1. Introduction

We consider the estimation of the regression coefficient $\beta_{12}$ in a linear regression:

$$E(X_{1i} | X_{2i}) = \mu_1 + \beta_{12} X_{2i},$$

on the basis of a random sample of observations, some of which are missing due to non-response. We assume that response depends, in a way to be specified, on values of a third variable, $X_3$, which is observable for the whole sample and which is related to the variables $X_1$ and $X_2$.

Two alternative approaches to this problem have been considered. The econometric approach, summarized by Heckman (1979), assumes response to be determined by an unobservable variable, $X_4$, which is the dependent variable in a linear regression relationship with $X_3$. Heckman (1976) proposes a two-stage (probit and OLS) estimator, which is asymptotically unbiased, under certain assumptions, within a conditional framework of inference (given the values of $X_2$ and $X_3$), and for which asymptotic expressions of variance are given by Lee, Maddala and Trost (1980).

The sampling theory approach, proposed by Nathan and Holt (1980) and by Holt and Smith (1979), to deal with the effect of sample design on regression analysis can be applied to this case. This is done by considering the response as obtained by sub-sampling from the main sample, via a non-informative design, with the sample distribution determined, in an unknown way, by the values of $X_3$. An overall MLE, under assumptions of multivariate normality, which is asymptotically unbiased, is proposed by Holt and Smith (1979) and some of its properties are shown to hold under less stringent assumptions by Nathan and Holt (1980).

In order to compare the estimators proposed, an overall model is set up, within which both approaches can be imbedded as special cases (section 2). The implications of each of the approaches with respect to the assumptions are examined in sections 3 and 4 and expressions for the conditional expectations and variances of the estimators are given in section 5. In section 6 an estimator is proposed and in section 7 the results of some simulation comparisons between the performances of the estimators under the full model assumptions and under certain relaxations of the assumptions are presented.

2. Overall Model

We formulate the model for univariate variables, for simplicity, but the extension to the multivariate case is immediate. We assume a finite sample of $N$ values of four variables with a multivariate normal distribution, i.e.:

$$(X_{1i}, X_{2i}, X_{3i}, X_{4i}) \sim N(\mu; \Sigma)$$

(i=1, ..., N)

where:

$$\mu = (\mu_1, \mu_2, \mu_3, \mu_4)$$

and:

$$\Sigma_{jk} = \sigma_{jk}; \sigma_{j3} = \sigma_{13}^2,$$

define $\beta_{ab}$, $\beta_{abc}$, $\sigma_{ab}$, $\sigma_{ab}^2$, $\sigma_{a,b,c}$, $\sigma_{abc}$, $\sigma_{ab,cd}$ as usual. The conditional distribution of $X_{1i}$, given $X_{2i}$:

$$X_{1i} | X_{2i} \sim N(\mu_1 + \beta_{12} (X_{2i} - \mu_2); \sigma_{12}^2),$$

defines the regression relationship of interest and $\beta_{12}$ is the parameter of interest. Observations on $X_{1i}$ are available only for a sub-sample of size $n$, defined by the indicator variable, $d_i$, which is determined by the variable $X_{4i}$:

$$d_i = \begin{cases} 1; & X_{4i} \geq 0 \\ 0; & \text{otherwise.} \end{cases}$$

Observations on $X_{4i}$ are not available for any unit (except via $d_i$), while observations on $X_{3i}$ are available for all units of the main sample (i=1, ..., $N$). Observations on $X_{2i}$ may be available for all $i=1, ..., N$, but those for which $d_i=0$ are not in general useful for estimating $\beta_{12}$.

