Covariances for Bivariate Selection Model Second-Step Estimates

Consider the bivariate latent variable model with normal disturbances,

(1) $$y^* = x\beta + \varepsilon ,$$
$$w^* = z\alpha + \sigma\nu ,$$

where x and z are vectors of exogenous variables, not necessarily all distinct, α and β are parameter vectors, again not necessarily all distinct, and σ is a positive parameter. The interpretation of y^* is latent desired hours of work, and of w^* is latent log potential wage. The disturbances ε and ν have a standard bivariate normal distribution

(2) $$\begin{bmatrix}\varepsilon \\ \nu\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right) ,$$

with zero means, unit variances, and correlation ρ.

There is a nonlinear *observation rule* determined by the application that maps the latent variables into observations. A typical rule might be "Observe y = 1 and $w = w^*$ if $y^* > 0$; observe y = -1 and nominally w = 0 when $y^* \le 0$". This could correspond, for example, to an application where the event of working (y = 1) or not working (y = 0) is observed, but the wage is observed only if the individual works ($y^* > 0$).

The event of working is given by a *probit* model, P(y|x) = Φ(yxβ). The conditional log likelihood of w given participation is

(3) $$l^2(\alpha,\beta,\sigma,\rho) = -\log\sigma + \log\varphi\left(\frac{w-z\alpha}{\sigma}\right) + \log\Phi\left(\frac{x\beta+\rho\left(\frac{w-z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) - \log\Phi(x\beta).$$

See Chapter 2 for the derivation of this likelihood. Writing

(4) $$w = z\alpha + \sigma\mathbf{E}(\nu|x,z,y=1) + \xi$$

for the observations with participation defines a disturbance ξ that by construction has conditional mean zero, given x,z, in the selected sample. Using the property that the conditional expectation of ν given y = 1 equals the conditional expectation of ν given ε, integrated over the conditional density of ε given y = 1, plus the property of the normal that $d\varphi(\varepsilon)/d\varepsilon = -\varepsilon\cdot\varphi(\varepsilon)$, one has

(5) $$\mathbf{E}\{w|z,y=1\} = z\alpha + \sigma\mathbf{E}\{\nu|y=1\} = z\alpha + \sigma \int_{-x\beta}^{+\infty} \mathbf{E}\{\nu|\varepsilon\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= z\alpha + \sigma\rho \int_{-x\beta}^{+\infty} \varepsilon\varphi(\varepsilon)d\varepsilon/\Phi(x\beta) = z\alpha + \sigma\rho\varphi(x\beta)/\Phi(x\beta) \equiv z\alpha + \lambda M(x\beta),$$

where $\lambda = \sigma\rho$ and $M(c) = \varphi(c)/\Phi(c)$ is called the inverse Mill's ratio. Further, using the relationship

(6) $$\mathbf{E}(\nu^2|\varepsilon) = \text{Var}(\nu|\varepsilon) + \{\mathbf{E}(\nu|\varepsilon)\}^2 = 1 - \rho^2 + \rho^2\varepsilon^2,$$

and the integration-by-parts formula

(7) $$\int_{-c}^{+\infty} \varepsilon^2\varphi(\varepsilon)d\varepsilon = - \int_{-c}^{+\infty} \varepsilon\varphi'(\varepsilon)d\varepsilon = -c\varphi(c) + \int_{-c}^{+\infty} \varphi(\varepsilon)d\varepsilon = -c\varphi(c) + \Phi(c),$$

one obtains

(8) $$\mathbf{E}\{(w-z\alpha)^2|z,y=1\} = \sigma^2\mathbf{E}\{\nu^2|y=1\} = \sigma^2 \int_{-x\beta}^{+\infty} \mathbf{E}\{\nu^2|\varepsilon\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= \sigma^2 \int_{-x\beta}^{+\infty} \{1 - \rho^2 + \rho^2\varepsilon^2\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta) = \sigma^2\{1 - \rho^2 + \rho^2 - \rho^2 x\beta\varphi(x\beta)/\Phi(x\beta)\}$$

$$= \sigma^2\{1 - \rho^2 x\beta\varphi(x\beta)/\Phi(x\beta)\} = \sigma^2\{1 - \rho^2 x\beta\cdot M(x\beta)\}.$$

Then,

(9) $$\mathbf{E}\ \{[w - z\alpha - \boldsymbol{E}\{w-z\alpha|z,y=1\}]^2|z,y=1\} = \mathbf{E}\{(w-z\alpha)^2|z,y=1\} - [\mathbf{E}\{w-z\alpha|z,y=1\}]^2$$

$$= \sigma^2\{1 - \rho^2 x\beta\varphi(x\beta)/\Phi(x\beta) - \rho^2\varphi(x\beta)^2/\Phi(x\beta)^2\} = \sigma^2\{1 - \rho^2 M(x\beta)[x\beta + M(x\beta)]\}.$$

Then (4) is a regression equation that with β known so that $M(x\beta)$ is a known transformation of observations has $E(\xi|x,z,y=1) = 0$ for the participant sub-population, but $E(\xi^2|x,z,y=1)$ given by (9) and heteroscedastic..

