Univariate versus multivariate modeling

of panel data [*]

Juan Carlos Bou

Department of Business Administration and Marketing

Universitat Jaume I

and

Albert Satorra

Department of Economics and Business

Universitat Pompeu Fabra. Barcelona

February 21, 2014

Running title: panel data modeling

Full Addresses:

Juan Carlos Bou

Department of Business Administration and Marketing

Universitat Jaume I

Avinguda Sos Baynat s/n

12071 Castelló, Spain

Tel. ++34-964387109

Fax. ++34-964728629

Email: bou@uji.es

Albert Satorra

Universitat Pompeu Fabra

Department of Economics and Business

Ramon Trias Fargas 25-27

08005 Barcelona, Spain

Tel. ++34 93 542 1758

Fax. ++34 93 542 1746

Email: albert.satorra@upf.edu

Abstract

Panel data can be arranged into a matrix in two ways, called 'long' and 'wide' formats (LF and WF). The two formats suggest two alternative model approaches for analyzing panel data: (i) univariate regression with varying intercept; and (ii) multivariate regression with latent variables (a particular case of structural equation model, SEM). The present paper compares the two approaches showing in which circumstances they yield equivalent—in some cases, even numerically equal—results. We show that the univariate approach gives results equivalent to the multivariate approach when restrictions of time invariance (in the paper, the TI assumption) are imposed on the parameters of the multivariate model. It is shown that the restrictions implicit in the univariate approach can be assessed by chi-square difference testing of two nested multivariate models. In addition, common tests encountered in the econometric analysis of panel data, such as the Hausman test, are shown to have an equivalent representation as chi-square difference tests. Commonalities and differences between the univariate and multivariate approaches are illustrated using an empirical panel data set of firms' profitability as well as a simulated panel data.

1 Introduction

Panel data are widely used in social and behavioral sciences, economics and other disciplines. Such data arises when a set of variables are observed on a sample of units (people, firms, households, geographical areas, etc.) at several time points. In an example from organizational research discussed in Section 3, a sample of firms is observed and their profitability and other accounting measures (expenditure on research and development (R&D), investment on advertising, etc.) are recorded for ten consecutive years. Researchers use this type of data to investigate the relationship of a dependent variable Y (e.g. profitability) on a set of covariates (firm's size, expenditure on R&D, etc.). We distinguish two types of covariates, those that are defined for each year within the firm and thus are time-varying (e.g., investment in R&D), and those that are defined for the firm and are time-invariant (e.g., the year the firm was founded). We denote these covariates as X and Z respectively.

A panel data set is a three-dimensional array with units (e.g., firms), variables (characteristics of the firm) and time (e.g., ten years) as its dimensions. It can be represented as periodic two-dimensional arrays (of firms $\times$ variables) which can be stacked vertically (long format, LF) or horizontally (wide format, WF) to form an enlarged two-dimensional array collecting the data for each period (year, quarter, month, etc.). In LF the same individual is repeated in several rows (rows can be reordered to arrange the data for an individual in consecutive records), while in WF rows correspond to distinct individuals and variables are repeated horizontally. These two ways of arranging the data are illustrated in Tables 1 and 2 and discussed later in this paper. Note that in LF, the unit of analysis is the observation of an individual at a point in time, while in the WF the sequence of observations of an individual at the various time points is regarded as an indivisible unit of analysis.

The LF and WF are associated with two apparently distinct modeling approaches. The LF calls for the univariate regression of Y on covariates X and Z with an intercept that varies with the individuals. The WF calls for multivariate regression models with latent variables (factor analysis or general forms of SEM) that allow structuring the temporal interdependence of the variables. The variable Y that is observed at several time points defines a vector $y = (Y_1, \ldots, Y_T)$ which is the dependent vector-variable of the multivariate regression. In the paper, the terminology LF and WF will be used a synonym of univariate and multivariate modeling respectively.

LF is quite common in econometrics. A univariate regression model with a unit specific intercept was proposed in Balestra and Nerlove (1966) and Wallace and Hussain

(1969). The varying intercept is either regarded as a set of fixed parameters (the so-called 'fixed-effect' model, 'FE'), or as a random sample from a distribution (the so-called 'random-effect' model, 'RE'). The 'FE' and 'RE' formulations lead to different estimators of the regression parameters for the time-varying covariates included in X; the regression parameters for the time-invariant covariates Z are not estimated in the 'FE'. For details of the LF econometric approach, see the chapters on panel data in classic econometric manuals (e.g., Green, 2003; Wooldridge, 2002) or the recent books devoted wholly to panel data, by Arellano (2003), Hsiao (2003) and Baltagi (2008).

The WF approach has also been discussed in econometrics. Chamberlain (1982) proposes the method of moments for panel data; Anderson and Hsiao (1982) fit a general dynamic model for panel data by maximum likelihood; Anderson (1987, 1989) give general results for *asymptotic robustness* in latent variable models that include the panel models as a special case. Satorra (2002) and Papadopoulos and Amemiya (2005) further extend results on asymptotic robustness to multiple group and correlated samples for panel data models. See also the recent work of Bai (2013), where factor analysis for dynamic panel data is proposed.

