log F\left(\frac{Z_1\delta + \Sigma_{12}\Sigma_{22}^{-1}u}{(1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{\frac{1}{2}}}\right) - \frac{1}{2}\log|\Sigma_{22}|$$

$$- \frac{1}{2}u'\Sigma_{22}^{-1}u + \log|J|$$

Since our model is triangular,

$$J = \begin{bmatrix} 1 & 0 \\ \beta_3 & 1 \end{bmatrix}, \quad \log|J| = 0,$$

and we are left with a multivariate least squares type likelihood function.[19] Thus the combined log likelihood function

$$\log L = -\frac{s}{2}\log|\Sigma_{22}| - \sum_{i=1}^{s} u_i'\Sigma_{22}^{-1}u_i$$

$$+ \sum_{i=1}^{s}\log F\left[\frac{Z_{1i}\delta + \Sigma_{12}\Sigma_{22}^{-1}u_i}{\left(1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{\frac{1}{2}}}\right]$$

$$+ \sum_{i=s+1}^{n}\log[1 - F(Z_{1i}\delta)]$$

is to be maximized with respect to δ, γ_2, γ, β, and the elements of Σ_{12} and Σ_{22}^{-1}. This is achieved using the algorithm described in Section III.[20]

The sample used and the wage equation to be estimated in this section differ from that considered in the previous sections in three ways: (1) Not to treat too many problems at once we limit our sample here to those having IQ scores. As can be seen from the results in Section III, this should not lead to much bias in our results. (2) We limit the working and out-of-school subsample to those who have good wage data and are working full time (35 hours or more per week).[21] And (3) because we shall be treating schooling as endogeneous, and because our experience measure, though independently computed, cannot be taken as independent of schooling, we reparameterize the wage function, using age as a variable instead of experience.

We estimate this model for the two latest years of our panel, 1971 and 1973.[22] As the result of the restrictions described above our samples consist of 2,176 and 2,419 young men in 1971 and 1973, respectively, with 1,570 and 2,057 of them or 72 and 85 percent, respectively, having left school and working full time.

Table 8 gives the results of estimating the three equation model outlined above for these samples. Part A shows the estimates of the first two equations. The major determinants of being in-the-sample (out of school) are age, IQ, military experience, and the culture index. There is, apparently, an attempt to make up for lost schooling due to military service and possibly also a response to the schooling subsidies available to army veterans. Total (completed) schooling is strongly related to parental background, IQ, and negatively to military service. Other things being equal, young blacks had more schooling than would have been predicted for them by their measured parental background and IQ.

The restriction that family background variables should have proportionally similar coefficients in the in-sample and completed schooling equations is satisfied only approximately. The ratio of the respective coefficients in the two equations (for the 1973 estimates) moves from -.13 for FOMY 14 to -.3 for FED. Culture, with a ratio of -.7 has apparently an independent additional effect on the speed with which desired schooling is accomplished. A formal test of the proportionality restriction for four family background coefficients (FED, MED, FOMY 14, and SIBS) across the two equations yields the following $\chi^2(4)$ test statistics: 9.4 for the 1971 estimates and 6.6 for the 1973 ones. Since the critical (.05) value for $\chi^2(4)$ is 9.5, these results are not too surprising and are consistent with the model outlined above.

Part B of Table 8 presents several estimates of the wage equation, starting with OLS, going to a two equation sample selection model similar to the one estimated in the previous

Table 8. Estimates of a Three Equation Model for NLS Males, 1971 and 1973

1971: N = 2,176 (1,570 out of school
with wage data and good IQ)

1973: N = 2,419 (2,057 out of school
with wage data and good IQ)