3. The Sampling Theory Approach

The sampling theory model considers the marginal distribution of $(X_{1i}, X_{2i}, X_{3i})$ and derives the overall MLE of $\beta_{12}$ under the assumption of non-informativeness of the sample design, as defined by:

$$(X_{1i}, X_{2i}) \mid (X_{3i}, d_i=1) \sim (X_{1i}, X_{2i}) \mid X_{3i}. \quad (3.1)$$

This is equivalent to the condition:

$$\sigma_{24,3}^2 = \sigma_{14,3}^2 = 0 \quad (3.2)$$

The overall MLE is obtained by DeMets and Halperin (1977) as:

$$\beta_{12}^2 = \frac{s_2 + \frac{s_{23}}{s_3} \left( s_3 - 1 \right)}{s_2 + \frac{s_{23}}{s_3} \left( s_3 - 1 \right)}, \quad (3.3)$$

where:

$$s_{ab} = \frac{1}{n} \sum_{i=1}^{N} d_i (X_{a1} - \bar{X}_a) (X_{b1} - \bar{X}_b)$$

and:

$$s_a^2 = \frac{1}{n} \sum_{i=1}^{N} d_i (X_{a1} - \bar{X}_a)^2;$$

$$\bar{X}_a = \frac{1}{n} \sum_{i=1}^{N} d_i X_{a1};$$

$$\bar{X}^2 = \frac{1}{n} \sum_{i=1}^{N} (X_{3i} - \bar{X}_3)^2;$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_{3i}$$

(a,b = 1,2,3).
4. The Econometric Approach

Heckman (1976) considers the joint conditional distribution of \((X_{li}, X_{4i})\), given \((X_{2i}, X_{3i})\), induced by (2.1), in the form of the regression relationships:

\[ X_{li} = \mu_i + \beta_{12}(X_{2i} - \mu_2) + u_{li}, \]
\[ X_{4i} = \mu_4 + \beta_{43}(X_{3i} - \mu_3) + u_{4i}, \]

(4.1)

where:

\( (u_{li}, u_{4i}) \mid (X_{2i}, X_{3i}) \sim N(0, \Sigma) \)

(4.2)

\[ \Sigma = \begin{pmatrix} \sigma_{12,2}^2 & \sigma^{14.23} \\ \sigma^{14.23} & \sigma_{43}^2 \end{pmatrix} . \]

The derivation of the estimator is made under the implied assumptions:

\[ (X_{li}, X_{4i}) \mid (X_{2i}, X_{3i}) \sim X_{li} \mid X_{2i}, (X_{4i} \mid X_{3i}) \sim X_{4i} \mid X_{3i} , \]

(4.3)

which is equivalent to the conditions:

\[ \sigma_{13.2} = 0. \]

(4.4)

Note that this implies that the parameters of (4.2) become:

\[ \sigma_{13.2}^2 = \sigma_{14.2}^2 + \sigma_{43}^2 ; \]
\[ \sigma_{14.2}^2 = \sigma_{14.3}^2 . \]

(4.5)

Heckman (1976) obtains an estimator of \( \beta_{12} \) in two stages. At the first stage, probit analysis is used to estimate the parameters of the conditional probability that the unit is sampled, given \( X_{3i} \):

\[ P(d_i = 1 \mid X_{3i}) = \phi(-Z_i), \]

where \( Z_i = \frac{[\mu_4 + \beta_{43}(X_{3i} - \mu_3)]}{\sigma_{43}} \) and \( \phi \) is the standard normal d.f. Let \( \hat{Z}_i \) be the estimate of \( Z_i \) obtained at this stage. At the second stage, \( \hat{\beta}_{12} \) is estimated by OLS from the regression relationship:

\[ E(X_{li} \mid X_{2i}, d_i = 1) = \mu_i + \hat{\beta}_{12}(X_{2i} - \mu_2) + \frac{\sigma_{14.3}}{\sigma_{43}} \hat{Z}_i, \]

(4.7)

where \( \hat{\beta}_{12} = \frac{\phi(\hat{Z}_i)}{\phi(-\hat{Z}_i)} \) is estimated at the first stage (and \( \phi \) is the standard normal p.d.f.).