The WF perspective has a long tradition in the behavioral sciences. To our knowledge, the first reference to a proposal for WF analysis of panel data is Jöreskog (1978), where a SEM model for panel data is specified and an application to the estimation of a labor supply function is discussed. Jöreskog fits the model by maximum likelihood (ML) using one of the first versions of the computer program LISREL (Jöreskog and Sörbom, 1978). The model fitted was a dynamic panel data model with a random unit intercept. See Bollen and Curran (2006) and Montfort, Oud and Satorra (2007) for examples of panel data analysis using the WF approach in the behavioral sciences.

Comparison of the LF and WF approaches to panel data has also been investigated. Ejrnaes and Holm (2008) show that the univariate 'FE' and 'RE' specifications can be analyzed using SEM. Bollen and Brand (2010) apply the LF and WF perspectives to a National Longitudinal Survey of Youth, showing that the 'FE' and 'RE' regression analyses can be reproduced using SEM.

The present paper aims to enhance the above mentioned comparison of LF and WF approaches. Our aim is to give a perspective by which users of the two approaches can understand their commonalities and differences. For the comparison we define an assumption of time invariance (TI) of parameters (the TI assumption in Section 3.2) that ensures correspondence between the univariate and multivariate approaches. In contrast to the previous comparison, we will not require 'FE' and 'RE' formulations to be distinguished; to do this, we will need to exploit the above mentioned results on asymptotic

robustness. In addition, to better illustrate the commonalities and differences of the two approaches, and the conditions for their equivalence, we not only analyze an empirical data set, but we also use simulated data. The analysis with simulated data shows the critical role of the TI assumption for the validity of the univariate approach, and illustrates the potential of the multivariate approach to deal with deviations from TI (such as heteroscedasticity, autocorrelation, etc.).

The illustrations use `Stata` software for the univariate approach and `Mplus` software for the multivariate approach; both are widely used by researchers of the univariate and multivariate approaches, respectively.[1] To help practitioners, the code of all the analyses are provided in an Appendix.[2]

The structure of the paper is as follows. Section 2 describes the two data formats. Section 3 describes the models. Section 4 presents the analysis with empirical data. Sections 5 compares the approaches using simulated panel data. Section 6 concludes with a discussion. The computer codes used in the illustrations are provided in the appendixes.

2 LF and WF

Consider a panel data set comprising n units, T time points and a set of variables. The variables can be of two types: *time varying* (which are characteristics of the units that vary with time), and *time invariant* (unit characteristics that are constant across time); suppose we have p_1 time-varying variables and p_2 time-invariant ones.

The data set in LF is arranged into an $nT \times (p_1 + p_2)$ matrix, in which a row corresponds to the values of a unit recorded at a time point. In WF, we have an $n \times (Tp_1 + p_2)$ matrix, in which each unit is represented by a single row and time repeated variables produce different columns. See Table 1 for an illustration of LF with three variables X, Y and Z. The first and second columns of the table are the unit and time indexes, the additional three columns contain the values of variables X, Y and Z. Each row is a combination of unit of analysis (*i*) and time point (*t*), so that x_{it}, y_{it} and z_i are the values of variables X, Y and Z for unit i at time t. Note that variables X and Y are allowed to take different values for each combination i and t, while variable Z is constant within each individual i. In the illustration used in Section 4, investments in R&D (`r&d`) and advertising expenses (`adv`) are examples of time-varying variables, since they vary with unit (firm) and time (year), while age of the firm (`age`) (years in existence, taking 2002 as the reference year) is

---

[1]Other software packages such as `xtr` in R for LF, and `LISREL`, `EQS`, `sem` of R, `sem` of `Stata` for WF could equally have been used.

[2]To help in teaching, the simulated data set is also available in a web site.

an example of a time-invariant variable.

An essential feature of LF panel data is the likely association (or statistical dependence) among rows that belong to the same unit (e.g., the same firm). So, the LF panel data is likely to produce what is called positive *intra-unit correlation*, also known as *intra-class correlation (ICC)* among rows.

———————— Tables 1 and 2 around here ————————

In the case of WF all the observations for a single individual produce just a single row of data and repeated measures of the same variable give rise to new columns of the data matrix. Table 2 illustrates the WF for the same variables X, Y and Z of Table 1. The time-varying variable X, measured at four time points, produce the columns X_1, X_2, X_3, X_4; idem for variable Y, which produces Y_1, Y_2, Y_3, Y_4. The time-invariant variable Z, however, has just one column of the data matrix associated to it. In WF, rows correspond to different units (individuals) of analysis so typically they can be assumed to be statistically independent.

Note that for a fixed number of units n, when the number of time points T increases, LF increases the number of rows by a factor T (the number of rows is nT), although the number of columns remain constant. This contrast with WF, where the number of rows remain constant (n) while the number of columns increases by a factor of T; the number of columns is $Tp_1 + p_2$. It should be made clear, however, that LF and WF are just two equivalent forms of presenting the same data. Sofware routines are available to convert WF data to LF data, and vice-versa.[3]

The two alternative modeling approaches, the univariate and the multivariate, are described in the next section.