Part A: In-sample and Completed Schooling Equations

Variables	In-sample (Out of School) 1971	In-sample (Out of School) 1973	Completed Schooling 1971	Completed Schooling 1973
IQ	-.028 (.002)	-.016 (.002)	.066 (.004)	.068 (.003)
FED	-.046 (.012)	-.029 (.013)	.092 (.018)	.097 (.016)
MED	-.007 (.014)	-.019 (.014)	.078 (.018	.090 (.017)
FOMY 14	-.018 (.018)	-.013 (.020)	.134 (.028)	.103 (.025)
Culture	-.156 (.048)	-.185 (.051)	.231 (.064)	.265 (.059)
SIBS	.013 (.015)	.015 (.016)	-.075 (.022)	-.067 (.019)
Age	.160 (.011)	.122 (.012)	.070 (.021)	.071 (.016)
AFEXP	-.076 (.034)	-.138 (.036)	-.183 (.066)	-.188 (.057)
Black	-.405 (.103)	-.346 (.116)	.838 (.164)	.744 (.141)
σ	1.0	1.0	1.73	1.81
ρ_{12}			-.43	-.36

Table 8--Continued

Part B: Different Estimates of the 1971 and 1973 Wage Equations

Part B1--1971

	Coefficients (standard errors)			
Variables	OLS	2 equation MLE	3SLS	3 equation MLE
Schooling	.019 (.005)	.020 (.005)	.048 (.015)	.070 (.016)
IQ	.0024 (.0008)	.0047 (.0009)	.0003 (.0013)	.0006 (.0014)
Age	.045 (.003)	.036 (.004)	.042 (.003)	.032 (.004)
Black	-.128 (.028)	-.104 (.030)	-.137 (.028)	-.124 (.033)
AFEXP	-.009 (.011)	-.006 (.011)	-.002 (.012)	.004 (.012)
σ	.347	.364	.350	.369
ρ_{13}		-.505		-.391
ρ_{23}			-.094	-.176
Part B2--1973				
Schooling	.019 (.004)	.020 (.005)	.051 (.014)	.069 (.014)
IQ	.0033 (.0007)	.0045 (.0008)	.0007 (.0013)	.0005 (.0013)
Age	.042 (.003)	.036 (.003)	.040 (.003)	.034 (.003)
Black	-.107 (.028)	-.090 (.032)	-.115 (.028)	-.103 (.033)
AFEXP	-.008 (.010)	+.0006 (.010)	-.001 (.011)	.011 (.012)
σ	.388	.404	.392	.412
ρ_{13}		-.509		-.464
ρ_{23}			-.116	-.175

Additional variables in equation: constant, SMSA, RNS 66.

Dependent variables: eq. 1 D = dummy variable for being out of school and working among all the persons interviewed in either 1971 or 1973 with good IQ data and other minor data restrictions; eq. 2 S71, S73 = highest grade of schooling completed for out-of-school youths; eq. 3 LW71, LW73 = logarithm of the wage rate per hour on current or last job.

2 equation MLE = Joint estimation of equation 1 and 3, assuming $\rho_{23} = 0$.

3SLS = Three stage least squares estimates of equations 2 and 3, ignoring sample selection bias problems.

Table 8--Continued

3 equation MLE = Joint maximum likelihood estimates of equations 1 through 3.

σ = Estimated standard deviations of the disturbances in the relevant equations. In the probit equation $\sigma_1 = 1$ by definition.

ρ's = Estimated correlation coefficients of the disturbances across the different equations.

section (assuming that $\rho_{12} = \rho_{23} = 0$), then to a Three Stage Least Squares estimate allowing for the endogeneity of schooling but not for potential sample selection bias in the wage equation, and finally to the complete 3 equation model MLE results. Note that we start out with a significantly lower schooling coefficient, because we are holding age rather than experience constant. A rough adjustment is just to add the age coefficient to the schooling coefficient.[23] The effect of allowing for selection bias is rather striking in these samples, especially once simultaneity is also allowed for. The schooling coefficient is raised by about .02 in both 1971 and 1973 (comparing column 4 to 3). The biggest impact, however, is from allowing the schooling variable to be endogeneous. It raises its coefficient by about .03 and drives the IQ coefficient to insignificance.[24] The effect is a bit smaller when one looks at the holding experience constant schooling coefficient, adding in the age coefficient: the estimated rate of return to schooling rises from 6.1 (6.4) percent in the 1973 (1971) OLS equation to 10.3 (10.2) percent in the 3 equation MLE model. Quite a difference. The rise in the coefficient of schooling occurs because those with a higher disturbance in the wage equation are more likely to be in school and out of the sample and because the disturbances in the schooling and wage equation are negatively correlated.

