Comments on

Wansbeek, Tom; Meijer, Erik

Published in:
TEST

DOI:
10.1007/s11749-007-0050-1

Publication date:
2007

Test (2007) 16: 33–36
DOI 10.1007/s11749-007-0050-1

DISCUSSION

Comments on: Panel data analysis—advantages and challenges

Tom Wansbeek · Erik Meijer

Published online: 6 March 2007

Professor Hsiao is to be congratulated with an excellent article surveying many, sometimes quite recent issues regarding panel data. He does so with great virtuosity, accuracy, and detail. Evidently, as he already mentions himself, no survey can do justice to the huge literature in the field. We would like to supplement his article with three points that we think are useful for applied researchers.

1 Structural equation models and software

In his Sect. 4, Professor Hsiao discusses several, mostly linear, models and their estimation. It is suggested that a considerable amount of technical analysis of specific cases is necessary to find a satisfactory estimator. However, many models, here and elsewhere, can be viewed as *structural equation models* (SEMs), for which widely available software can provide efficient estimators. SEM is a general framework for models with latent variables. There are several equivalent general model structures, of which the so-called LISREL model (after the LISREL program) is the most well

This comment refers to the invited paper available at:
http://dx.doi.org/10.1007/s11749-007-0046-x.

The authors thank Arie Kapteyn for his helpful comments.

T. Wansbeek (✉) · E. Meijer
Department of Econometrics, University of Groningen, Groningen, The Netherlands
e-mail: T.J.Wansbeek@rug.nl

E. Meijer
RAND Corporation, Santa Monica, CA, USA
e-mail: meijer@rand.ord

known:

$$\eta_i = \alpha + B\eta_i + \Gamma\xi_i + \zeta_i,$$
$$y_i = \tau_y + \Lambda_y\eta_i + \varepsilon_i,$$
$$x_i = \tau_x + \Lambda_x\xi_i + \delta_i,$$

where η_i is a vector of endogenous latent (unobserved) variables for observation i, ξ_i is a vector of exogenous latent variables, y_i and x_i are vectors of observed variables, ζ_i, ε_i, and δ_i are disturbances or errors, α, τ_y, and τ_x are vectors of intercept parameters, B and Γ are matrices of regression coefficients among the latent variables, and Λ_y and Λ_x are matrices of coefficients, called *factor loadings*, linking observed and latent variables. The first equation is a simultaneous equations regression model for the latent variables, whereas the second and third equations are factor analysis submodels, also jointly called the *measurement model*. Through the latter, errors-in-variables models fall into this class (e.g., Wansbeek 2001), but less tangible latent constructs like technical efficiency (e.g., Ahn et al. 2001) can also be tackled in this way. By imposing restrictions on the general SEM structure a wide variety of specific models can be generated, including dynamic ones. Already Jöreskog (1978) showed how panel data models can be written as SEMs. Nevertheless, the potential of SEMs in econometrics in general, and in panel data analysis in particular, remains underexploited to a surprising degree.

Originally, the SEM framework included only linear models, but state of the art software also includes facilities for ordinal dependent variables, in which y_i and x_i are replaced by the latent variables y_i^* and x_i^* and the relations between the starred and unstarred variables are of the familiar threshold type, as in ordinal probit models. Mixture models, stratified and clustered samples, and full information estimation with missing data have also been studied in the literature and are features of some of the more advanced programs. See, e.g., Wansbeek and Meijer (2000) and the references therein for an extensive discussion of this type of model. The most widely used SEM software packages are LISREL (http://www.ssicentral.com), EQS (http://www.mvsoft.com), Mplus (http://www.statmodel.com), Mx (http://www.vcu.edu/mx/), and Amos (http://www.spss.com/amos/).

The actual way in which a particular model can be written as a SEM is sometimes quite complicated algebraically. But, fortunately, applied researchers typically do not have to make these translations explicitly, because the software allows for a more intuitive model specification, either through almost literally writing the equations or through graphical user interfaces. This approach works best with large N and small T and can be applied with random individual effects and fixed time effects. Some programs also allow random coefficients (across individuals) in addition to random effects. The Stata package GLLAMM (http://www.gllamm.org) and the related book by Skrondal and Rabe-Hesketh (2004) are based on this idea of viewing random effects and random coefficients as latent variables.