3   Models

3.1   Univariate (regression) approach

The basic framework of the univariate approach is the regression equation

$$y_{it} = \alpha_i + \beta x_{it} + \gamma z_i + \epsilon_{it}, \tag{1}$$

---

[3]For example, the so-called `reshape` commands available in both proprietary and free software, e.g. `Stata` and `R`, respectively.

where $i = 1, \ldots, N$ indexes individuals (units) and $t = 1, \ldots, T$ indexes time points. The scalar y_{it} and the vectors x_{it} $(1 \times k)$ and z_i $(1 \times q)$ are, respectively, the values of the dependent variable Y and (time-dependent and time-independent, respectively) covariates $x = (X_1, \ldots, X_k)$ and $z = (Z_1, \ldots, Z_q)$ for unit i at time t. Even though this is a multiple regression model, it is univariate in nature since it relates to a single dependent variable Y. The vectors β $(k \times 1)$ and γ $(q \times 1)$ are vectors of regression coefficients for time-varying and time-invariant covariates respectively, and α_i is an intercept parameter that is allowed to vary across units (this variation will be further discussed below). The error term ϵ_{it} is assumed to be centered and i.i.d. with respect to i and t, variance σ_ϵ^2, and independent of the Xs, Zs and the varying intercept parameter α_i.[4]

When $\alpha_i \equiv \alpha$, i.e., in the case of a non-varying intercept, standard OLS regression produces consistent estimates for β and γ. When the intercept α_i varies across units, OLS regression does not ensure consistency. In the econometric literature the variation of α_i is known as *unobserved heterogeneity* and two basic formulations have been given. In the so-called 'FE' formulation the α_i's are viewed as unobserved unit-characteristics that are non-stochastic and fixed over hypothetical replications of the data set. In that case, consistent estimation of γ is not possible, but OLS on transformed data may produce a consistent estimator of β; this is the so-called *within estimator* (WE) of β (described below). Alternatively, in the 'RE' formulation, the α_i's are viewed as i.i.d. realizations of a random variable, say $\underline{\alpha}$, of mean α and variance σ_α^2; in this case, consistent estimates of β and γ are obtained using mixed-effect regression under the assumption that $\underline{\alpha}$ is independent of the covariates. This leads to the mixed-effect estimator (MEE) of β and γ (described below). Even though the terms 'FE' and 'RE' abounds in the econometric literature on panel data – and in the software for the analysis of these models – they are surrounded by certain ambiguity. In classic manuals on econometric theory such as Green (2003) and Wooldridge (2002) we read, respectively,

> "It should be noted that the term 'fixed' as used here indicates that the term does not vary over time, not that it is nonstochastic, which need not be the case" (Green, 2003, p. 285)

> "In the traditional approach to panel data models, c_i [our α_i] is called a 'random effect' when it is treated as a random variable and a 'fixed effect' when it is treated as a parameter to be estimated for each cross section observation i. Our view is that discussions about whether the c_i should be treated as random variables or as parameters to be estimate are wrongheaded for micro-econometric

---

[4]In the multivariate formulation of this model (Section 3.2) we will be able to view the ϵ_{it}s as a variable which is distinct with t, thus with variance that may vary with t.

panel data applications." (Wooldridge, 2002, p. 251-252)

In this paper all variables (observable or latent) will be regarded on the same footing in terms of their 'randomnes' : either they vary across units, or across time, or across units and time. We use the term *'fixed'* synonymously with inferences conditional to the values of those variables. On this issue, Mundlak (1978) makes some clarifying observations:

"[Mundlak] proposes to remedy the situation by first indicating that the whole approach which calls for a decision on the nature of the effect, whether it is random or fixed, is both, arbitrary and unnecessary. Without a loss in generality, it can be assumed from the outset that the effects are random and view the FE inference as a conditional inference, that is, conditional on the effects that are in the sample. It is up to the user of the statistics to decide whether he wants inference with respect to the population of all effects or only with respect to the effects that are in the sample. This view unifies the two approaches in a well defined form and eliminates any arbitrariness in deciding about 'nature', in a way which is influenced by the subsequent choice of a 'desirable' estimator [...] when the model is properly specified, the GLSE [our MEE] is identical to the 'within' estimator. Thus there is only one estimator. The whole literature which has been based on an imaginary difference between the two estimators, starting with Balestra and Nerlove is based on an incorrect specification which ignores the correlation between the effects and the explanatory variables." (Mundlak, 1978, p. 70)

The next two subsections describes the standard econometric approaches for the univariate model (1) with varying intercept α_is.

