The Economic and Social Review, Vol. 27, No. 1, October, 1995, pp. 33-42

Estimating Equations with Information Loss on at Least One Dependent Variable

DENIS CONNIFFE*
The Economic and Social Research Institute

Abstract: Efficient, or joint, estimation of a pair of linear equations with the same explanatory variables reduces to separate estimation of each equation. This is no longer the case if information has been lost on at least one dependent variable; for example, if all that is recorded is whether some value is exceeded or not, or if information has been lost completely. This paper considers a class of problems, for which the efficient estimates can be deduced by simple intuitive arguments and which occurs frequently in practice. Some members of the class have already appeared in the literature, but have not been related, while others are new.

I INTRODUCTION

It is well known that joint (or efficient) estimation of two (or more) equations with the same explanatory variables, and measured on the same n observations having the form

$$\begin{aligned} y_{1i} &= x_i' b_1 + u_{1i} \\ y_{2i} &= x_i' b_2 + u_{2i} \end{aligned}, \qquad (1)$$

where u_{1i} and u_{2i} will be assumed bivariate normal, just reduces to single equation estimation of each (Anderson, 1958; Zellner, 1962). Of course, it is demonstrated in texts that if there are cross equation restrictions on

Paper presented at the Ninth Annual Conference of the Irish Economic Association.

*Thanks for helpful comments are due to Andrew Chesher (Bristol), Patrick Honohan (ESRI), and to conference participants.

coefficients, or if explanatory variables differ between equations, this may no longer be the case. But such cases are actually just further manipulation of the ordinary least squares solutions, appropriate only if strong assumptions hold. For example, if

$$b_1' = (b_{11}', b_{12}'),\ b_2' = (b_{21}',\ b_{22}') \text{ and } x_i' = (x_{1i}',\ x_{2i}')$$

(1) becomes

$$\begin{aligned} y_{1i} &= x_{1i}'\, b_{11} + x_{2i}'\, b_{12} + u_{1i} \\ y_{2i} &= x_{1i}'\, b_{21} + x_{2i}'\, b_{22} + u_{2i} \end{aligned} \qquad (2)$$

These still have the same explanatory variables, so the usual ordinary least squares formulae apply for coefficients (call them $b_{ij}*$) as do the usual variances or covariances (call them v_{ij}). Now a symmetry constraint, $b_{12} = b_{21}$, would just mean that

$$\begin{pmatrix} b_{11}* \\ b_{12}* \\ b_{21}* \\ b_{22}* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{12} \\ b_{22} \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \qquad (3)$$

or

$$B^* = HB + W,$$

where B denotes the reduced coefficient vector $B' = (b_{11}, b_{12}, b_{22})$. The solution of (3) is a simple GLS estimator itself.

$$\hat{B}_H = [H'V^{-1}H]^{-1}H'V^{-1}B^*$$

where V is the variance covariance matrix of the B*.

Again, the model (2) would become seemingly unrelated regressions if it could be assumed that $b_{12} = b_{21} = 0$ and then

$$\begin{pmatrix} b_{11}* \\ b_{12}* \\ b_{21}* \\ b_{22}* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{22} \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \qquad (4)$$

or $$B^* = K B_k + W$$

giving $$\hat{B}_k = [K'V^{-1}K]^{-1}K'V^{-1}B^*. \quad (5)$$

Thus, given that the appropriate estimators for the common explanatory variable case are known, it is quite easy to proceed to cases that just reduce the number of coefficients through assumptions. Of course, the assumptions should be tested, which is another reason for starting from the common explanatory variable case.

However, efficient estimation of two equations with common explanatory variables does not always reduce to single equation estimators. This paper is concerned with one class of situations, where simple intuitive arguments often suggest the appropriate estimators, but which occur frequently in practice. As a starting point, suppose the situation represented by (1) did once hold, but information was lost on one or both of y_1 and y_2. For example, all measures of y_2 might be lost except the signs. Then there is just a binary variable z and the problem becomes that of efficient joint estimation of a linear equation for y_1 and a probit model for z. This was solved by Chesher (1984). Or, all information might be lost on a sub-set n–r of the y_2 values, the remaining r being fully observed. This is the two linear equations with unequal numbers of observations problem. The appropriate estimators were given in Conniffe (1985). But, as will be seen, there are many other possibilities.

