Econometric Methods for Panel Data

Based on the books by BALTAGI: *Econometric Analysis of Panel Data* and by HSIAO: *Analysis of Panel Data*

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

May 11, 2016

Outline

Binary dependent variables

If the dependent variable y is binary (0/1) and some covariates are continuous, the customary interpretation is a model of the basic form (pooled)

$$P(y_{it} = 1) = F(x_{it}'\beta),$$

where $P(.)$ denotes a probability and $F(.)$ a cumulative distribution function for a symmetric distribution, often the normal distribution ('probit', $F(.) = \Phi(.)$) or the logistic distribution ('logit', $F(x) = e^x/(1 + e^x)$). These models essentially assume that $y_{it}^* = x_{it}'\beta + u_{it}$ is a latent (unobserved) variable, and that

$$y_{it} = \begin{cases} 1 & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \leq 0. \end{cases}$$

Logit and probit: what is special in a panel?

- Individual effects μ_i represent important characteristics that are ignored in cross-section analysis. For fixed T and $N \to \infty$, they cannot be estimated consistently (incidental parameters with fixed effects). Worse, this implies a severe bias for ML estimates for β;
- The main problem is that the model is highly nonlinear. Purging the variables by a within transformation does not work;
- With random effects, ML is no more a simple GLS procedure. RE with logit/probit typically implies numerical integration or simulation procedures.

Fixed-effects logit: conditional ML

Consider the fixed-effects *logit* model

$$P(y_{it} = 1) = F(x_{it}'\beta + \mu_i),$$

with $F(.)$ denoting the logistic c.d.f.

In this FE *logit* model, $\sum_{t=1}^{T} y_{it}$ is a *minimal sufficient statistic* for the incidental parameters μ_i. Maximization of the likelihood conditional on this statistic defines an estimator that does not depend on the effects.

This conditional ML estimator works well. For small T, it constitutes a simple generalization of usual cross-section logit estimation. For the probit case, the estimator is less convenient. For large T, the ML estimator is nearly unbiased.

A Hausman test for fixed effects

For $H_0 : \mu_i \equiv 0$, a logit estimator that does not account for any effects is consistent and efficient. It corresponds to 'pooled OLS' in the linear model. Conditional ML is consistent but inefficient.

With fixed effects $H_A : \exists i \quad \text{s.t.} \quad \mu_i \neq 0$, the basic logit estimator becomes inconsistent. Conditional ML is consistent.

This is the situation that admits a Hausman test. The statistic will be distributed χ^2_K under H_0.

An alternative idea: optimizing incidence

The *maximum score estimator* targets the incidence between predicted and observed events of $y = 1$. These estimators have been advocated by MANSKI and can be applied to panels. A problem is weighting, as there may be more/less cases with $y_{it} = 1$ than $y_{it} = 0$.

This estimation framework does not use statistics, for inference one may bootstrap after estimation.

Random effects probit models

Again, RE probit/logit models cannot be estimated by feasible GLS. There is no closed form for the ML estimate.

Random effects models for binary dependent variables are complex, as maximizing the likelihood requires the evaluation of multiple integrals. Gauss distributions have simple forms for some of these integrals, thus probit RE is preferred to logit RE. 'RE logit' often refers to assuming logistic ν_{it} and normal μ_i.

There are several techniques available to lighten the burden of multiple integration:

- Traditional quadrature techniques;
- The *method of simulated moments* (MSM) uses Monte Carlo simulation to evaluate the integrals, an otherwise common practice in mathematics.

No Mundlak in RE probit

By contrast to linear RE models, RE probit models will not become equivalent to FE probit if covariates and effects are correlated. Special routines for this case have been developed. Others recommend padding the model with additional regressors $x_i' = (x_{i1}', \ldots, x_{iT}')$.

Aspects of dynamic logit/probit panels

Generally, estimation of dynamic logit/probit panel model does not imply new technical difficulties. ML and approximations are known to perform well.

HECKMAN suggested using dynamic binary panels to discriminate between *spurious state dependence* and *true state dependence*.

In *spurious state dependence*, for example, lagged events of unemployment increase the probability of more such events because of individual characteristics that favor unemployment (the individual effects may be correlated with the observed regressors);

True state dependence is reflected in the coefficient on lagged unemployment, taking effects and autocorrelation into account.

Wooldridge's dynamic RE probit

WOOLDRIDGE's model assumes

$$\begin{aligned} & P(y_{it}|x_t, \mu_i, y_{i,0}, y_{i1}, \ldots, y_{i,t-1}) \\ = \quad & \Phi(x_{it}'\beta + \lambda y_{i,t-1} + x_i'\delta + \gamma y_{i,0} + \gamma_0 + \varepsilon_i), \end{aligned}$$

with $y_{i,0}$ denoting a pre-sample starting value and $\Phi(.)$ the normal c.d.f. Note that x_i contains all time points for the (strictly) exogenous covariate x, all of which are used as explicit regressors in probit estimation. The coefficient λ represents true state dependence. The part $x_i'\delta + \gamma y_{i,0} + \gamma_0 + \varepsilon_i = \mu_i$ represents the individual effect.

Tobit in panels

The most common model for censored data is the *Tobit model* that assumes

$$y_{it}^* = x_{it}'\beta + \mu_i + \nu_{it},$$

with $\nu_{it} \sim i.i.d.\ N(0, \sigma_\nu^2)$ and

$$y_{it} = \begin{cases} y_{it}^* & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \leq 0, \end{cases}$$

which is easily generalized to other forms of censoring. It is closer to regression than the binary model, and at least the (ML–based) FE Tobit is reported to perform well.

Extensions of the basic Tobit

- *Tobit II* assumes an equation-type rule for a second latent variable that determines whether y be observed. Such models with two individual effects, for each equation one, are often handled by complex estimators based on first-order differencing to eliminate the effects;
- *Dynamic* Tobit models suffer from poor performance of standard ML–based estimators and are often handled by GMM estimation.