The signs of the correlations among the disturbances require some care in interpretation. Note that they are all negative. That $\rho_{12} < 0$ is understandable since being out-of-school, other things equal, reduces the level of completed schooling. That $\rho_{23} < 0$ has been observed also in other contexts (cf. Griliches, 1977). It can be interpreted as the result of errors in measurement in S and of the dependence of both wage rates and completed schooling on unmeasured initial human capital levels. These initial human capital levels have, however, opposite signs in the two equations, since a higher initial level requires less additional investment (formal schooling) to attain the ultimately desired level. The negative sign on ρ_{13} is also not too surprising since it implies that, other things equal, the better people (in terms of the wage equation disturbance) are still in school. It appears, however, to contradict the expected relationship among the signs, since both ρ_{23} and ρ_{12} are also negative. But that can be explained with recourse to the model outlined above. Recalling that $e_1 = u - e_2$, $\rho_{13} < 0$ implies that Cov $ue_3 <$ Cov $e_2e_3 < 0$, i.e., that the "speed" with which actual schooling is accomplished is also negatively correlated with the unobserved components of the wage equation.[25] This could be due to the same reasons as discussed above for the negative correlation between e_2 and e_3 and to a possible misspecification of the wage equation. Those who go through school faster may have lower wages because they tend to choose jobs with more non-pecuniary benefits (such as teaching) or more on-the-job training. In any case, there is no necessary inconsistency in the observed signs of the ρ's.

The estimated biases are not only of substantial size but also statistically significant. One can test the single equation OLS estimates against the 3 equation MLE's by asking the question whether the estimated $\hat{\rho}_{13}$, $\hat{\rho}_{23}$, $\hat{\rho}_{12}$, could all be zero. The estimated $\chi^2(3) = 77$ for 1973 is highly significant. It is unlikely that these results could have arisen by chance from a population with no correlation among the disturbances of the different equations. We get a similar answer when we consider these correlations in pairs or singly. For example, the restriction that both sample selection correlations be zero yields a $\chi^2(2)$ of 67 and 66 for 1973 and 1971 samples, respectively. Similarly, setting $\rho_{23} = 0$, results in a $\chi^2(1)$ of 6 and 10 in the 1973 and 1971 samples, respectively. All rather unlikely events.

Having tested and decisively rejected the joint null hypothesis of no correlation among the disturbances of all three equations, it is also interesting to ask if the estimates of the wage equation alone differ significantly. We are primarily interested in this equation and would like to assess the specific effect of our procedure on our estimates of it. To do so, we use a specification test developed by Hausman (1976). Under the null hypothesis of no self-selection, three stage least squares (3SLS) estimates of the schooling and wage equation are asymptotically efficient. Thus, Hausman's lemma 2.1 applies so that the asymptotic variance of the difference between the ML and 3SLS estimates is the difference in their respective variances. Performing a large sample χ^2 test with 6 degrees of freedom on the estimates of the coefficients of the wage equation for 1971 and 1973 we find $\chi^2(6)$'s of 125 and 93, respectively,which are highly significant. Thus self-selection is indeed important. Testing only the difference in the sum of the schooling and age coefficients, which is on the order of .01, yields $\chi^2(1)$'s equal to 14.7 and 6.4 in 1971 and 1973, respectively, again indicating a statistically significant difference. It is clear that for this population, self-selection has a substantial effect on the estimated returns to schooling.

We have developed, in this section, a three equation wage determination model allowing for both sample selection and simultaneity, and estimated it in a substantive context, illustrating the seriousness of such biases and the computational feasibility of the suggested estimation methods in a large-sample many-variables context. The main reservation about this model is that it does not take fully intc account the dynamic aspect of the going-to-school decision. The school-vs-work equation is treated as a once and for all decision independent of the time sequence of previous events. But in any period, being in school, being at work, and the accumulated level of schooling all depend on a string of such previous decisions. It is beyond the scope of this paper, however, to try to develop a fully integrated optimizing-over-time schooling and work decision model, though we hope to do so in the future.