Denoting sample variances and covariances involving \( \hat{Z}_i \) by

\[ s^2 = \frac{1}{n} \sum_{i=1}^{N} d_i (\hat{Z}_i - \bar{Z})^2, \]
\[ s_{al} = \frac{1}{n} \sum_{i=1}^{N} d_i (\hat{Z}_i - \bar{Z})(X_{ai} - \bar{X}_a) \quad (a = 1, 2, 3), \]

(4.8)

and the conditional sample variances and covariances by:

\[ s_{ab,c} = s_{ab} - s_{ac}/s_c^2 ; \]

the resulting estimator can be written as:

\[ \hat{\beta}_{12} = \frac{s_{12} - s_{14.3}/s_4}{s_{14.3}/s_4} \]

(4.9)

5. Conditional Expectations and Variances of the Estimators

Under the general model of section 2, the conditional expectation of \( \hat{\beta}_{12} \), given the values of \( X_2 = (X_{21}, \ldots, X_{2N}) \), \( X_3 = (X_{31}, \ldots, X_{3N}) \) and \( d = (d_1, \ldots, d_N) \), can be derived as:

\[ E(\hat{\beta}_{12} \mid X_2, X_3, d) = \beta_{12} + \frac{\beta_{13.2} s_{23}^2 / s_3^2 + (s_{23}/s_3)(\sigma_{23}/s_3) s_{31} + s_{21} \sigma_{23}^2}{s_{23}/s_3 + (s_{23}/s_3)(\sigma_{23}/s_3)} \]

(5.1)

where:

\[ \lambda = \frac{(\sigma_{14.2}^2/\sigma_{43}^2) \phi(\hat{\lambda}_i) - \phi(-\hat{\lambda}_i)}{\phi(\hat{\lambda}_i)}, \]

(5.2)

\[ \lambda = \frac{[\mu_4 + \beta_{43}(X_{3i} - \mu_3)]}{\sigma_{43}} \]

(5.3)

and \( \sigma_{alb} \), \( \sigma_{alb} \), \( (a, b = 1, 2, 3) \) are defined as in (4.8) with \( \hat{\lambda}_i \) replaced by \( \hat{\lambda}_i \). Similarly the conditional expectation of \( \hat{\beta}_{12} \) can be derived as:

\[ E(\hat{\beta}_{12} \mid X_2, X_3, d) = \beta_{12} + \frac{\beta_{13.2} s_{23}^2 / s_3^2 + (s_{23}/s_3)(\sigma_{23}/s_3) s_{31} + s_{21} \sigma_{23}^2}{s_{23}/s_3 + (s_{23}/s_3)(\sigma_{23}/s_3)} \]

(5.4)

\[ \lambda = \frac{[\mu_4 + \beta_{43}(X_{3i} - \mu_3)]}{\sigma_{43}} \]

(5.5)

The expectations under special conditions are as follows:

1) If \( s_{a}^2 = s_{a}^2 \):

\[ E(\hat{\beta}_{12} \mid X_2, X_3, d) = \beta_{12} + \frac{\beta_{13.2} s_{23}^2 / s_3^2 + (s_{23}/s_3)(\sigma_{23}/s_3) s_{31} + s_{21} \sigma_{23}^2}{s_{23}/s_3 + (s_{23}/s_3)(\sigma_{23}/s_3)} \]

(5.6)

2) If the conditions of the sampling theory model (3.2) hold:

\[ E(\hat{\beta}_{12} \mid X_2, X_3, d) = \beta_{12} + \frac{\beta_{13.2} s_{23}^2 / s_3^2 + (s_{23}/s_3)(\sigma_{23}/s_3) s_{31} + s_{21} \sigma_{23}^2}{s_{23}/s_3 + (s_{23}/s_3)(\sigma_{23}/s_3)} \]

(5.7)