Before proceeding to examine them, it is essential to appreciate why joint estimation of (1) does reduce to single equation estimation. Least squares (or maximum likelihood, given normality) estimation of a single linear equation, with dependent variable y and explanatory variables x_1 and x_2, lead to minimisation of

$$\Sigma(y_i - x'_{1i}c_{1.2} - x'_{2i}c_{2.1})^2$$

and so to the conditions

$$\Sigma(x_{ji}(y_i - x'_{1i}c^*_{1.2} - x'_{2i}c^*_{2.1}) = 0 \quad j = 1,2.$$

But these imply that

$$c_1^* = c_{1.2}^* + d_2^* c_{2.1}^*$$

and

$$c_2^* = c_{2.1}^* + d_1^* c_{1.2}^* \quad (6)$$

where c_1^* and c_2^* are the vectors of simple regression coefficients of y on x_1 and x_2 respectively and d_2^* and d_1^* are the matrices of simple regression coefficients of x_1 on x_2 and of x_2 on x_1.

Returning to the pair of Equations (1), bivariate normality of u_1 and u_2 implies that y_{2i}, conditionally on y_{1i}, is normal with mean

$$x_i'b_2 + \delta(y_{1i} - x_i'b_1)$$

or

$$x_i'(b_2 - \delta b_i) + \delta_{1i}$$

where $\delta = \sigma_{12} / \sigma_{11}$, and with variance $\sigma_{22}(1-\rho^2)$. So in a regression of y_2 on x *and* y_1 the coefficient of x, $b^*_{x.y1}$, estimates $b_2 - \delta b_1$ and the coefficient on y_1, $b_{y1.x}$, estimates δ. So a "new" estimator of b_2 is

$$\hat{b}_2 = b^*_{x.y1} + (\delta b_1) = b^*_{x.y1} + b_1^* b^*_{y1.x} \tag{7}$$

where b_1^* is the estimator of b_1 from the first equation: the "simple" coefficient of y_1 on x. But from (6), the right hand side of (7) is just b_2^*, so $\hat{b}_2 = b_2^*$.

It is true that a similar result can hold for non-linear equations (Gallant, 1975), but only if the non-linear functional form is the same for both equations and the disturbance terms are additive normal. But these are very restrictive conditions.

II CASES IN THE LITERATURE

Chesher (1984) considered joint estimates of linear and probit models. This could occur, for example, if using household expenditure data to model consumption of two commodities such as food and a major consumer durable. Most households will not purchase (or replace) the major durable at all during the survey period, so a (0,1) variable will result. An underlying y_2 variable can be hypothesised, however, perhaps the service derived from the durable. Since it is not observed, σ_{22} can be taken as unity and the occurrence of z = 1 as corresponding to $y_2 \geq 0$. Another way of viewing this is that it is $b_2 / \sqrt{\sigma_{22}}$ that is identifiable. The single equation estimates of b_1 and b_2 are obtained by ordinary least squares and probit analysis respectively. Denote the OLS of b_1 by b_1^* and the single equation probit estimator by b_2^p. The latter is the estimator obtained by maximising the likelihood

$$\Pi\left\{1 - F(-x_i'b_2)\right\}^{z_i} \left\{F(-x_i'b_2)\right\}^{1-z_i},$$

where F is the cumulative distribution function of the standard normal, so that $F(-x_i'b_2)$ is the probability that $z_i = 0$.

When considering efficient estimators for b_1 and b_2, the idea that information lost on a system cannot lead to improved precision will be employed. If y_2, rather than z, was available the situation would be (1) and b_1^* would be the efficient estimator. Therefore, it must still be the efficient estimator when y_2 is not available. However, we can perhaps do better than b_2^p. Since y_{2i}, conditionally on y_{1i}, has mean $x_i'(b_2 - \delta b_1) + \delta y_{1i}$ and variance $1-\rho^2$ with $\rho^2 = \sigma_{12}^2 / \sigma_{11}$, a probit analysis employing x and y_1 can obtain $b_{x.y1}^p$ and $b_{y1.x}^p$ that estimate

$$\frac{b_2 - \delta b_1}{(1-\rho^2)^{\frac{1}{2}}} \text{ and } \frac{\delta}{(1-\rho^2)^{\frac{1}{2}}} \text{ respectively.}$$