VI. SUMMARY

The consequences of missing data and sample selection are analyzed best within the context of the consistency and (asymptotic) efficiency of the estimates. In a large panel of individual respondents such as the NLS Young Men panel data consistency of the estimates is by far the more important concern since the large sample size guarantees relatively accurate estimates. There is no need, however, to use only consistent estimation methods since even for the rather complex models considered in this paper we find that maximum likelihood procedures are readily applied. They do not seem unreasonably more expensive than alternative consistent methods which do not utilize all of the data. Also, they seem to provide more powerful tests of non-randomness than do the consistent methods even for quite large samples. Thus, our first conclusion is that maximum likelihood is the appropriate estimator except perhaps for initial data exploration.

The first substantive problem considered in estimating the returns to schooling for the NLS Young Men is that IQ scores are missing for about one-third of the sample. Missing IQ is not a problem of self-selection; but reasons exist that indicate that it might not be missing at random. We find, however, that IQ is missing almost at random and therefore techniques which would lead to consistant estimates in the presence of non-randomness are probably unnecessary. Moreover, first order missing data type techniques which produce efficient estimates for the random missing data case are unlikely to lead to much improvement in such data. This happens because the "instruments" used in place of the missing data are unlikely to be very highly correlated with the missing data due to the large proportion of randomness in individual data.

The next problem considered is the changing composition of the sample over time which turns out to be the result of two opposing forces: sample attrition and sample accretion, that is, entrance into the labor forces of the fraction of the sample originally enrolled in school. In the earlier period of 1966-1969 we find sample selectivity to be an important factor. By the later period of 1971-1973 when school enrollment had decreased from 62 percent to 10 percent and sample attrition had increased from zero to 18 percent, little effect of sample selectivity could be found. The findings in this section should be viewed, however, as preliminary since we are mixing together disparate reasons for sample selection. A more complete model would separate the different sources of sample selection into schooling, military, and sample attrition decisions.

The most interesting of our models concentrates on a subsample of the NLS Young Men and considers the choice of work versus school. Since self-selection is certainly at work in the decision to continue one's education, non-randomness in the unobserved attributes of those people continuing in school is to be expected. Estimators which do not account for this self-selection would lead to serious inconsistency in the estimates.

Using an expanded model which permits the simultaneous determination of schooling and earnings, we find that self-selection of those remaining in the sample and the correlation of the unobserved attributes across equations results in serious underestimation of the returns to schooling, on the order of 50 percent or more. Nor can these results have arisen by chance since statistical tests indicate significant non-randomness even for very small test sizes. Thus, we conclude that self-selection for additional schooling seems an important factor in estimating returns to schooling emphasizing again the importance of unobserved individual attributes in the wage determination process.

Lastly, we have also presented in some detail the statistical and computational methods used to deal with sample selection problems. These methods should prove useful to other researchers as they confront similar problems.

VII. FOOTNOTES

1. See Afifi and Elashoff (1966) for an earlier survey, Maddala (1977) Chap. 10, Sec. 4, for a survey from the econometric point of view, and Dempster et al. (1977) for the most recent state-of-the-art paper.

2. See Dagenais (1972), Haitovsky (1968), and Hester (1976).

3. See, among others, Amemiya (1973), Hanoch (1976), Hausman and Wise (1977), Hausman and Spence (1977), Heckman (1974, 1976), Maddala (1976), and Tobin (1958).

4. See Griliches (1976) for more detailed description of these data and for related work.

5. This point has been made by Kelejian (1969), among others.

6. This is an asymptotic expression, making no allowance for the ε's, the discrepancies due to the "first stage" β_{1a} and $\hat{\delta}_a$ not being equal to their population values. Hence, it too is an upper bound. It appears to be equivalent to Kelejian's (1969) formula 28, though we have not been able to reproduce his derivations exactly.

