9 Multilevel and Nonlinear Panel Data Models *

Olaf Hübler[1]

[1] Institute of Empirical Economic Research, University of Hannover
huebler@ewifo.uni-hannover.de

Summary: This paper presents a selective survey on panel data methods. The focus is on new developments. In particular, linear multilevel models, specific nonlinear, nonparametric and semiparametric models are at the center of the survey. In contrast to linear models there do not exist unified methods for nonlinear approaches. In this case conditional maximum likelihood methods dominate for fixed effects models. Under random effects assumptions it is sometimes possible to employ conventional maximum likelihood methods using Gaussian quadrature to reduce a T-dimensional integral. Alternatives are generalized methods of moments and simulated estimators. If the nonlinear function is not exactly known, nonparametric or semiparametric methods should be preferred.

9.1 Introduction

Use of panel data regression methods has become popular as the availability of longitudinal data sets has increased. Panel data usually contain a large number of cross section units which are repeatedly observed over time. The advantages of panel data compared to cross section data and aggregated time series data are the large number of observations, the possibility to separate between cohort, period and age effects. Furthermore, we can distinguish between intra- and interindividual effects and we can determine causal effects of policy interventions. New problems with panel data arise in comparison to cross section data by attrition, time-varying sample size and structural changes.

The modelling of panel data approaches distinguishes in the time dependence, in the assumptions of the error term and in the measurement of dependent variables. Due to the specific assumption consequences for the estimation methods follow. Apart from classical methods like least squares and maximum likelihood estimators, we find in panel data econometrics conditional and quasi ML estimators, GEE

*Helpful comments and suggestions from an unknown referee are gratefully acknowledged.

(generalized estimating equations), GMM (generalized methods of moments), simulated, non- and semiparametric estimators. For linear panel data models with predetermined regressors we can apply conventional techniques. The main objective is to determine and to eliminate unobserved heterogeneity. Two situations are distinguished: regressors and unobserved heterogeneity are independent or interact. Much less is known about nonlinear models. In many models not only simple first differences methods, but also conditional likelihood approaches fail to eliminate unobserved heterogeneity. As the specification of nonlinearity is often unknown, non- and semiparametric methods are preferred.

We distinguish between several types of panel data models and proceed from general to more specific models. The endogenous variable y_{it} can be determined by the observed exogenous time invariant ($\tilde{x}_i$) and time-varying (x_{it}) variables, unobserved time invariant regressors (α_{i*}) and a time-varying error term (u_{it}). The term $m(\cdot)$ tells us that the functional relation is unknown, i. e. nonparametric approaches are formulated, which may vary between the periods ($m_t(\cdot)$). If y_{it} is not directly determined by $\tilde{x}_i$, x_{it}, α_{i*} and u_{it}, but across an unobservable variable, we call this a latent model, expressed by $g(\cdot)$. Furthermore, this relation may be time-varying ($g_t(\cdot)$). This generalized time-varying nonparametric latent model can be presented by

$$y_{it} = g_t[m_t(\tilde{x}_i, x_{it}, \alpha_{i*}, u_{it})]\,. \tag{9.1}$$

Simplifications are possible and can lead to conventional linear models with individual effects and time varying coefficients.

9.2 Parametric Linear and Multilevel Models

Standard panel data analysis starts with linear models

$$y_{it} = x'_{it}\beta + u_{it} \qquad i = 1, ..., N \quad t = 1, ..., T\,, \tag{9.2}$$

where y is the dependent variable, x is a K×1 regressor vector, β is a K×1 vector of coefficients and u is the error term. The number of cross section observations is N and these units are repeatedly measured. When the cross sectional data are only pooled over T periods the coefficients can be estimated by OLS under classical assumptions about the error term. If an unobserved time invariant individual term α_i is incorporated, model (9.2) turns into

$$y_{it} = x'_{it}\beta + \alpha_i + \epsilon_{it} =: x'_{it}\beta + u_{it}\,. \tag{9.3}$$

The methods which are developed for this purpose depend on the assumptions of the error term, the regressand, the regressors and the coefficients of the model. Some panel data sets cannot be collected every period due to lack of resources or