Limited Dependent Variables and Panel Data

Christina Zauner

June 24^{th}, 2009

Structure

Motivation

Many economic questions involve the explanation of binary variables, e.g.:

- explaining the participation of women in the labor market
- explaining retirement decisions
- studying the effect of unemployment on the level of satisfaction (see WINKELMANN & WINKELMANN (1998))

$\Rightarrow$ Limited Dependent Variables are mainly used in **cross-sectional panels** (large N, small T), i.e. for microeconometric questions.

Introduction

Binary choice models are formulated in terms of a **latent variable**:

- e.g. the choice of participating in the labor market (the variable of interest) depends on whether the offered wage is greater than the unobserved reservation wage of the given person (the latent variable)

Formally,

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* = x_{it}'\beta + u_{it} > 0 \\ 0 & \text{if } y_{it}^* = x_{it}'\beta + u_{it} \leq 0 \end{cases}$$

where x_{it} is a vector of explanatory variables for person i at time t, where β is a vector of coefficients to be estimated and where u_{it} is a error which will be specified below.

Assumptions and Problems of the Fixed Effects Estimator

Assume that $u_{it} = \mu_i + \nu_{it}$, where:

- μ_i is an unknown parameter to be estimated
- ν_{it} is a remainder error which is assumed to be i.i.d. across individuals and time with a symmetric cumulative distribution function $F(\cdot)$

The incidental parameters problem:

- In the linear case, the estimators for μ_i are not consistent if $N \to \infty$ but T remains fixed. Nevertheless, β can be estimated consistently.
- With qualitative data, the estimators for β and μ_i are however not independent $\Rightarrow$ both μ_i and β are inconsistent in cross-sectional panels

Maximum Likelihood Estimation

The log-likelihood function is given by:

$$\begin{aligned} \log L\left(\beta, \mu_1, \ldots, \mu_N\right) = & \sum_{i,t} y_{it} \log\left[Pr\left\{y_{it} = 1\right\}\right] + \\ + & \sum_{i,t} \left(1 - y_{it}\right) \log\left[Pr\left\{y_{it} = 0\right\}\right] \end{aligned}$$

where

$$\begin{aligned} Pr\left\{y_{it} = 1\right\} &= Pr\left\{y_{it}^* > 0\right\} \\ &= Pr\left\{\mu_i + x_{it}'\beta + \nu_{it} > 0\right\} \\ &= Pr\left\{-\nu_{it} < \mu_i + x_{it}'\beta\right\} \\ &= F\left(\mu_i + x_{it}'\beta\right) \end{aligned}$$

For small T and large N, maximizing the log-likelihood function yields however inconsistent estimators!

Conditional Maximum Likelihood Estimation

In order to obtain a consistent estimator also for small T, a **sufficient statistic** t_i for μ_i is needed:

- Conditional on t_i, the likelihood contribution of individual i does not depend any longer on μ_i, but only on β, i.e. the density is such that
  $f\left(y_{i1}, \ldots, y_{iT} | t_i, \mu_i, \beta\right) = f\left(y_{i1}, \ldots, y_{iT} | t_i, \beta\right)$
- Maximizing the conditional log-likelihood function based on $f\left(y_{i1}, \ldots, y_{iT} | t_i, \beta\right)$ with respect to β yields a consistent estimator for β.

HOWEVER: For the probit model (remainder error normally distributed), there is no sufficient statistic for μ_i !!! $\Rightarrow$ It is not possible to estimate β consistently in cross-sectional panels!

The Fixed Effects Logit Model - An Example

For the fixed effects logit model, a sufficient statistic for μ_i is $t_i = \sum_t y_{it}$. (CHAMBERLAIN (1980)) Let $T = 2$:

- If $t_i = 0$, there is only one possible event, namely $y_{i1} = 0$ and $y_{i2} = 0$. (Similarly, for $t_i = 2$) Hence, individuals who do not change status over the time horizon do not enter the conditional log-likelihood function.
- It is sufficient to consider individuals with $t_i = 1$ with the possible events $(y_{i1}, y_{i2}) = (0, 1)$ and $(y_{i1}, y_{i2}) = (1, 0)$:
  - Using Bayes rule, it holds that:

$$Pr\left\{(0,1)\,|t_1 = 1, \mu_i, \beta\right\} = \frac{Pr\left\{(0,1)\,|\mu_i, \beta\right\}}{Pr\left\{(0,1)\,|\mu_i, \beta\right\} + Pr\left\{(1,0)\,|\mu_i, \beta\right\}}$$

  - Moreover,
    $Pr\left\{(0,1)\,|\mu_i, \beta\right\} = Pr\left\{y_{i1} = 0|\mu_i, \beta\right\} Pr\left\{y_{i2} = 1|\mu_i, \beta\right\}$

