Double-Hurdle Models with Dependent Errors and Heteroscedasticity

Julian Fennema and Mathias Sinning

Heriot-Watt University and RWI-Essen

Essen, April 2nd 2007
Tobit model

**Model**

Censoring of the dependent variable is traditionally dealt with using the Tobit model

\[
y_i = y_i^* = x_i \beta + \varepsilon_1 \quad \text{if} \quad y_i^* > 0 \quad \text{otherwise} \quad y_i = 0
\]

**Likelihood function**

\[
L = \prod_{0} \left(1 - \Phi \left(\frac{x_i \beta}{\sigma \varepsilon_1}\right)\right) + \prod_{+} \left(\frac{1}{\sigma \varepsilon_1} \phi \left(\frac{y_i - x_i \beta}{\sigma \varepsilon_1}\right)\right)
\]
Cragg’s formulation

Model

Cragg (1971) proposed the extension that the probability of a zero realisation, $1 - \Phi(.)$, is not directly to the density for a continuous realisation, $\phi(.)$, but instead governed by some other process.

\[ y_i = y_i^* = x_i\beta + \varepsilon_{1i} \]
\[ = 0 \]

if \[ x_i\beta + \varepsilon_{1i} > 0 \text{ and } z_i\alpha + \varepsilon_{2i} > 0 \]

if \[ x_i\beta + \varepsilon_{1i} \leq 0 \text{ and } z_i\alpha + \varepsilon_{2i} > 0 \]

or \[ x_i\beta + \varepsilon_{1i} > 0 \text{ and } z_i\alpha + \varepsilon_{2i} \leq 0 \]

or \[ x_i\beta + \varepsilon_{1i} \leq 0 \text{ and } z_i\alpha + \varepsilon_{2i} \leq 0 \]
Cragg’s formulation - errors

Independent

The original model made the assumption that the two error terms were jointly normal,

\[
\begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{pmatrix} \sim N(0, \Sigma), \text{ and uncorrelated, } \Sigma = \begin{pmatrix}
\sigma_{\varepsilon_1}^2 & 0 \\
0 & 1
\end{pmatrix}
\]

Likelihood function

\[
L = \prod_{i=0} \left(1 - \Phi \left( \frac{x_i \beta}{\sigma_{\varepsilon_1}} \right) \cdot 1 - \Phi (z_i \alpha) \right) \prod_{i=0} \left( \Phi (z_i \alpha) \frac{1}{\sigma_{\varepsilon_1}} \phi \left( \frac{y_i - x_i \beta}{\sigma_{\varepsilon_1}} \right) \right)
\]
Separability

Similar to that demonstrated by McDowell (2003) for count models in Stata, the separability of the likelihood function permits the use of a combination of Stata command to estimate this model: `truncreg` and `probit`
Jones’ extension

Correlated errors

This assumption has been relaxed in later work, e.g. Jones (1992), where \( \Sigma = \begin{pmatrix} \sigma^2_{\varepsilon_1} & \sigma_{\varepsilon_1 \rho} \\ \sigma_{\varepsilon_1 \rho} & 1 \end{pmatrix} \)

Likelihood function

\[
L = \prod_0^1 [1 - F_2(z_i \alpha, x_i \beta/\sigma, \rho)] \prod_+ \Phi \left( \frac{z_i \alpha + \frac{\rho}{\sigma} (y - x_i \beta)}{\sqrt{1 - \rho^2}} \right) \frac{1}{\sigma} \phi \left( \frac{(y - x_i \beta)}{\sigma} \right)
\]
Non-separable

Both parts of this likelihood function must, however, be maximised simultaneously; there is no two-step equivalent. This has been available in Stata, on an *ad hoc* basis since 2004 using the `dhurdle` command written for Stata 7.

Syntax

dhurdle y x1 x2, sel(d x1 t1)
### Comparison to Flood and Gråsjö

<table>
<thead>
<tr>
<th></th>
<th>True value</th>
<th>Stata Bias(%)</th>
<th>RMSE</th>
<th>Gauss Bias(%)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-0.2</td>
<td>-23.3</td>
<td>.665</td>
<td>-18.0</td>
<td>.836</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.2</td>
<td>-17.6</td>
<td>.152</td>
<td>3.7</td>
<td>.085</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1</td>
<td>.34</td>
<td>.081</td>
<td>2.7</td>
<td>.182</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>0.7</td>
<td>49.0</td>
<td>1.25</td>
<td>175</td>
<td>8.065</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.2</td>
<td>24.6</td>
<td>.881</td>
<td>-44.7</td>
<td>1.325</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-0.2</td>
<td>27.3</td>
<td>.392</td>
<td>-80.7</td>
<td>.570</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>2</td>
<td>1.38</td>
<td>.256</td>
<td>-.2</td>
<td>.227</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.5</td>
<td>-26.3</td>
<td>.261</td>
<td>31.2</td>
<td>.447</td>
</tr>
</tbody>
</table>
There is, by now, a wide variety of literature demonstrating that if the assumption of homoscedastic, normally-distributed, errors is violated then ML parameter estimates are inconsistent.

Solutions

Two extensions

1. Heteroscedastic errors
2. Non-normal errors
T reatment of zeroes
Double hurdle model
Extensions
Heteroscedasticity
Non-normality
Pipeline

**dhurdle** now can incorporate variance dependent on a set of independent variables.

**Syntax**

dhurdle y x1 x2, sel(d x1 t1) het(.)
Robinson (1982) showed that ML estimation of LDV models leads to inconsistent parameter estimates if the assumption of normally distributed errors does not hold.

Syntax

dhurdle y x1 x2, sel(d x1 t1) ihs
Work in progress

The final steps to complete the estimation package that are currently underway are

1. Finalise `predict` options for the double hurdle.
Testing 1, 2, 3

Using the IHS as the general form, the imposition of the following restrictions is feasible:

1. If $\gamma = 0$ then conventional formulation without transformation.
2. If $\sigma$ is constant then homoscedastic errors.
3. If $\rho = 0$ then independent double hurdle.
4. If $\prod \Phi(z_i\alpha)$ then no censoring present and model simplifies to a Heckman.
5. If $\Phi(z_i\alpha)$ and $\rho = 0$ then no censoring or selection present and model simplifies to a Tobit.