3.1.1 The within estimator (WE)

A simple way to estimate (1) in the presence of varying intercept α_i is to use OLS for the regression model implied in the within-unit data. Considering the individual averages of each variable, the dependent variable and all the explanatory variables, and (1), we obtain the 'between-unit' regression

$$\overline{y_i} = \overline{x_i}\beta + \overline{\epsilon_i} \tag{2}$$

for $i = 1, \ldots, N$, where $\overline{y_i} = \frac{1}{T}\sum_t y_{it}$, $\overline{x_i} = \frac{1}{T}\sum_t x_{it}$, and $\overline{\epsilon_i} = \frac{1}{T}\sum_t \epsilon_{it}$. Subtracting (2) from (1) gives

$$y_{it}^* = x_{it}^*\beta + \epsilon_{it}^* \tag{3}$$

7

where $t = 1, \ldots, T$, $i = 1, \ldots, N$, $y_{it}^* = (y_{it} - \overline{y_i})$, $x_{it}^* = (x_{it} - \overline{x_i})$ and $\epsilon_{it}^* = (\epsilon_{it} - \overline{\epsilon_i})$. This is a regression equation with the same regression coefficient β as in (1) but free of the problem of a varying intercept. The OLS estimator of β in (3) is a consistent estimator and it is known as the within estimator (WE).

Instead of differentiation with respect to the mean of the unit, we could have taken first-difference of all the variables. The first difference estimator (FDE) of β is the OLS estimator in model (3) when $y_{it}^* = (y_{it} - y_{i(t-1)})$, $x_{it}^* = (x_{it} - x_{i(t-1)})$, $\epsilon_{it}^* = (\epsilon_{it} - \epsilon_{i(t-1)})$ and $t = 2, \ldots, T$. In the econometric literature on panel data, WE and the FDE are two alternatives for what is called the 'FE' estimator. One important property is that the WE (idem, the FDE) is a consistent estimator regardless of possible dependence between the varying intercept and covariates. The WE (idem, FDE), however, does not inform on γ, a parameter that has been eliminated from the model by the transformation of the data. This can be a drawback when the impact of time-invariant variables Zs on the dependent variable Y is also of interest. In addition, sufficient variation of the time-varying covariates Xs is required to avoid collinearity problems (too large standard errors) in the estimation of β. Compared to WE, the FDE has the disadvantage of inducing serial correlation on the error terms; in practice, however, both WE and FDE tend to produce very similar results and for this reason, for simplicity in the illustration, we only report the WE. [5] [6]

3.1.2 Mixed-effect estimator (MEE)

We now consider the case where the α_is of (1) are assumed to be i.i.d. realizations of a random variable, say $\underline{\alpha}$, of mean α and variance σ_α^2. Estimation is performed using classic techniques for mixed-effects regression where a regression model is considered with some of the regression coefficients assumed to be fixed (constant across units) and others are assumed to be random (varying across units). In panel data analysis, a simple case of mixed-effect regression is used, namely the case where β and γ are the fixed parameters and the intercept α_i is a random parameter. An essential assumption in mixed-effects regression is that the random parameters are independent of the covariates. Mixed-effects regression models have a long tradition in the biometric literature. Estimation methods have been established based on ML or re-weighted least squares (RWLS) (Laird and Ware, 1982). In the econometric literature, the estimators of β and γ arising from the mixed-effect regression approach are called 'RE' estimators; here, we denote them mixed-effects estimators (MEE).

---

[5] The software `Stata` uses the function `xtreg` with the option `fe`.

[6] A related approach when there are no time invariant variables is to use a dummy for each individual, known as the Least Squares Dummy Variable (LSDV) model (Greene, 2003).

The distinction between the WE (or FDE) and the MEE estimators of β is that the former ensures consistency regardless of the correlation between α_i and the time-varying variables X, while the later produces a consistent estimator only under uncorrelation. The econometric literature on panel data points to this uncorrelation as the deciding criterium for 'FE' (WE or FDE) *versus* 'RE' (MEE) estimators of β. The Hausman test (Hausman, 1978) was devised for the choice between the two alternative estimators. The test is constructed from the difference between the 'FE' and 'RE' estimators of β: the first being consistent but inefficient under the null; the second being efficient under the null, but inconsistent under the alternative.[7] Rejection by the Hausman test suggests that α_i is correlated with the covariates and thus estimation of β should be based on the 'FE' estimator. The Hausman test is available in the standard software for univariate analysis of panel data. [8] In the multivariate approach described in the next section, the Hausman test will be matched to a chi-square difference test.

3.2 Multivariate approach to panel data

We now describe the basic multivariate model for the WF data. The alternative to the univariate regression (1) is the following multivariate regression model

$$
\begin{aligned}
y_{i1} &= \mu_1 + \eta_i + \beta_1 x_{i_1} + \gamma_{21} z_i + \epsilon_{i1} \\
&\vdots \qquad \vdots \\
y_{it} &= \mu_t + \eta_i + \beta_t x_{it} + \gamma_{2t} z_i + \epsilon_{it} \qquad\qquad (4)\\
&\vdots \qquad \vdots \\
y_{iT} &= \mu_T + \eta_i + \beta_T x_{iT} + \gamma_{2T} z_i + \epsilon_{iT}
\end{aligned}
$$

for the T-dimensional vector variable $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$. Here y_{it}, x_{it}, z_i and ϵ_{it} are the same quantities as in (1), and $\mu_t + \eta_i$, β_t and γ_t, are the time-varying versions of α_i, β and γ of (1), respectively.