3) If the conditions of the econometric model (4.4) hold:

\[ E(\hat{\beta}_{12} \mid X_2, X_3, d) = \beta_{12} + \frac{\beta_{13.2} s_{23}^2 / s_3^2 + (s_{23}/s_3)(\sigma_{23}/s_3) s_{31} + s_{21} \sigma_{23}^2}{s_{23}/s_3 + (s_{23}/s_3)(\sigma_{23}/s_3)} \]

(5.8)

Thus \( \hat{\beta}_{12} \) is conditionally unbiased if and only if both the conditions of the sampling theory model, (3.2), and those of the econometric model (4.4), hold.
The conditional variance of $\hat{\beta}_{12}$ can be derived, under the general model, as:

$$V(\hat{\beta}_{12}|X_2,X_3,d) = \sigma^2_{12.3} \{s_{2.3}^2 + (s_{2.3}^2/s_3)(\hat{\sigma}^2_3/s_3^2) + \sum_{i=1}^{N} d_i \lambda (\sigma^2_{i-\lambda} - \lambda) \}$$

where:

$$\sigma^2_{12.3} = \sigma^2_1 + \sigma^2_2 - \sigma^2_3$$

and $s_{2.3}^2 = (s_2/s_3)(\hat{\sigma}^2_3/s_3^2) + (s_2^2/s_3^2)(\hat{\sigma}^2_3/s_3^2)$.

(5.8)

while that of $\hat{\beta}^*_1$ can be derived as:

$$V(\hat{\beta}^*_1|X_2,X_3,d) = \sigma^2_{12.3} \{s_{2.3}^2 + (s_{2.3}^2/s_3)(\hat{\sigma}^2_3/s_3^2) \}$$

where:

$$\sigma^2_{12.3} = \sigma^2_1 + \sigma^2_2 - \sigma^2_3$$

and $s_{2.3}^2 = (s_2/s_3)(\hat{\sigma}^2_3/s_3^2) + (s_2^2/s_3^2)(\hat{\sigma}^2_3/s_3^2)$.

(5.9)

For special conditions we obtain:

1) If $\sigma^2_2 = s_2^2$:

$$V(\hat{\beta}_{12}|X_2,X_3,d) = \sigma^2_{12.3} \{s_{2.3}^2 + (s_{2.3}^2/s_3)(\hat{\sigma}^2_3/s_3^2) \}$$

(5.10)

while (5.9) does not change.

2) If the conditions of the sampling theory model (3.2) hold:

$$V(\hat{\beta}_{12}|X_2,X_3,d) = \sigma^2_{12.3} \{s_{2.3}^2 + (s_{2.3}^2/s_3)(\hat{\sigma}^2_3/s_3^2) \}$$

(5.11)

$$V(\hat{\beta}^*_1|X_2,X_3,d) = \frac{\sigma^2_{12.3}}{s_{2.3}^2} \sigma^2_{12.3}$$

(5.12)

3) If the conditions of the econometric model (4.4) hold:

The only changes in (5.8) and (5.9) are in the definitions of $\sigma^2_{12.3}$ and $\lambda_1$. 

6. An Estimator Based on the Overall Model

Consider the joint conditional distribution of $(X_{1i},X_{4i})$, given $(X_{2i},X_{3i})$, under the overall model (1.2). This distribution is completely specified as follows:

$$X_{1i} = \nu_1 + \beta_{12.3}(X_{2i} - \mu_2) + \beta_{13.2}(X_{3i} - \mu_3) + \nu_{1i}$$

$$X_{4i} = \nu_4 + \beta_{42.3}(X_{2i} - \mu_2) + \beta_{43.2}(X_{3i} - \mu_3) + \nu_{4i}$$

(6.1)

where:

$$\nu_{1i} \sim \nu(\mu_1,\sigma^2_{11})$$

$$\nu_{4i} \sim \nu(\mu_4,\sigma^2_{44})$$

$$X_{2i} \sim \nu(\mu_2,\sigma^2_{22})$$

$$X_{3i} \sim \nu(\mu_3,\sigma^2_{33})$$

$$\nu_{1i} \perp \nu_{4i} \perp X_{2i} \perp X_{3i}$$

$$\nu_{1i} \perp X_{2i} \perp \nu_{4i} \perp X_{3i}$$

and $\nu^*$ is defined by (4.2).