The Fixed Effects Logit Model - An Example

In the logit model it holds that

$$Pr\left\{y_{it}=1|\mu_i,\beta\right\}=\frac{\exp\left\{\mu_i+x_{it}'\beta\right\}}{1+\exp\left\{\mu_i+x_{it}'\beta\right\}}$$

and correspondingly:

$$Pr\left\{y_{it}=0|\mu_i,\beta\right\}=1-\frac{\exp\left\{\mu_i+x_{it}'\beta\right\}}{1+\exp\left\{\mu_i+x_{it}'\beta\right\}}=\frac{1}{1+\exp\left\{\mu_i+x_{it}'\beta\right\}}$$

Inserting this into the expression for $Pr\left\{(0,1)\,|t_i=1,\mu_i,\beta\right\}$ yields:

$$Pr\left\{(0,1)\,|t_i=1,\mu_i,\beta\right\}=\frac{\exp\left\{(x_{i2}-x_{i1})'\,\beta\right\}}{1+\exp\left\{(x_{i2}-x_{i1})'\,\beta\right\}}$$

and correspondingly:

$$Pr\left\{(1,0)\,|t_i=1,\mu_i,\beta\right\}=1-Pr\left\{(0,1)\,|t_i=1,\mu_i,\beta\right\}$$

The Fixed Effects Logit Model - An Example

Hence, in the case of $T = 2$ the fixed effects logit model is equivalent to a simple cross-sectional logit model using

- an endogenous variable $\tilde{y}_i$ which is one if y_{it} changes from 0 to 1 between $t = 1$ and $t = 2$ and zero if y_{it} changes from 1 to 0
- and the difference $x_{i2} - x_{i1}$ as the explanatory variable

Testing for Fixed Effects

Testing fixed individual effects versus no fixed individual effects can be done using a Hausmann test:

- The logit estimator ignoring individual effects is consistent and efficient under the null hypothesis of no individual effects and inconsistent under the alternative
- The fixed effects logit estimator is consistent under both hypotheses, but inefficient if there are no individual effects
- Under the null hypothesis the test statistic is χ^2_K distributed

Assumptions and Problems of the Random Effects Estimator

Usually, a probit specification is used for the random effects estimator, i.e. it is assumed that $u_{it} = \mu_i + \nu_{it}$, where μ_i is i.i.d. with $N\left(0, \sigma_\mu^2\right)$ and where ν_{it} is i.i.d. with $N\left(0, \sigma_\nu^2\right)$.

Computational Problems:

- Because of the individual random effect, the observations from individual i are correlated over time, i.e. $E\left(u_{it} u_{is}\right) = \sigma_\mu^2$ for $t \neq s$
- In order to obtain the probability contribution of any individual i, T-dimensional integrals are necessary $\Rightarrow$ this is not feasible if T is large

Assumptions and Problems of the Random Effects Estimator

Conditioning on the individual effects, this reduces to a one dimensional integral:

$$
\begin{aligned}
& f\left(y_{i1},\ldots,y_{iT}|x_{i1},\ldots,x_{iT},\beta\right) \\
= & \int_{-\infty}^{\infty} f\left(y_{i1},\ldots,y_{iT}|x_{i1},\ldots,x_{iT},\mu_i,\beta\right) f\left(\mu_i\right) d\mu_i \\
= & \int_{-\infty}^{\infty} \left[\prod_t f\left(y_{it}|x_{it},\mu_i,\beta\right)\right] f\left(\mu_i\right) d\mu_i
\end{aligned}
$$

which can be approximated through numerical integration (e.g. Gaussian quadrature procedure)

Simulation Estimation

- Unfeasible integrals can be replaced by unbiased Monte Carlo probability simulators (method of simulated moments (MSM) estimator)
- Therefore, it is necessary to simulate M^T potential choice sequences for each individual, where M is the number of possible choices each individual faces in any period $\Rightarrow$ also not feasible if T is large
- The MSM estimator is asymptotically as efficient as maximum likelihood

State Dependence or Unobserved Heterogeneity

Frequently, it is observed that people with a long history of unemployment are less likely to find a job. This might have two reasons:

- State Dependence: People who are unemployed for a long time loose their skills or become discouraged from searching a job
- Unobserved Heterogeneity: There is a selection mechanism present, meaning that people who share certain characteristics are those who are long-term unemployed

Testing for State Dependence

- Test for $\gamma = 0$ in the model $y_{it}^* = x_{it}'\beta + \gamma y_{it-1} + \mu_i + \tilde{\nu}_{it}$
- If $\gamma = 0$ is rejected, this can either be due to true state dependence or serially correlated remainder errors ν_{it}
- In order to find out whether true state dependence is present, it is possible to condition on lagged values of the explanatory variables without conditioning on the lagged state
- If conditioning on the lagged explanatory variables does not change the probability of $y_{it} = 1$, there is no true state dependence in the data
- If conditioning on the lagged explanatory variables changes the probability of $y_{it} = 1$, there is true state dependence in the data