Figure 1 is a path model representation of this set of equations. The unobserved individual effect is represented by the latent variable η_i in the circle. The triangle represents the variable constant to one; arrows emanating from the variable constant correspond to the time varying overall means, μ_ts. The variables in the squares correspond to the observed variables, and are Ys, Xs and Zs. Arrows in the figure connecting variables in

---

[7]The null is the 'RE', the alternative is the 'FE'.

[8]e.g., in `xtreg` of `Stata` or `plm` of `R`.

the squares or in the circle correspond to regression coefficients (the dependent variable is the one indicated by the arrow); double-headed arrows (both solid and dashed) correspond to correlations among variables; errors terms in the equations are represented by the ϵ_{it}s single-headed arrows pointing to the dependent variables. The figure shows that the regression coefficients of the η_i to the Ys are set fixed to 1.

──────────── Figure 1 around here ────────────

The univariate and multivariate representations (1) and (4) coincide when the following time invariance assumption holds: [9]

Time-Invariance (TI): $\mu_t = \mu$, $\beta_t = \beta$, $\gamma_t = \gamma$, $\alpha_i = \mu + \eta_i$, and ϵ_{it} i.i.d. (across i and t)

When TI is not imposed, the multivariate representation (4) is more general than (1). For example, in (4), the ϵ_{it}'s are different variables across t, with their variances possibly varying with t (encompassing thus heteroscedasticity in panel data) and/or possibly correlated across time (encompassing autocorrelation in panel data). Note that when X and Z are absent from the model, then (4) is a classical simple model: a 'confirmatory' single-factor model (Jöreskog, 1969).

We now discuss the fixed versus random 'nature' (Munlaik's 1978 terminology) of the η_i's in (4). Assume first the case where TI holds. We have the two options:

Fixed: The η_is are fixed unit characteristics; in which case, (4) is equivalent to (1) with α_i's of a fixed 'nature'. No additional assumption (specification) needs to be made regarding possible dependency between α_i's and covariates.

Random: The η_i's are i.i.d. random realizations of a latent variable, say $\underline{\eta}$. In this case, we need to be specific about the possible correlation between $\underline{\eta}$ and covariates. We have the two options:

A: $\underline{\eta}$ is uncorrelated with X and Z;

B: $\underline{\eta}$ is possibly correlated with X and Z

In both cases, $\underline{\eta}$ is assumed to be uncorrelated with the error terms ϵ's of (4).

Specification (Random & (A)) corresponds to (1) with α_i random. In this case, SEM estimates of β and γ of (4) coincide with the MEE estimator that was discussed in the context of the univariate formulation (1).

---

[9]When in (1) we have a dummy variable for time, then the TI assumption would drop the restriction $\mu_t = \mu$ and $\alpha_i = \mu + \eta_i$ would be changed to $\alpha_i = \eta_i$

In the specification (Random & (B)), the multivariate model (4) is not identified without fixing the parameters of correlation between $\underline{\eta}$ and the vector Z (In Figure 1 this restriction is represented by $\phi(Z, \eta) = 0$ in the dashed double-headed arrow). If that restriction is added (the simple restriction is to insert zero correlation) the SEM estimator of β to be discussed below is equivalent to the WE (or FDE) developed for the univariate model of (1). We will see that this zero-correlation restriction does not change the inferences on SEM estimates of β. Not only do the SEM estimates of β (and the s.e. of estimates) not change with the value where we fit that correlation, but the estimates of β do not change even when we suppress Z from the model altogether.

We see that the multivariate model (4) reproduces both the MEE and the WE estimators depending on whether we use specification (A) or (B) respectively. Since specification (A) is nested within (B), a classic chi-square difference test in SEM can be developed as an equivalent to the Hausman test discussed for the univariate approach.

The SEM approach to the estimation and inference of the multivariate model (4) is as follows. Let S denote the covariance matrix of the observables variables, Σ the population probability limit of S, θ a vector that collects the independent parameters of the model, and $\Sigma = \Sigma(\theta)$ the covariance structure function implied by the model (4). An estimate $\hat{\theta}$ can be obtained by minimizing the discrepancy between S and $\hat{\Sigma} = \Sigma(\hat{\theta})$. Two widely used discrepancy functions (that correspond to weighted least squares (WLS) and ML estimation respectively) are

$$F_{WLS}(\theta) = (s - \sigma)'W(s - \sigma)$$

and

$$F_{ML}(S, \Sigma(\theta)) = \ln | \Sigma(\theta)S^{-1} | + \mathrm{tr}\left\{ S\Sigma(\theta)^{-1} \right\} - p$$

where p is the number of observed variables, s and σ are the vectors of non-redundant elements of the matrices S and Σ, and W is the (possibly sample dependent) weight matrix. In least squares estimation, W is the identity matrix. This approach produces parameter estimates that can be shown to be consistent and asymptotically normal, and inferences that are (asymptotically) free of distributional assumptions are available. Normal theory and distribution free (robust) s.e. as well as robust chi-square goodness-of-fit tests are available (for technical details, see Satorra and Bentler, 1990, 1994; and Satorra, 2002).