The Heckman two-step procedure estimates (4) by OLS, with a consistent estimator for β obtained from the probit selection model plugged into the inverse Mills ratio formula. This regression estimates only the product $\lambda \equiv \sigma\rho$, but consistent estimates of σ and ρ could be obtained in a further step: The OLS regression does not correct for heteroscedasticity or the presence of earlier stage estimates. It is nevertheless consistent, but printed out standard errors are not consistent.

The general theory of two-step GMM estimation can be used to get consistent estimates of the covariance matrix for the estimates in (4). Suppose a random sample of observations (x,y,z,w), with x and z interpreted as explanatory variables, y as the dependent variable determining selection, and w as the dependent variable determining the continuous outcome, given selection. Suppose in the first stage one estimates the parameter vector β by b_N satisfying

(10) $$0 = \mathbf{E}_N h(b_N;x,y),$$

where $\mathbf{E}_N$ denotes empirical expectation (or sample average) and $h(\beta;x,y) = \nabla_\beta \log\Phi(yx\beta)$ is the score of the marginal likelihood of y.

Suppose in the second stage one estimates a parameter vector $\theta = (\alpha,\lambda)$ with $\lambda = \sigma\rho$ by coefficients $t_N = (a_N,c_N)$ obtained by applying OLS to the model $w = z\alpha + \lambda M(xb_N) + \zeta$, where $M(xb_N) = \varphi(xb_N)/\Phi(xb_N)$ is an inverse Mills ratio evaluated at xb_N. This regression corresponds to the moments

$$(11)\qquad 0 = \mathbf{E}_N g(t_N,b_N;x,y,z,w) \equiv \mathbf{E}_N \begin{bmatrix} z' \\ M(xb_N) \end{bmatrix} (w - za_N - c_N M(xb_N)).$$

Note that the g moments do not depend on y; this will simplify formulas later. Make a Taylor's expansion of both the first-stage and the second-stage moment conditions around the true β, α, and λ. Suppress the x,y,z,w arguments to simplify notation:

$$(12)\qquad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sqrt{N} \begin{bmatrix} \boldsymbol{E}_N h(\beta) \\ \boldsymbol{E}_N g(\theta,\beta) \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix} \sqrt{N}\ (b_N - \beta) - \begin{bmatrix} 0 \\ C \end{bmatrix} \sqrt{N}\ (t_N - \theta) + o_p,$$

where $A = -\text{plim}\ \mathbf{E}_N \nabla_\beta h(\beta)$, $B = -\text{plim}\ \mathbf{E}_N \nabla_\beta g(\theta,\beta)$, and $C = -\text{plim}\ \mathbf{E}_N \nabla_\theta g(\theta,\beta)$.

The term $\sqrt{N} \begin{bmatrix} \boldsymbol{E}_N h(\beta) \\ \boldsymbol{E}_N g(\theta,\beta) \end{bmatrix}$ is asymptotically normal by a central limit theorem, with a covariance matrix $\begin{bmatrix} \Omega_{hh} & \Omega_{hg} \\ \Omega_{gh} & \Omega_{gg} \end{bmatrix}$, with $\Omega_{hh} = \text{plim}\ \mathbf{E}_N h(\beta)h(\beta)'$, $\Omega_{hg} = \text{plim}\ \mathbf{E}_N h(\beta)g(\theta,\beta)'$, and $\Omega_{gg} = \text{plim}\ \mathbf{E}_N g(\theta,\beta)g(\theta,\beta)'$. Solve the first block of equations and substitute the solution $\sqrt{N}\ (b_N - \beta) = A^{-1} \sqrt{N}\ \mathbf{E}_N h(\beta) + o_p$ into the second block to obtain

$$(13)\qquad 0 = \sqrt{N}\ \{\mathbf{E}_N g(\theta,\beta) - BA^{-1}\mathbf{E}_N h(\beta)\} - C \sqrt{N}\ (t_N - \theta) + o_p.$$

The term in braces on the right-hand-side of this expression has an asymptotic covariance matrix

$$(14)\qquad \Omega_{gg} - BA^{-1}\Omega_{hg} - \Omega_{gh}A'^{-1}B' + BA^{-1}\Omega_{hh}A'^{-1}B'.$$

