Journal of Econometrics 184 (2015) 111–123

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data☆

Francesco Bartolucci [a], Federico Belotti [b], Franco Peracchi [b,c,*]

[a] University of Perugia, Italy
[b] University of Rome Tor Vergata, Italy
[c] EIEF, Italy

ARTICLE INFO

Article history:
Received 14 May 2013
Received in revised form
21 April 2014
Accepted 2 September 2014
Available online 16 September 2014

JEL classification:
C12
C33
C35

Keywords:
Generalized linear models
Longitudinal data
Fixed-effects
Hausman-type tests
Self-reported health
Health and Retirement Study

ABSTRACT

Recent literature on panel data emphasizes the importance of accounting for time-varying unobservable individual effects, which may stem from either omitted individual characteristics or macro-level shocks that affect each individual unit differently. In this paper, we propose a simple specification test of the null hypothesis that the individual effects are time-invariant against the alternative that they are time-varying. Our test is an application of Hausman (1978) testing procedure and can be used for any generalized linear model for panel data that admits a sufficient statistic for the individual effect. This is a wide class of models which includes the Gaussian linear model and a variety of nonlinear models typically employed for discrete or categorical outcomes. The basic idea of the test is to compare two alternative estimators of the model parameters based on two different formulations of the conditional maximum likelihood method. Our approach does not require assumptions on the distribution of unobserved heterogeneity, nor it requires the latter to be independent of the regressors in the model. We investigate the finite sample properties of the test through a set of Monte Carlo experiments. Our results show that the test performs well, with small size distortions and good power properties. We use a health economics example based on data from the Health and Retirement Study to illustrate the proposed test.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

A distinctive feature of panel data modeling is the treatment of unobserved heterogeneity, which is typically interpreted as the effect of unobservable factors on the outcome of interest. The simplest way of dealing with this form of heterogeneity is to include in the model time-invariant unobservable individual (i.e., unit-specific) effects. Assuming that these effects are constant over time, however, may be difficult to justify in certain applications. For example, Stowasser et al. (2011) convincingly argue that the dynamic pattern of self-reported health status can be better modeled by introducing a latent time-varying individual-specific health component. Clearly, biased parameter estimates may result if the individual effects are assumed to be time-invariant when in fact they are not. This is especially true in the case of long panels.

Linear panel data models with time-varying individual effects have been studied, among others, by Holtz-Eakin et al. (1988), Chamberlain (1992) and Ahn et al. (2001, 2013) in a large n and small T framework, and by Bai (2009), Bonhomme and Manresa (2012) and Kneip et al. (2012) in a large n and large T framework; see Ahn et al. (2013) for a detailed review of this literature.

On the other hand, only a few studies have tried to relax the assumption of time-invariant individual effects in nonlinear settings. For example, Heiss (2008) proposes a limited dependent variable model with time-varying effects which are assumed to follow a first-order autoregressive process with parameters that are common across sample units, while Bartolucci and Farcomeni (2009) present a multivariate extension of the dynamic logit model based on time-varying individual effects which are assumed to follow a time-homogeneous Markov chain for every sample unit. Although

☆ We thank Bill Greene, Pravin Trivedi, the Editor, an Associated Editor and two anonymous referees for their constructive suggestions. We also thank participants at the 2012 Annual Health Econometrics Workshop and the 2013 Italian Congress of Econometrics and Empirical Economics, and seminar participants at the University of Padua (Department of Statistical Sciences) and the Max Planck Institute of Economics for useful comments. We are grateful to Florian Heiss for allowing us to use his `arldv` Stata package.

* Correspondence to: Department of Economics and Finance, University of Rome Tor Vergata, via Columbia 2, 00133 Rome, Italy. Tel.: +39 06 7259 5934; fax: +39 06 2040 219.
E-mail address: franco.peracchi@uniroma2.it (F. Peracchi).

http://dx.doi.org/10.1016/j.jeconom.2014.09.002
0304-4076/© 2014 Elsevier B.V. All rights reserved.

the specification in Heiss (2008) is parsimonious (it uses only one additional parameter with respect to a standard random-effects model) and perhaps more easily justifiable in many applications, the discrete approach adopted by Bartolucci and Farcomeni (2009) results in a model that is more flexible and tends to fit the data better; see Bartolucci et al. (2011) for more detailed comments. Unlike the linear case, however, both approaches are computationally demanding. Further, the first approach requires strong parametric assumptions on the distribution of the random effects. Therefore, practitioners may find it useful to carry out a preliminary test for the presence of time-invariant unobserved heterogeneity before estimating this type of models.

In this paper, we present a simple test for the null hypothesis of time-invariant individual effects in generalized linear models (GLMs) for panel data. This class of models is quite broad and includes the Gaussian linear model and a variety of nonlinear models typically employed for discrete or categorical outcomes, such as logit, probit, Poisson and negative binomial regression models. The basic idea of the test is to compare two alternative estimators of the model parameters based on two different formulations of the conditional maximum likelihood method. It extends to GLMs with canonical link the suggestion by Wooldridge (2010, p. 325) of comparing the fixed-effects and the first-difference estimators as a way of formally testing violations of strict exogeneity.

Because our test is a pure specification test[1] based on the comparison of two alternative estimators of the same parameter vector, we refer to it as a Hausman-like test. Unlike the standard version of the Hausman test (Hausman, 1978), however, we compare estimators that are both inconsistent under the alternative. In fact, as pointed out by Ruud (1984), what matters for a specification test to have power is that it is based on estimators that diverge under the alternative (that is, their difference converges in probability to a nonzero limit), and that the sampling variance of their difference is sufficiently small. We show that, since our alternative estimators depend on different functions of the data, they generally converge in probability to different points in the parameter space when the individual effects are time-varying. Thus, our test has power against a variety of alternatives resulting in time-varying individual effects, such as omitted time-varying regressors, failure of functional form assumptions, and general misspecification of the systematic part of the model. Clearly, when the inconsistency of both estimators is the same, as in the case of a panel with only two waves, our test has no power.