Asymptotic robustness (AR) theory for estimates and test statistics developed under the assumptions of normality of latent and independent variables of the model is available in Satorra (2002). This AR theory shows that, under certain conditions, inferences

regarding parameter estimates like the regression coefficients remain valid when latent variables deviate from the normality assumption, even when they are assumed to be fixed (i.e., when inferences are conditional to the values of those latent variables). Applied to our model, this AR theory concludes that the distribution of the SEM estimates for the vectors β and γ are robust not only to non-normality of $\underline{\eta}$ but even to the 'nature' (fixed or random) of the η_i's. The validity of statistical inferences for the regression coefficients of X and Z, regardless of whether the η_i's are viewed as fixed or random, has special relevance for our discussion. It sheds light on the traditional econometric debate on the choice between 'fixed' or 'random' effects models. We understand that the relevant distinction is not between 'fixed' or 'random', but between the specifications (A) or (B) pointed above (i.e., of whether or not we restrict the correlation among η_i's and the X). That is, without loss of generality, we specify (4) with η_i assumed to be a random variable with the specification of either (A) or (B).

3.2.1   Multivariate representation of the WE and MEE

The multivariate model (4) with additional restrictions produces estimates that can be seen to be equivalent to the WE and MEE (the 'FE' and 'RE' specifications, respectively, in econometric parlance). For both estimators we require the restriction TI of time invariance of parameters discussed above.

The multivariate results that are equivalent to the univariate MEE are obtained using the specification that η_i is uncorrelated with the X's and the Z's. This is represented in the path diagram shown in Figure 1. In this specification a dashed double-headed arrow (i.e., a covariance) between the time-varying and time-invariant variables and η_i is represented by a set of parameters ϕ. The specifications (4) and (1) are equivalent when the parameters ϕ are fixed to zero.

To reproduce the WE using the multivariate approach we now need to introduce possible correlation between the η_i and the time-varying covariates Xs. This is represented in the path diagram of Figure 1 but adding the possible correlation of η and Xs as parameters of the model. The correlation of η and Zs continues to be set to zero because of the need for identification of the model. The WE arises when η is freely correlated with the covariates and the residuals are uncorrelated over time. To replicate the WE (or FDE) of the regression coefficients of Xs, the time-invariant variables (the Z) can be kept in the model with the specification of being uncorrelated with η, or can simply be suppressed from the model. Whatever the approach, the estimates for the regression coefficients of the Xs (and their s.e.) remain unchanged. The same specification, but with zero correla-

tion between X_{it}s and η_i and allowing time invariant variables Zs in the model, match the MEE results.

The restrictions required for the multivariate model to produce results equivalent either to WE or MEE are over-identifying restrictions on (4). These restrictions are required for attaining comparable estimators with the univariate approach. General multivariate models that do not impose these restrictions are presented in the next section.

3.2.2 Multivariate dynamic panel data models

The multivariate specification (4) could be expanded to include an autoregressive structure to Y_t. The following equations express a dynamic (autorregressive) model specification for panel data. Possible correlation among the unobserved heterogeneity η_i and the time-varying variables X is also allowed.

$$
\begin{aligned}
y_{i1} &= \mu_1 + x_{i1}\gamma_{11} + z_i\gamma_{21} + \eta_i + \epsilon_{i1} \\
&\vdots \quad\quad \vdots \\
y_{it} &= \mu_t + \beta y_{it-1} + x_{it}\gamma_{1t} + z_i\gamma_{2t} + \eta_i + \epsilon_{it} \\
&\vdots \quad\quad \vdots \\
y_{iT} &= \mu_T + \beta y_{iT-1} + x_{iT}\gamma_{1T} + z_i\gamma_{2T} + \eta_i + \epsilon_{iT}
\end{aligned}
\tag{5}
$$

The path diagram representation of this model is shown in Figure 2. The model can be easily estimated using the SEM approach (see Bou and Satorra (2009a) for an application of this model to profitability data).

———————— Figure 2 around here ————————

The multivariate SEM specification can be shown to encompass the dynamic panel data model specifications of Anderson and Hsiao (1982) (also called 'serial correlation' and 'state dependence' models). We could also add an equation where the time-invariant variables have an effect not on the observable Ys but on the latent factor η_i. This is a model that assumes full mediation of η_i (the unobserved individual effect) in the relationship among Z and the dependent variable Y, (this type of model is investigated using SEM in Stoel, van den Wittenboer & Hox, 2004). In the context of firm profitability data, Bou and Satorra (2007, 2009b, 2010) use SEM to fit these general dynamic panel data models, in cases of single or multiple group data, and also hierarchical multi-level data. The path diagram of this model is represented in Figure 3.