Substituting b_1^* for b_1 and σ_{11}^*, the estimator obtained from the error mean square of the ordinary regression analysis for y_1, for σ_{11}, gives

$$\hat{b}_2 = \frac{b_{x.y1}^p + b_1^* b_{y1.x}^p}{\left\{1 + \sigma_{11}^* (b_{y1.x}^p)^2\right\}^{\frac{1}{2}}}. \tag{8}$$

As Chesher remarked, (8) can be calculated given any statistical or econometric software package that contains a probit analysis routine. So can an estimate of its asymptotic variance, $V(b_2)$. The corresponding estimator for $\delta = \sigma_{12} / \sigma_{11}$ is

$$\hat{\delta} = \frac{b_{y1.x}^p}{\left\{1 + \sigma_{11}^* (b_{y1.x}^p)^2\right\}^{\frac{1}{2}}}. \tag{9}$$

The case where a sub-set of y_2's has been lost completely, so that only r observations have measurements on both y_1 and y_2 and another n–r have measurements on y_1, is also common in practice. It can arise with survey data when answering one question is more time consuming, difficult or perhaps unpleasant, than answering another. But it can also occur with time series data. The author's most recent encounter with the situation arose from looking at Irish data on smoking. Data on total tobacco consumed are available from excise sources for a long time period. Data on the number of smokers exist only since regular surveys commenced in 1972. But it could be interesting to model both total tobacco consumption (y_1) and number of smokers (y_2) in terms of x variables like price, national income, population etc.

The same logic as before implies that $\hat{b}_1 = b_1^*$ where b_1^* is the ordinary least squares estimate based on all n observations. Again, the conditional distribution of y_2 given y_1 implies that $b^{r*}_{x.y1}$ estimates $b_2 - \delta b_1$ and $b^{r*}_{y1.x}$ estimates δ, where the superscript r reminds that these are the ordinary least squares estimates based on the r "complete" observations. Now, as for (7),

$$\hat{b}_2 = b^{r*}_{x.y1} + b_1^* b^{r*}_{y1.x} \tag{10}$$

and note that b_1^* the estimator of b_1 based on all observations has been substituted for b_1. But from (6)

$$b_2^{r*} = b^{r*}_{x.y1} + b_1^{r*} b^{r*}_{y1.x}$$

so that (10) may be rewritten

$$\hat{b}_2 = b_2^{r*} - (b_1^{r*} - b_1^*) b^{r*}_{y1.x}. \tag{11}$$

This is the estimator discussed by Conniffe (1985), where the asymptotic variance and indeed the exact small sample variances are given. Other published literature could be considered to describe situations falling within the class. For example, if for both y_1 and y_2 only the signs were available for all n, the multivariate probit model of Ashford and Sowdon (1970) would apply.

III OTHER CASES

One quite feasible situation is where a survey question presents such difficulty that although the question (the y_2 variable) has been reduced to (0,1) format (business surveys, for example, often reduce questions about expectations of future activity to "above or below normal") the question is unanswered by some respondents. Then there are t (<n) observations for which the probit analysis with x and y_1 as explanatory variables can be conducted. However, all n observations may be available for a y_1 variable and hence for estimating b_1. As before, b_1^* is the estimator for b_1 and it is intuitively obvious that (8) is the estimator for b_2, but with the difference that, although $b^p_{x.y1}$ and $b^p_{y1.x}$ are based on the r observations, the estimates b_1^* and σ_{11}^* are based on all n observations.

Another possibility is that we seek full information on y_2, but provide a (0,1) option for respondents unable or unwilling to provide full information. Then there could be r observations on both y_2 and y_1 and n–r observations on z and y_1. Considering the r observations alone, b_2^{r*}, the ordinary least

squares estimator, is clearly efficient and considering the n–r observations separately, the efficient estimator is given by (8), but with all components calculated only over the n–r observations. However, this actually estimates $b_2/\sqrt{\sigma_{22}}$ and so needs multiplying by the square root of σ_{22}^{r*}, the standard estimator of σ_{22} from the r observations, to give the estimator $\tilde{b}_2$. Since b_{22}^{r*} and σ_{22}^{r*} are independent and the r and n–r observations are assumed independent, an efficient combined estimator is obtained by the usual rule of combining inversely as the variances.

$$V_{n-r}(V_r + V_{n-r})^{-1} b_2^{r*} + V_r(V_r + V_{n-r})^{-1}\tilde{b}_2 \tag{12}$$

where V_r and V_{n-r} are used to denote the variances of the estimates for the r and n–r observations and $V_{n-r} = \sigma_{22}V_c + b_2 b_2'/2r$, where V_c denotes the variance of the Chesher estimator (8). In practice, the variances are replaced by estimates.

