Recent Advances in the Field of Trade Theory and Policy Analysis Using Micro-Level Data

July 2012

Bangkok, Thailand

Cosimo Beverelli
(World Trade Organization)

Content

a) Binary dependent variable models in cross-section

b) Binary dependent variable models with panel data

c) Examples of firm-level analysis

b) Binary dependent variable models in cross-section

- Binary outcome
- Latent variable
- Linear probability model (LMP)
- Probit model
- Logit model
- Marginal effects
- Odds ratio in logit model
- Maximum likelihood (ML) estimation
- Rules of thumb

Binary outcome

- In many applications the dependent variable is not continuous but qualitative, discrete or mixed:
  - Qualitative: car ownership (Y/N)
  - Discrete: education degree (Ph.D., University degree,…, no education)
  - Mixed: hours worked per day

- Here we focus on the case of a binary dependent variable
  - Example with firm-level data: exporter status (Y/N)

Binary outcome (ct'd)

- Let y be a binary dependent variable:

$$y = \begin{cases} 1 & \textit{with probability } p \\ 0 & \textit{with probability } 1 - p \end{cases}$$

- A regression model is formed by parametrizing the probability p to depend on a vector of explanatory variables $\boldsymbol{x}$ and a $K \times 1$ parameter vector β
- Commonly, we estimate a conditional probability:

$$p_i = \Pr[y_i = 1|\boldsymbol{x}] = F(\boldsymbol{x}_i{}'\beta) \qquad (1)$$

where $F(\cdot)$ is a specified function

Intuition for $F(\cdot)$: latent variable

- Imagine we wanted to estimate the effect of $\boldsymbol{x}$ on a continuous variable y^*
- The "index function" model we would like to estimate is:

$$y_i{}^* = \boldsymbol{x}_i{}'\beta - \varepsilon_i$$

- However, we do not observe y^* but only the binary variable y

$$y = \begin{cases} 1 & if\ y^* > 0 \\ 0 & otherwise \end{cases}$$

Intuition for $F(\cdot)$: latent variable (ct'd)

- There are two ways of interpreting ${y_i}^*$:

1. Utility interpretation: ${y_i}^*$ is the additional utility that individual i would get by choosing $y_i = 1$ rather than $y_i = 0$

2. Threshold interpretation: ε_i is a threshold such that if $\boldsymbol{x}_i{}'\beta > \varepsilon_i$, then $y_i = 1$

- The parametrization of p_i is:

$$
\begin{aligned}
p_i = \Pr[y = 1|\boldsymbol{x}] &= \Pr[y^* > 0|x] = \Pr[\,\boldsymbol{x}'\beta - \varepsilon > 0|x] \\
&= \Pr[\varepsilon < \boldsymbol{x}'\beta] = F[\boldsymbol{x}'\beta]
\end{aligned}
$$

where $F(\cdot)$ is the CDF of ε

Linear probability model (LMP)

- The LPM does not use a CDF, but rather a linear function for $F(\cdot)$
- Therefore, equation (1) becomes:

$$p_i = \Pr[y_i = 1|\boldsymbol{x}] = \boldsymbol{x}_i{}'\beta$$

- The model is estimated by OLS with error term ε_i
- From basic probability theory, it should be the case that $0 \leq p_i \leq 1$
- This is not necessarily the case in the LPM, because $F(\cdot)$ in not a CDF (which is bounded between 0 and 1)
    - Therefore, one could estimate predicted probabilities $\hat{p}_i = \boldsymbol{x}_i{}'\hat{\beta}$ that are negative or exceed 1