We feel, however, that further extending the discussion on general dynamic models would take us away from the main purpose of the present paper, which is simply to compare the univariate and multivariate approaches to panel data in the context of a simple model. The next section discusses the two panel data alternatives using an empirical data set.

4 Illustration with firms' profitability data

The previous theoretical discussion will now be illustrated with an analysis involving empirical data. The application aims to study the relationship of a measure of a firms' profitability (variable Y, roa) with characteristics of the firm. We use a panel data set of firms with ten years of data. A brief description of the panel data used is presented below.

4.1 Data and variables

The data comes from the *Encuesta Sobre Estrategias Empresariales (ESEE)* (Survey on Business Strategies), a survey carried out annually by the Spanish Ministry of Industry to collect information on Spanish manufacturing firms. The data analyzed correspond to a period of 10 consecutive years, from 1993 to 2002. For the sake of this illustration, we only retained the firms with complete data for the whole period of observation. The panel data analyzed is composed of $n = 560$ firms and $T = 10$ periods with information on variables now described.

The dependent variable Y is *Return on Assets* (roa) (the ratio of annual net income to total assets), a measure of firm profitability that is widely used in organization research (see, for example, Schmalensee, 1985; Rumelt, 1991; McGahan and Porter, 1997, for exhaustive details on this variable). This is a time-varying variable in the panel. The following time-varying covariates (Xs) are used: *Capacity Utilization* (cu) (ratio of the utilized productive capacity to the total installed capacity of the firm); *R&D Intensity* (r&d) (the firms' annual R&D expenditures divided by annual sales); *Advertising Intensity* (adv) (ratio of the firm's total advertising expenses to sales); *Size of the firm* (size) (log of number of employees). Just one time-invariant covariate (Z) is included in the data analyzed: *Age of the firm* (age) (the number of years since the company was created, taking 2002 as the reference year).

Previous studies (e.g., Capon, Farley and Hoenig, 1990) documented the above covari-

14

ates (the *X*s and *Z*) to have a significant impact on firms' profitability. Specifically, `cu` has been reported to positively influence `roa` via the reduction of direct costs (see, e.g., Aaker and Jacobson, 1987); `adv` enhances the relationship between customers and the firm, and it is known to be positively associated with `roa` (e.g., Farris and Reibstein, 1979; Andras and Srinivasan, 2003; Phillips, Chang and Buzzell, 1983); `r&d` is expected to be positively related to firm profitability (Kirner, Kinkel and Jaeger, 2009; Thornhill, 2006), since investments in R&D lead to the development of new products with more customer appeal (Brown and Eisenhardt, 1995) and lower manufacturing costs (Sinclair, Keppler and Cohen, 2000). Finally, the literature has reported a positive association between `size` and profitability due to economies of scale and exploitation of market power by large firms (Buzzell, Gale and Sultan, 1975).

4.2   The two data layouts: LF and WF

Tables 3 and 4 show the two data layouts for a selection of three variables: namely `roa`, `cu` and `age`; the variables *Y*, *X* and *Z*, respectively. The dimension of the table for the complete LF data set would be 5600 × 5, where the first two columns indicate unit and time and the other three columns are the variables *Y*, *X* and *Z*; the number of rows (5600) is the result of 560 (the number of firms) multiplied by 10 (the number of time points). The dimension of the WF table for *Y*, *X* and *Z* data would be 560 × 22; one row for each unit (firm), and columns corresponding to: the first column is the index of the firm; columns 2-11 are the `roa` for 10 different years; columns 12 -21 are the `cu` for 10 different years; the last column of the table corresponds to the values of the time-invariant covariate `age`. The complete WF profitability data considered would have 52 columns, we would need to add 30 additional columns corresponding to the 10 measures of the time-variant variables `r&d`, `adv`, `size`. We see that Tables 3 and 4 show exactly the same data, i.e., the data corresponding to firms 1, 2,3, and 560 for years 1993, 1994 and 2002. As mentioned before, the LF and WF data are two equivalent representations of the same information.

———————— Tables 3 and 4 around here ————————

4.3   Results for the univariate approach

This section presents the results of the analysis of firms' profitability using the standard univariate panel data approach. The analysis is undertaken using regular options of `Stata`, a widely used software package for this type of analysis. We report the typical

'fixed' and 'random' effects specifications, i.e., respectively, the WE and MEE analyses discussed previously (FDE analysis produced results very similar to WE and are not presented here for the sake of brevity). These univariate results are shown in the first half of Table 5. The `Stata` code to produce these results is available in Appendix 1.