Yet another possible situation would arise from the previous case if some respondents failed to provide any information on y_2. Then there could be r observations on y_2 and y_1, t on z and y_1 and n–r–t extra observations on y_1. From the t observations on z and y_1 an estimator of b_2 can be based on the product of the Chesher estimator (8) and an estimator of σ_{22}. From the remaining n–t observations, the r on y_2 and y_1 and the n–r–t on y_1, an estimator $\hat{b}_2$, would be based on (11). Conniffe (1985) also described estimation of σ_{22} for this data pattern, giving all relevant formulae. Then an efficient overall estimator of b_2 is

$$V_{n-t}(V_t + V_{n-t})^{-1}\tilde{b}_2 + V_t(V_t + V_{n-t})^{-1}\hat{b}_2$$

where V_t is the variance of $\tilde{b}_2$ and V_{n-t} is the variance of $\hat{b}_2$.

So far, all examples have retained their ordinary least squares estimates of b_1, evaluated over all n observations, as the efficient estimator of b_1. So let us consider loss of information on both y_1 and y_2. Suppose a business survey seeks a (0,1) reply to one question and a quantitative reply to another. It could easily happen that quantitative answers are not available from some respondents who do provide replies to the (0,1) question. Then there are r observations with y_1 and z and n–r more with z recorded. An efficient estimator for b_2 (taking $\sigma_{22} = 1$) can be deduced along the lines previously employed. The r observations provide an estimator $\tilde{b}_2$ and the n–r provide an independent probit estimator b_2^p. Then an efficient combination is

$$b_2^+ = V_{n-r}(V_r + V_{n-r})^{-1}\ \tilde{b}_2 + V_r(V_r + V_{n-r})^{-1} b_2^p$$

where V_r and V_{n-r} are the variances of $\tilde{b}_2$ and b_2^p respectively.

As regards estimation of b_1, analogy with (10) suggests an estimator

$$b_1^{r*} - A(\tilde{b}_2 - b_2^+) \tag{13}$$

that is, a modification of the OLS estimator from the r observations by a multiple of the difference between the estimators of b_2 based on the r and n observations. Choosing A to minimise the (large sample) variance of (13) leads to

$$b_1^{r*} - \frac{\sigma_{12}}{\sigma_{11}} V_1 V_r^{-1}(\tilde{b}_2 - b_2^+) \tag{14}$$

where V_1 is the conventional variance (matrix) of b_1^{r*}. Replacing σ_{12}/σ_{11} by (9) and variances by estimates gives the estimator of b_1. Note that if y_2 was measured, this would reduce to (11) with b_1 and b_2 interchanged.

Since V_r must be larger for unmeasured y_2 then for measured y_2, the adjustment to the OLS estimate of b_1 ought to be smaller than in the measured case, which is an intuitively plausible result. The argument used to derive (14) is perhaps not sufficient to *prove* that it is a fully efficient estimator; that is, that its asymptotic variance attains the lower bound provided by the variance of the maximum likelihood estimator. However, a fully rigorous proof is possible, but requires some rather complicated mathematical detail.

It may be worth mentioning that one case of this example has already been examined in great detail. This is when the (0,1) variable is itself the indicator of the presence or absence of y_1. The context has been the possibility of sample selection bias (Heckman, 1979). However, that is a very special case indeed and will not be pursued here.

All the examples considered have the property of being computationally feasible, requiring just routines for ordinary least squares and probit analysis. But if we do not worry about computational matters, there are obviously many more cases that can be considered. Perhaps y_2 is replaceable by a pair of binary variables (z_1 z_2) or perhaps a single z can take a discrete set of values. Perhaps y_2 is fully observable up to some value and only larger values are indistinguishable. Again, both y_1 and y_2 could reduce to binary measurements for some sets of observations. There seem to be plenty of possibilities.

IV GENERALISATIONS AND RESERVATIONS

As mentioned in the Introduction, the assumption of common explanatory variables in both equations may sometime be regarded as an initial hypothesis, so that the estimators discussed in the preceding section may themselves be utilised to estimate the parameters of more restricted models. For example, once initial estimates of b_1 and b_2 have been obtained it is possible to test and estimate models that assume $b_{12} = b_{21} = 0$ where $b_1' = (b_{11}', b_{12}')$, $b_2' = (b_{21}', b_{22}')$ by exactly the same procedure as given in (4) and (5).