7. This differs from the suggestion discussed by Hester (1976) by including also x_2 among the list of instruments. The exclusion of the other x's from his instrumental variable estimator may account for his somewhat strange and biased results.

8. We have not discussed explicitly using y_b to estimate the missing x_{1b}. The full-information maximum-likelihood procedure would do so implicitly. To econometricians using y to estimate missing x values looks suspiciously like an invitation to simultaneity bias. But a complete maximum likelihood procedure which assumes that both y and all the x's are multivariate normal, would use *all* the information in the sample [see, e.g., the description of the E-M algorithm in Dempster et al. (1977)], but it would not use the constructed $\hat{x}_{1b}$ directly in a regression of y on x_1 and x_2. Rather it would use such an $\hat{x}_{1b}$ to fill in the *covariances* in the X'X matrix, where the fact that $\hat{x}_{1b}$ may depend on e (the disturbance in y) does not matter and rely on a more elaborate procedure for getting an estimate of its variance, where it does matter.

9. The original sample consisted of 5,225 individuals. One hundred thirty-one observations (2.5%) were eliminated because of missing scores on the "knowledge of the world of work" test (124) and the rest because of missing schooling information as of 1966.

10. It is similar to the models considered by Amemiya (1973), Hanoch (1976), Hausman and Spence (1977), Heckman (1976), Maddala (1976), and Nelson (1975), among others.

11. This expression is given in Heckman (1976) and Hanoch (1976) among other places. We present the above derivation at

some length because we found their explications rather cryptic.

12. A program using the algorithm described above was written in Fortran IV (double precision) for the IBM 370/168. It required approximately 40 seconds ($14.60) and took 6 iterations to converge from starting values given by OLS (for the IQ equation) and probit (for the sample selection equation). Note that this was a rather large-scale problem with N = 5094 (S = 3324) and 12 and 8 independent variables in the IQ and sample selection equations, respectively. For comparison, estimation of the probit equation, calculation of the inverse Mills ratio, and OLS estimation of the wage equation including this ratio, cost $10.60.

13. This differs from the "ρ is significant" finding in Table 2, because of differences in asymptotic approximations to the information matrix.

14. We shall deal with the out-of-school and non-interview selectivity biases in subsequent sections of this paper.

15. The estimated "ability bias" is somewhat larger in the combined sample: .012 versus .009 in the "good IQ" subsample alone. This is due, in part, to the schooling coefficient being somewhat lower in sample B, as already noted above, and to the fact that the observed IQ scores are likely to be subject to significant measurement error. In such a context, using predicted IQ may be better than using actual. Thus, both the coefficient of IQ and the fit go up when we substitute T_3 for IQ in the sample as a whole. This contradicts the model maintained in the text. See Griliches (1977) for further discussion of this topic and for estimates which allow for both errors in IQ and the endogeneity of schooling.

16. To treat several causes of not being in the sample differentially is a bit beyond the state of the art at the moment. One could, of course, compute a multinomial probit or logit analysis and get differential coefficients for the various variables. The difficulty, however, is in construction of the appropriate Mills ratio variables which would now be the result of truncation on several correlated margins. See Hausman and Wise (1977b) for an initial attack on a related problem.

17. Discussion of such models arose originally in the context of studies of female labor force participation. See Gronau (1974) and Heckman (1974), among others.

18. See Amemiya (1974), Hausman and Wise (1977), and Nelson and Olson (1977) for attacks on similar problems.

19. See Hausman (1975) for a more detailed exposition of this point.

20. Least squares estimates of γ_2, γ_3, β, σ_2^2, σ_3^2 and probit estimates of δ were used as starting values with ρ_{12}, ρ_{13}, ρ_{23} set to zero initially. For one of the samples discussed below, s = 2057 and n = 2419, the program converged in 18 iterations, with an approximate cost of $56.00 on the Harvard-MIT IBM 370/168 computer.

21. In an earlier version of this paper we restrict the sample to the subset of high school graduates, focusing on the going-to-college margin and truncating S from below at

$S \geq 12$. The results were similar to those reported below and perhaps even more striking. But since truncating S from below introduces additional statistical problems we have reverted in this version to the analysis of the complete range of S.