It is worth emphasizing three features of our test. First, it does not require assumptions on the distribution of unobserved heterogeneity, nor it requires the latter to be independent of the regressors in the model. Second, it can be easily implemented using standard statistical software, as the test statistic is a simple quadratic form involving the difference of the parameter estimates and consistent estimates of their asymptotic variances and covariance.[2] Third, it does not require assumption on how time-invariant regressors enter the model, as the conditional likelihood function does not depend on them.

The remainder of this paper is organized as follows. Section 2 introduces our test in the case of a linear panel data model and analyzes its power properties in this simple setting. Section 3 presents our general statistical framework for the test. Section 4 investigates the small sample properties of the proposed test through a set of Monte Carlo experiments. Section 5 provides an empirical illustration based on data from the Health and Retirement Study. Finally, Section 6 offers some conclusions.

[1] A pure specification test one that places little structure on the alternative hypothesis; see Cox and Hinkley (1974) and Ruud (1984) for a detailed discussion.

[2] We implemented the proposed test in a series of R and Stata functions which are available from the corresponding author upon request.

2. The test in the case of linear panel data models

Consider a balanced panel where n units, drawn at random from a given population, are observed for T periods. For each sample unit $i = 1, \dots, n$, we denote by $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$ the vector of observations on the outcome of interest and by $\boldsymbol{X}_i$ the matrix of observations on k time-varying regressors. The tth row of $\boldsymbol{X}_i$ is denoted by $\boldsymbol{x}_{it} = (x_{it1}, \dots, x_{itk})'$.

Under the null hypothesis of time-invariant unobserved heterogeneity, our model for the data is the standard linear panel data model

$$y_{it} = \alpha_i + \boldsymbol{\beta}'\boldsymbol{x}_{it} + \epsilon_{it}, \quad i = 1, \dots, n, \ t = 1, \dots, T, \tag{1}$$

where α_i is a time-invariant unobservable individual effect and the error vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{iT})'$ is assumed to be mean independent of $\boldsymbol{X}_i$. Note that, at this stage, we make no other assumption on the ϵ_{it}, so they may be heteroskedastic or serially correlated for a given i. Under our set of assumptions, a consistent estimator of $\boldsymbol{\beta}$ is the fixed-effects (FE) estimator

$$\hat{\boldsymbol{\beta}}_1 = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i' \tilde{\boldsymbol{X}}_i \right)^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i' \tilde{\boldsymbol{y}}_i,$$

with $\tilde{\boldsymbol{X}}_i = \boldsymbol{L}\boldsymbol{X}_i$ and $\tilde{\boldsymbol{y}}_i = \boldsymbol{L}\boldsymbol{y}_i$, where $\boldsymbol{L}$ is the $T \times T$ symmetric idempotent matrix that transforms a vector into deviations from the time average of its elements. An alternative consistent estimator of $\boldsymbol{\beta}$ is the first-difference (FD) estimator

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i=1}^{n} \Delta\boldsymbol{X}_i' \Delta\boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \Delta\boldsymbol{X}_i' \Delta\boldsymbol{y}_i,$$

where $\Delta\boldsymbol{X}_i = \boldsymbol{P}\boldsymbol{X}_i$, $\Delta\boldsymbol{y}_i = \boldsymbol{P}\boldsymbol{y}_i$ and $\boldsymbol{P}$ is the $(T-1) \times T$ matrix that transforms a vector into first differences. Both estimators may be regarded as OLS estimators based on different transformations of the original data. Since we allow the ϵ_{it} to be heteroskedastic or serially correlated, neither estimator is efficient under the null hypothesis,[3] although both are consistent.

2.1. The test statistic

To test the null hypothesis of time-invariant unobserved heterogeneity we propose a Hausman-type test based on the difference $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2$ between the FE and the FD estimators. In fact, comparing the FE and FD estimators via a Hausman test is mentioned by Wooldridge (2010, p. 325) as one way to formally detect violations of strict exogeneity,[4] although he does not study in detail the power properties of the test and its possible generalization to nonlinear models.

Under the null hypothesis of time-invariant unobserved heterogeneity,

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{12}' & \boldsymbol{V}_2 \end{bmatrix} \right).$$

This implies that the asymptotic null distribution of $\sqrt{n}\hat{\boldsymbol{\delta}} = \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)$ is Gaussian with mean zero and variance $\boldsymbol{V}_0 = \boldsymbol{V}_1 + \boldsymbol{V}_2 - \boldsymbol{C}_{12} - \boldsymbol{C}_{12}'$. A consistent estimator of $\boldsymbol{V}_1$ is

$$\widehat{\boldsymbol{V}}_1 = \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i' \tilde{\boldsymbol{X}}_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i' \hat{\boldsymbol{\epsilon}}_{i1} \hat{\boldsymbol{\epsilon}}_{i1}' \tilde{\boldsymbol{X}}_i \right) \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i' \tilde{\boldsymbol{X}}_i \right)^{-1}, \tag{2}$$

[3] The FE estimator is more efficient when the errors in (1) are homoskedastic and serially uncorrelated, while the FD estimator is more efficient when they follow a random walk.

[4] It follows that our test has power against a broad class of alternatives resulting in endogeneity, such as time-varying individual effects, omitted time-varying regressors, failure of functional form assumptions and general misspecification of the systematic part of the model.

Download English Version:

https://daneshyari.com/en/article/5095898

Download Persian Version:

https://daneshyari.com/article/5095898

Daneshyari.com