Other kinds of generalisation follow from considering more than two equations. By considering conditional distributions of y_2, given y_1 and y_3 say, another range of situations can be examined, in some of which probit analyses using x, y_1 and y_3 as explanatory variables would certainly feature. The issues associated with generalisation of (11) to multiple equations have been discussed by Conniffe (1985) and at least some of the deductions would carry over to the class of models considered here.

Turning to reservations about the use of the estimators discussed here, there are clearly assumptions implicit in the combination of estimators over different sub-samples. If y_1 and y_2 are fully measured in r observations and information on y_2 is lost in some fashion in another n–r, is there a danger that the r and n–r are really samples from different populations? This question has been addressed rather differently in mainline statistics and in econometrics. In the former, the question has usually been (for example, Rubin, 1976): when can we utilise the extra information in incomplete observations? The estimators based on the r observations are presumed to be valid and the issue is whether the incompleteness in the n–r observations is a warning that the underlying population has changed. In econometrics, Heckman (1979) has asked a rather different question: could the completeness of the r observations indicate that selection bias has produced a sample that no longer represents the population?

However, it is not difficult to test the assumptions involved, using the Hausman (1978) line of argument. The efficient estimator, say $\hat{b}_2$, and a single equation estimator, say b_2^*, differ in their asymptotic variances, but the actual estimates ought not to be too different, assuming that both are consistent estimators of the same parameter. So the difference $\hat{b}_2 - b_2^*$ ought not be large relative to its standard error. The old result of Fisher (1925) that the asymptotic covariance between an efficient and an inefficient estimator equals the asymptotic variance of the efficient estimator, implies that the variance of the difference is the difference of the variances. So implementation of Hausman tests is easy in this context.

Mentioning "asymptotic" does raise another issue. The idea of "efficiency" relates to minimum variance in large samples and in small samples the potential gain can be more than offset by the extra estimation error involved in estimating joint models rather than single models. In the case of linear equations with extra observations on y_1, where (11) is the efficient estimator of b_2, the asymptotic variance would suggest that (11) is always better than the single equation estimator b_2^{r*} provided the correlation between y_1 and y_2 is non-zero. But the exact small sample variance (Conniffe, 1985) shows that in small samples the correlation needs to be appreciable before there is any gain. Probit and other non-linear estimation methods are not amenable to the mathematical devices used to obtain exact small sample formulae, but simulation or other numerical approaches could clarify matters.

However, possible failure of asymptotic properties to hold in small samples is by no means confined to the class of models being considered in this paper. Every branch of econometrics utilises methods originally (and still in some cases) justified only by asymptotic arguments. Reservation about realisable gains in small samples should stimulate relevant investigations, rather than deter employment of intuitively plausible adjustments to standard formulae, which is what I hope I have shown the estimators in this paper to be.

REFERENCES

ANDERSON, T.W., 1958. *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

ASHFORD, J.R., and R.R. SOWDON, 1970. "Multivariate Probit Analysis", *Biometrics*, Vol. 26, pp. 535-546.

CHESHER, A., 1984. "Improving the Efficiency of Probit Estimators", *The Review of Economics and Statistics*, Vol. 66, pp. 523-527.

CONNIFFE, D., 1985. "Estimating Regression Equations with Common Explanatory Variables but Unequal Numbers of Observations", *Journal of Econometrics*, Vol. 27, pp. 179-196.

FISHER, R.A., 1925. "Theory of Statistical Estimation", *Proceedings of the Cambridge Philosophical Society*, Vol. 22, pp. 700-725.

GALLANT, A.R., 1975. "Seemingly Unrelated Nonlinear Regressions", *Journal of Econometrics*, Vol. 3, pp. 35-50.

HAUSMAN, J.A., 1978. "Specification Tests in Econometrics", *Econometrica*, Vol. 46, pp. 1,251-1,271.

HECKMAN, J., 1979. "Sample Selection Bias as Specification Error", *Econometrica*, Vol. 47, pp. 153-161.

RUBIN, D.B., 1976. "Inference and Missing Data", *Biometrika*, Vol. 63, pp. 581-592.

ZELLNER, A., 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", *Journal of the American Statistical Association*, Vol. 57, pp. 348-368.