Note that this reduces to (4.1)-(4.2) under the assumptions (4.3).

An estimator based on the more general model above can be derived in an analogous manner to that used in the econometric approach, described in section 4 as follows:

a) Use probit analysis to estimate the parameters of the conditional probability of a unit's being observed, given both $X_{24i}$ and $X_{34i}$:

$$P(d_i=1|X_{24i},X_{34i}) = \phi(z_i)$$

(6.3)

where $z_i$ is defined by (5.1). Note that, under this model, the probit estimator of $z_i$ is just $\hat{z}_i$ - the probit estimator of $Z_i$ obtained under the simplifying assumptions.

b) The parameters of the following regression relationship:

$$E[X_{1i}|X_{24i},X_{34i},d_i=1] = \nu_1 + \beta_{12.3}(X_{2i} - \mu_2) + \beta_{13.2}(X_{3i} - \mu_3) + \nu_{1i}$$

(6.4)

are estimated by OLS after substituting $\hat{z}_i$ instead of $Z_i$. Let $\hat{\beta}_{12.3}$ and $\hat{\beta}_{13.2}$ be the estimators of $\beta_{12.3}$ and $\beta_{13.2}$ thus obtained.

c) Let $\hat{\beta}_{32}$ be the OLS estimator of $\beta_{32}$ obtained from the observations $(X_{24i},X_{34i})$; i=1,...,N, i.e.

$$\hat{\beta}_{32} = \frac{\Sigma (X_{24i} - \overline{X}_{24})(X_{34i} - \overline{X}_{34})}{\Sigma (X_{24i} - \overline{X}_{24})^2}$$

(6.5)

Then the estimator of $\hat{\beta}_{12} - \hat{\beta}_{12}^0$ is obtained from the relationship:

$$\hat{\beta}_{12} = \hat{\beta}_{12.3} + \hat{\beta}_{32} \hat{\beta}_{13.2}$$

(6.6)

as:

$$\hat{\beta}_{12} = \hat{\beta}_{12.3} + \hat{\beta}_{32} \hat{\beta}_{13.2}$$

(6.7)

The conditional expectation of $\hat{\beta}_{12}^0$, given $X_{24},X_{34}$ and $d$ can be shown to be:

$$E(\hat{\beta}_{12}|X_{24},X_{34},d) = \beta_{12.3} + \hat{\beta}_{32} \beta_{13.2}$$

$$+ \frac{\Sigma (X_{24i} - \overline{X}_{24})(X_{34i} - \overline{X}_{34})}{\Sigma (X_{24i} - \overline{X}_{24})^2}$$

(6.8)

Since $\hat{\beta}_{32}$ is a consistent estimator of $\beta_{32}$ and since the last two terms of (6.8) vanish as $N \rightarrow \infty$ (because $\overline{X}_{24}$ is a consistent estimator of $X_{24}$), $\hat{\beta}_{12}^0$ is a consistent estimator of $\hat{\beta}_{12}$. This can also
be seen from Heckman's (1979) result, if both $X_2$ and $X_3$ are used as explanatory variables in both equations.