22. Originally we estimated our model only for 1973 but given the rather striking results we decided to check them against the 1971 data to convince ourselves that they were not a fluke.

23. For the OLS results the schooling and age coefficients add up to .019 + .042 = .061 for 1973 while the schooling coefficient holding experience rather than age constant is .050. Both are comparable to the results reported in the previous section.

24. The IQ coefficient is probably underestimated, since no allowance is made for errors in its measurement. See Griliches (1977) for additional discussion of these issues and related results.

25. If one respecifies the model in terms of $e_1 = u - e_2$ than the implied (for 1973) $\rho_{u2} = .84$ and $\rho_{u3} = -.46$, with $\sigma_u = 1.72$.

VIII. REFERENCES

Afifi, A. A. and R. M. E. Elashoff, 1966. "Missing Observations in Multivariate Statistics--I. Review of the Literature." Journal of the American Statistical Association 61(315):595-605.

Amemiya, T., 1973. "Regression Analysis When the Dependent Variable Is Truncated Normal." Econometrica 41(5):997-1016.

__________ 1974. "Multivariate Regression and Simultaneous Equation Models When the Dependent Variables Are Truncated Normal." Econometrica 42(6):999-1012.

Berndt, E. B., B. Hall, R. Hall, and J. A. Hausman, 1974. "Estimation and Inference in Nonlinear Structural Models." Annals of Economics and Social Measurement 3:653-65.

Dagenais, M. G., 1972. "Asymptotic Behavicr and Small Sample Performance: Experiments on Regression Parameter Estimation with Incomplete Observations." European Economic Review 3:389-98.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977. mum Likelihood from Incomplete Data Via the EM Algorithm." Journal of the Royal Statistical Society, Series B, 39(1):1-22.

Griliches, Z., 1976. "Wages of Very Young Men." Political Economy 84(4, pt. 2):S69-85.

__________ 1977. "Estimating the Returns to Schoo. Econometric Problems." Econometrica 45(1):1-.

Gronau, R., 1974. "Wage Comparisons--A Selectivity Bias. Journal of Political Economy 82:1119-43.

Haitovsky, Y., 1968. "Missing Data in Regression Analysis. Journal of the Royal Statistical Society, Series B, 30(1):67-82.

Hanoch, G., 1976. "A Multivariate Model of Labor Supply: Methodology for Estimation." Rand Corporation.

Hausman, J. A., 1975. "An Instrumental Variable Approach to Full Information Estimators." Econometrica 43(4):727-39

__________ 1976. "Specification Tests in Econometrics." Econometrica, in press.

Hausman, J. A. and A. M. Spence, 1977. "Non-Random Missinq Data." Massachusetts Institute of Technology, Working Paper No. 185.

Hausman, J. A. and D. A. Wise, 1977a. "Social Experimentation, Truncated Distribution, and Efficient Estimation." Econometrica 45(4):919-38.

__________ 1977b. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences." Econometrica, in press.

Heckman, J. D., 1974. "Shadow Prices, Market Wages, and Labor Supply." Econometrica 42:679-94.

__________ 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." Annals of Economics and Social Measurement 5(4):475-92.

Hester, D. D., 1976. "Estimation from Incomplete Samples." University of Wisconsin-Madison, SSRI Workshop Series Paper 7608.

Kelejian, H. H., 1969. "Missing Observations in Multivariate Regression: Efficiency of a First-Order Mehtod." Journal of the American Statistical Association 64(4): 1609-616.

Maddala, G. S., 1976. "Self-Selectivity Problems in Econometric Models." University of Florida Working Papers in Econometrics, No. 76-77-06.

__________ 1977. Econometrics. New York: McGraw Hill.

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

Nelson, F. and L. Olson, 1977. "Specification and Estimation of a Simultaneous-Equation Model with Limited Dependent Variables." California Institute of Technology, Social Sciences Working Paper, No. 149.

Tobin, J., 1958. "Estimation of Relationship for Limited Dependent Variables." Econometrica 26(1):29-36.