Then, $\sqrt{N}\ (t_N - \theta) = C^{-1}\ \sqrt{N}\ \{\mathbf{E}_N g(\theta,\beta) - BA^{-1}\mathbf{E}_N h(\beta)\} + o_p$ has asymptotic covariance matrix

$$\text{(15)} \qquad acov(t_N) = C^{-1}\{\Omega_{gg} - BA^{-1}\Omega_{hg} - \Omega_{gh}A'^{-1}B' + BA^{-1}\Omega_{hh}A'^{-1}B'\}C'^{-1}$$

This is the general formula for covariances of second-step estimators, and there will be some simplification for the application. In general, if B = 0, there is no correction; this is the "block diagonality" case where θ can be estimated consistently even if the estimator of β is not consistent. In the bivariate selection problem, letting $m(x\beta) = dM(x\beta)/d(x\beta)$,

$$\text{(16)} \qquad B = -\mathbf{E}\,\nabla_\beta \begin{bmatrix} z' \\ M(x\beta) \end{bmatrix} (w\text{-}z\alpha\text{-}\lambda M(x\beta))$$

$$\text{(17)} \qquad = -\lambda\mathbf{E} \begin{bmatrix} z'xm(x\beta) \\ M(x\beta)xm(x\beta) \end{bmatrix} - \mathbf{E}_{x,z} \begin{bmatrix} z' \\ m(x\beta)x' \end{bmatrix} \mathbf{E}_{w|x,z}(w\text{-}z\alpha\text{-}\lambda M(x\beta)) = -\lambda\mathbf{E} \begin{bmatrix} z'xm(x\beta) \\ M(x\beta)xm(x\beta) \end{bmatrix}$$

.

Then B = 0 if $\lambda = \sigma\rho = 0$; i.e., if $\rho = 0$ so that selection is "at random" and is independent of the determination of the latent w. A useful implication of this result is that for testing the hypothesis that $\rho = 0$, one is in the block diagonal case under the null, so that the covariance correction for first-stage estimation does not enter the determination of standard errors. Further, the w regression is homoscedastic when $\rho = 0$. Then, a conventional T-test for $\lambda = 0$ can be carried out using the standard OLS statistics, without any corrections for the estimation of β in the inverse Mills ratio or for heteroscedasticity.

It will be useful to work out the form of the C matrix,

$$\text{(18)} \qquad C = -\mathbf{E}\,\nabla_\theta \begin{bmatrix} z' \\ M(x\beta) \end{bmatrix} (w\text{-}z\alpha\text{-}\lambda M(x\beta)) = \begin{bmatrix} z'z & z'M(x\beta) \\ M(x\beta) & M(x\beta)^2 \end{bmatrix} .$$

Note that C is just the "X′X" matrix for the second-stage regression. If the regression were homoscedastic, then C would equal Ω_{gg}. However, it is not, and Ω_{gg} is the array appearing in the center of White's robust covariance matrix estimator for regressions with heteroscedasticity of unknown form.

If g does not depend on y, then $\Omega_{gh} = \mathbf{E}_{x,z}\mathbf{E}_{w|x,z}\{g\cdot\mathbf{E}_{y|x}h\} = 0$. This is true for the bivariate selection application. Then, one has the simplification

$$\text{(19)} \qquad acov(t_N) = C^{-1}\{\Omega_{gg} + BA^{-1}\Omega_{hh}A'^{-1}B'\}C'^{-1}$$

The fact that the first stage of estimation in the bivariate selection problem was maximum marginal likelihood gives a further simplification. One has $h(\beta;x,y) = \nabla_\beta \log\Phi(yx\beta)$, and hence $A = -\mathbf{E}\nabla_{\beta\beta}\log\Phi(yx\beta)$. The information equality for maximum likelihood then implies $A = \mathbf{E}(\nabla_\beta\log\Phi(yx\beta))(\nabla_\beta\log\Phi(yx\beta))' = \Omega_{hh}$, and the covariance matrix for t_N simplifies further to

(20) $$\text{acov}(t_N) = C^{-1}\{\Omega_{gg} + B(\Omega_{hh})^{-1}B'\}C'^{-1}.$$

All the terms of this covariance matrix could be estimated from sample analogs, computed at the consistent estimates. The following table summarizes consistent estimators for the various covariance terms; recall that $\mathbf{E}_N$ denotes empirical expectation (sample average):

Matrix	Estimator
C	$-\mathbf{E}_N\nabla_\theta g(t_N,b_N)$
B	$-\mathbf{E}_N\nabla_\beta g(t_N,b_N)$
A	$-\mathbf{E}_N\nabla_\beta h(b_N)$
Ω_{hh}	$\mathbf{E}_N h(b_N)h(b_N)'$
Ω_{gh}	$\mathbf{E}_N g(t_N,b_N)h(b_N)'$
Ω_{gg}	$\mathbf{E}_N g(t_N,b_N)g(t_N,b_N)'$