The conditional variance of $\hat{\beta}_{12}$, given $X_2, X_3$ and $d$ can be shown to be:

\[
V(\hat{\beta}_{12} \mid X_2, X_3, d) = \frac{s_{21}^2 + s_{31}^2}{2s_{23}^4} \{s_{23}^2(1 + s_{32}^2 s_{21}^2 / s_{32}^2)\}^2 \\
+ \frac{g^2}{n} \sum_{i=1}^{N} d_i \lambda_i (\hat{\lambda}_i - \lambda_i) \\
\left(1 - \frac{s_{32}^2 s_{21}^2}{s_{31}^2} (X_4_i - X_4) \right) \\
- \frac{s_{23}^2 s_{32}^2}{s_{31}^2} (X_3_i - X_3) \\
\left(\hat{\lambda}_i - \lambda_i\right) \}^2 \} (6.9)
\]

7. Comparisons of the Estimators

The proposed estimators can be computed without the assumptions implied by the models under which they are derived. Their conditional biases and variances can be computed both under the general model assumptions and under sub-models. However, the results cannot be compared analytically due to the fact that analytic expressions for the probit estimators $\hat{\lambda}_i$ are not available. Thus simulation comparisons have to be used.

An initial small simulation study was carried out with respect to the estimators $\hat{\beta}_{12}$ and $\hat{\beta}^*_{12}$. Four sets of population parameters, specified in Table 1 by their correlation matrices, were used. Under the first neither the econometric model assumptions (4.4) nor the sampling model assumptions (3.2) hold. The second set represents conditions close to those of the econometric model: $|r_{1,2}|, |r_{2,4}| < 0.01$, while for the remaining two sets the exact econometric model and the exact sampling model assumptions hold, respectively.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Correlation Matrix</th>
</tr>
</thead>
</table>
| 1. General | $0.762 \ 0.626 \ 0.418$
| & $0.590 \ 0.288$
| & $0.139$
| & $0.177$ $0.100$ $0.143$
| 2. Approximate | $0.762 \ 0.626 \ 0.418$
| E-conomic Model Conditions | $0.704 \ 0.274$
| & $0.139$
| & $0.009$ $0.005$ $0.143$
| 3. Exact | $0.762 \ 0.626 \ 0.418$
| E-conomic Model Conditions | $0.822 \ 0.262$
| & $0.139$
| & $s_{24.3} = 0$ $s_{14.3} = 0.143$
| 4. Exact | $0.762 \ 0.626 \ 0.200$
| Sampling Model Conditions | $0.590 \ 0.188$
| & $0.139$
| & $0.177$ $0.100$ $0.414$

Table 2: Simulation Study Results

| Parameter Set | Approx. Exact Exact General Econometric Econometric Sample |
|---------------|-----------------|-----------------|
| 1. General | $0.036 \ 0.041 \ 0.046 \ -0.002$
| Bias | $(0.0025)$ $(0.0011)$ $(0.0003)$ $(0.0028)$
| $\hat{\beta}^*_{12}$ | $0.226 \ 0.176 \ -0.004 \ 0.206$
| $(0.0004)$ $(0.1115)$ $(0.0001)$ $(0.0008)$ |
| Relative Error | $0.024 \ 0.024 \ 0.024 \ 0.023$
| $(0.0002)$ $(0.0002)$ $(0.0001)$ $(0.0002)$ |
| Root M.S.E. | $0.048 \ 0.052 \ 0.024$
| $(0.0022)$ $(0.0011)$ $(0.0003)$ $(0.0009)$ |

The results, though limited in their scope, give an indication that the sample model estimator, $\hat{\beta}^*_{12}$, may be more robust than the econometric model estimator, $\hat{\beta}_{12}$. The latter has high mean square error (due to high bias) when the econometric model conditions do not hold, while $\hat{\beta}^*_{12}$ maintains a low level of mean square error even when the conditions for its consistency do not hold. Since $\hat{\beta}^*_{12}$ is an explicit single-stage estimator, this might make it preferable to $\hat{\beta}_{12}$. However further studies on the performances of the estimators for different departures from model assumptions have to be carried out.

8. Acknowledgements

The author is grateful to Sheila Ben-Tuvya and to Esther Sharon for programming and running the simulations of section 7. Research for this paper was partially supported by research agreements between the Israel Central Bureau of Statistics and the U.S.-Israel Binational Agricultural Research and Development Fund and the Israel National Council of Research and Development.
9. References


