Online Appendix for "By choice or by force? Uncovering the nature of informal employment in urban Mexico"

Robert Duval-Hernández

Open University of Cyprus, CIDE, and IZA

November 21, 2020

This note presents in more detail the econometric models estimated in the paper. Further information can be found in Tunalı (1986).

1 Discrete Choice Model

For an individual i, denote the latent propensity to apply to a formal job by V_i^a, and the propensity of being hired by V_i^h. These propensities depend on characteristics of the workers, some of which are observable to the econometrician. In particular, assume that the following structure holds:

$$V_i^a = Z_i\gamma_a + u_{ia} \tag{1}$$

$$V_i^h = X_i\gamma_h + u_{ih}, \tag{2}$$

where Z_i and X_i are vectors of observable individual characteristics, and u_{ia} and u_{ih} are random terms capturing other unobservable factors. Assume that the vector of error terms (u_{ia}, u_{ih}) follows a standard bivariate normal law with correlation parameter ρ; i.e. the vector has zero mean and variance-covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{3}$$

The above propensities in equations (1) and (2) are not observable. Instead it is only observed whether they are positive or not, i.e. only the following indicators are observed

$$D_{ij} = \begin{cases} 1 & \text{if } V_i^j > 0 \\ 0 & \text{if } V_i^j \leq 0 \end{cases} \tag{4}$$

for $j \in \{a, h\}$.

Denote by $G(\cdot, \cdot; \cdot)$ the standard bivariate normal distribution and by $F(\cdot)$ its univariate counterpart. Also, let $C_a = Z\gamma_a$ and $C_h = X\gamma_h$.

Given the above assumptions and notation, one can characterize the probability that a given individual with explanatory variables (Z_i, X_i) is either a formal, $i \in \mathcal{F}$, an involuntary informal, $i \in \mathcal{I}$, or a voluntary informal, $i \in \mathcal{V}$, worker.[1] These probabilities are given by the following equations:

Formal worker

$$\begin{aligned} P(D_a = 1, D_h = 1) &= P(V^a > 0, V^h > 0) \\ &= P(u_a > -Z\gamma_a, u_h > -X\gamma_h) \\ &= G(C_a, C_h; \rho) \end{aligned} \tag{5}$$

Involuntary informal worker

$$\begin{aligned} P(D_a = 1, D_h = 0) &= P(V^a > 0, V^h \leq 0) \\ &= P(u_a > -Z\gamma_a, u_h \leq -X\gamma_h) \\ &= G(C_a, -C_h; -\rho) \end{aligned} \tag{6}$$

Voluntary informal worker

$$\begin{aligned} P(D_a = 0) &= P(V^a \leq 0) \\ &= P(u_a \leq -Z\gamma_a) \\ &= F(-C_a) \end{aligned} \tag{7}$$

With these probabilities, the likelihood function of the discrete choice problem used in the paper can be formulated as

$$\mathcal{L} = \prod_{D_a=0} [1 - F(C_a)] \cdot \prod_{\substack{D_a=1 \\ D_h=0}} G(C_a, -C_h; -\rho) \cdot \prod_{\substack{D_a=1 \\ D_h=1}} G(C_a, C_h; \rho). \tag{8}$$

This model is sometimes called a bivariate probit with sample selection (see for instance Van de Ven and Van Praag, 1981).

Furthermore, this framework can be extended as to estimate earnings regressions adjusted for sample selectivity. In particular, assume the existence of segment-specific earnings functions

$$\log y_s = X_s \beta_s + u_s$$

where y denotes earnings, X_s is a vector of observable characteristics, u is an unobservable residual, and the subscript s denotes the type of worker under consideration, namely: formal, involuntary informal, and voluntary informal. To estimate earnings equations adjusted for sample selectivity bias, assume the joint multivariate normality of the error terms u_a, u_h and the three errors terms e_s, for $s \in \{\mathcal{F}, \mathcal{I}, \mathcal{V}\}$.

[1] From now on, the individual subscript i is dropped for simplicity.

This is a switching regression model with earnings functions defined for all the employed population. However, for any given individual only one realization of earnings is observed, depending on the market segment at which he or she ends up being employed. More precisely, the segment switching rule is the following:

$$\log y = \begin{cases} \log y_{\mathcal{V}} & \text{if } D_a = 0 \\ \log y_{\mathcal{I}} & \text{if } D_a = 1 \text{ and } D_h = 0 \\ \log y_{\mathcal{F}} & \text{if } D_a = 1 \text{ and } D_h = 1 \end{cases}$$

Due to the fact that the segment allocation is not random, when estimating the earnings equations through Ordinary Least Squares in a second stage one needs to correct for sample selectivity bias. In particular, one needs to estimate

$$\begin{array}{ll} \log y = X_{\mathcal{V}}\beta_{\mathcal{V}} + \rho_a^{\mathcal{V}}\hat{\lambda}_0 + \tilde{u}_{\mathcal{V}} & \text{if } D_a = 0 \\ \log y = X_{\mathcal{I}}\beta_{\mathcal{I}} + \rho_a^{\mathcal{I}}\hat{\lambda}_{a\mathcal{I}} + \rho_h^{\mathcal{I}}\hat{\lambda}_{h\mathcal{I}} + \tilde{u}_{\mathcal{I}} & \text{if } D_a = 1 \text{ and } D_h = 0 \\ \log y = X_{\mathcal{F}}\beta_{\mathcal{F}} + \rho_a^{\mathcal{F}}\hat{\lambda}_{a\mathcal{F}} + \rho_h^{F}\hat{\lambda}_{hF} + \tilde{u}_{\mathcal{F}} & \text{if } D_a = 1 \text{ and } D_h = 1 \end{array} \quad (9)$$

The selection correction terms are given by

$$\lambda_0 = -\frac{f(C_a)}{F(-C_a)}$$

$$\lambda_{a\mathcal{I}} = \frac{f(C_a)F(-C_h^*)}{G(C_a,-C_h;-\rho)} \quad \lambda_{h\mathcal{I}} = -\frac{f(C_h)F(C_a^*)}{G(C_a,-C_h;-\rho)} \quad (10)$$

$$\lambda_{a\mathcal{F}} = \frac{f(C_a)F(C_h^*)}{G(C_a,C_h;\rho)} \quad \lambda_{h\mathcal{F}} = \frac{f(C_h)F(C_a^*)}{G(C_a,C_h;\rho)}$$

where C_a^* and C_h^* are given by

$$C_a^* = \frac{C_a - \rho C_h}{(1-\rho^2)^{\frac{1}{2}}} \qquad C_h^* = \frac{C_h - \rho C_a}{(1-\rho^2)^{\frac{1}{2}}}.$$

Since the selectivity correction terms λ's are estimated using information arising from the parameters estimates of the discrete choice model, then the standard errors of the regressions (9) must be adjusted to account for the presence of generated regressors. The formulas to do this, as well as the derivation of the full model can be found in Tunalı (1986).

References

Tunalı, I. (1986). A General Structure for Models of Double Selection and an Application to a Joint Migration/Earnings Process with Remigration. *Research in Labor Economics 8*(Part B), 235–282.

Van de Ven, W. P. and B. M. Van Praag (1981). The demand for deductibles in private health insurance. A probit model with sample selection. *Journal of Econometrics 17*(2), 229–252.