2 Attrition

An important problem in panel data analysis, which is not explicitly mentioned in Professor Hsiao's article, is *attrition* or dropout of the study, so that for some respon-

dents only measures on the first few time points are available. This is most problematic if the probability of dropout is related to the variables of interest; as elsewhere in econometrics, selection on an endogenous variable induces inconsistency of estimators if no precaution is taken. So endogenous attrition must be explicitly modeled to obtain consistent estimators of the parameters of interest. The topic was pioneered by Hausman and Wise (1979), who investigated the potential effect, on the estimation of earnings functions, of attrition in the Gary income maintenance experiment. The attrition process they considered was generalized by Ridder (1990), who considered the possibility of attrition depending on lagged variables. Recent contributions include Hirano et al. (2001), who show the potential of using refreshment samples in order to distinguish between various forms of attrition, and Das (2004), who provides a nonparametric approach. A recent overview of the topic of incomplete panels, of which the attrition literature forms an important subset, is given by Baltagi and Song (2006).

3 Panel data on aging, retirement, and health

Professor Hsiao mentions some well-known panel data sets that have proven useful for economic analysis, most notably the NLS and the PSID. A fairly recent exciting development for economists, epidemiologists, sociologists, and researchers in many other fields is the emergence of a worldwide concerted effort of collecting panel data about aging, retirement, and health in many countries. This started with the Health and Retirement Study in the USA (HRS; http://www.rand.org/labor/aging/dataprod/, http://hrsonline.isr.umich.edu/), which is a bi-annual panel data set, with currently seven waves available (1992–2004). It was followed by the English Longitudinal Study of Ageing (ELSA; currently 2002 and 2004 available; http://www.ifs.org.uk/elsa/) and the Survey of Health, Ageing, and Retirement in Europe (SHARE; currently the first wave, 2004, available; http://www.share-project.org/), which covers 11 continental European countries, but more European countries, as well as Israel, will be added. Other countries are developing similar projects, in particular, several Asian countries.

These data sets are collected with a multidisciplinary view and thus contain lots of information about people of (approximately) 50 years and over and their households. Among others, this involves labor history and present labor force participation, income from various sources (labor, self-employment, pensions, social security, assets), wealth in various categories (stocks, bonds, pension plans, housing), various aspects of health (general health, diseases, problems with activities of daily living and mobility), subjective predictions of retirement, and actual retirement. These studies are set up such that the data are highly comparable across countries, so that, in addition to cross-sectional comparisons and comparisons over time, comparisons across countries can be made as well.

Using these data, researchers can study various substantive questions that cannot be studied from other (panel) studies, such as the development of health at older age, and the relation between health and retirement. Furthermore, due to the highly synchronized questionnaires across a large number of countries, it becomes possible to study the role of institutional factors, like pension systems, retirement laws, and social security plans, on labor force participation and retirement.

References

Ahn SC, Lee YH, Schmidt P (2001) GMM estimation of linear panel data models with time-varying individual effects. J Econom 101:219–255

Baltagi BH, Song SH (2006) Unbalanced panel data: a survey. Stat Pap 47:493–523

Das M (2004) Simple estimators for nonparametric panel models with sample attrition. J Econom 120:159–180

Hausman JA, Wise DA (1979) Attrition bias in experimental and panel data: the Gary income maintenance experiment. Econometrica 47:455–473

Hirano K, Imbens GW, Ridder G, Rubin DB (2001) Combining panel data sets with attrition and refreshment samples. Econometrica 69:1645–1659

Jöreskog KG (1978) An econometric model for multivariate panel data. Ann INSEE 30–31:355–366

Ridder G (1990) Attrition in multi-wave panel data. In: Hartog J, Ridder G, Theeuwes J (eds) Panel data and labor market studies. North-Holland, Amsterdam, pp 45–68

Skrondal A, Rabe-Hesketh S (2004) Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Chapman & Hall/CRC, Boca Raton

Wansbeek TJ (2001) GMM estimation in panel data models with measurement error. J Econom 104:259–268

Wansbeek TJ, Meijer E (2000) Measurement error and latent variables in econometrics. North-Holland, Amsterdam