4.3.1 WE results

Table 5 shows WE, standard errors, and z-values for the regression coefficients of the variables included in the model. The estimates for regression coefficients of `cu` , `adv` and `size` are statistically significant. As expected the coefficients are positive in the case of `cu`; however, we found negative values for both `adv` and `size`. In addition, the estimate for `r&d` is negative, though non-significant. The table also shows estimates for the variance of the α_is and the variance of the disturbance term of the equation, that is $\sigma^2_{\alpha_i} = 34.34$ and $\sigma^2_{\epsilon_{it}} = 70.06$, respectively. [10] We also see the estimate of the intra-unit correlation coefficient (the `rho = 0.33`, which is the "intraclass correlation coefficient (ICC)" of `roa` within firms). [11] From `Stata` we also obtain (thought not shown in the table) the correlation between the intercept parameter and the linear combination of the covariates, which is –.0.42, a value that would tend to support the WE approach (which does not restrict that correlation to be zero).

————————— Table 5 around here —————————

4.3.2 MEE results

Table 5 shows the MEE results for two models, one that includes the time-invariant variable Z (that is, `age`) and another without Z. In both models, the random intercept is assumed to be uncorrelated with the covariates.

We observe substantial differences between the results of WE and MEE (with and without Z). There is a large difference between the estimates of the regression coefficient for `size`, –2.042 (WE) versus -0.894 (MEE), as well a noticeable difference in their s.e., MEE being the most accurate one. The variable `adv` (advertising intensity) that was significant with WE is not significant with MEE. The estimates of the regression coefficient of `cu` are, however, very similar. Estimates of the variance of the varying intercept and the

---

[10]The output of `Stata` reports the square root of those values (i.e. the standard deviations), but we transformed them to estimates of variances for the sake of comparison with the results we will obtain later using the multivariate model approach.

[11]Since that analysis is based on the assumption of a fixed random intercept, for the estimates of variances, intraclass correlation, and correlation between intercept and covariates, `Stata` does not produce s.e.

disturbance term are also shown in the table. When comparing MEE results with and without Z, we observe the same estimate of the mean of the random intercept, and only a slight difference in the estimates of the variance (variance associated to a model with Z is slightly smaller than without Z). The variance of α for MEE however is much smaller than the sampling variance of the intercept in the WE approach. The variance of the disturbance terms is similar in value among the three univariate models. We see that adding the time-invariant variable Z in the model does not change the MEE estimates of the other regression coefficients. In contrast to WE, with MEE we can obtain estimates for the regression coefficient of the time-invariant variables.

Several specification tests are available in the univariate approach. One specification test which is widely used in the classic univariate econometrics of panel data is the Hausman test. With our data, the Hausman test value is 34.70 for $df = 4$, to be compared with the χ_4^2 distribution.[12] The null hypothesis is rejected (p-value < 0.05) suggesting thus the need for the WE (see for example, chapter 13 in Green, 2003, for more details on using the Hausman test in panel data analysis). Since we observe significant differences between WE and MEE, the results of the Hausman test suggest support for the results of WE.

4.4   Results for the multivariate approach

The multivariate modeling approach discussed in Section 3 will now be illustrated with the firms' profitability data. We first analyze a multivariate model that reproduces the WE and MEE results shown in the first half of Table 5. The multivariate results are shown in the second half of the table. After that we will expand the analysis to a general multivariate SEM model. The results reported in this section were obtained using the `Mplus` software (Muthén and Muthén, 1998-2012), though other SEM software packages, e.g., EQS (Bentler, 2006) or LISREL (Jöreskog and Sörbom, 2006), could have been used. The code for the analysis using `Mplus` is reproduced in Appendix 2.

4.4.1   Multivariate equivalent to WE and MEE

The multivariate models to be analyzed are represented as a path diagram in Figure 1 of Section 3. The dependent variable `roa` (now a different variable for each time point) corresponds to the boxes Y_ts ($t = 1, 2, \ldots, 10$). Each of the time-varying covariates, `cu`, `r&d`, `adv` and `size`, corresponds to ten different variables; these are the boxes X_ts. The (possible vector) of time-invariant variable Z (in our application, `age`) is also represented in the path

---

[12]Appendix 1.C. shows the computation of this statistic on our data using `Stata`.

diagram. Double arrows connecting the Xs to η, a vector of covariances $\phi(X, \eta)$, represent parameters of correlation. The path diagram shows shows also the possible non-zero correlations among Z and η (parameters $\phi(Z, \eta)$) and among Z and X.

As explained in Section 3, to attain the equivalence between the multivariate and the univariate approach we require imposing the time invariance hypothesis TI, Z to be uncorrelated with η, and the specifications (A) or (B) regarding the correlation between the random intercept and covariates. The specification (B) where $\phi(X, \eta)$ is a vector of free parameters of the model, yields results equivalent to the univariate analysis WE; the specification $\phi(X, \eta) = 0$, yields results equivalent to MEE. The SEM results for the ML estimation method are shown in the second half of Table 5. [13]

Comparison of results shown in the first and second halves of Table 5 shows a clear match between univariate and multivariate results. Remember that numbers in the first half of the table were obtained using the classic methods of the univariate approach to panel data; the second half were obtained by completely different algorithms, the standard methods for SEM analysis. The coincidence is in both parameter estimates as well as s.e. (henceforth z-values). The multivariate SEM analysis adds a chi-square goodness-of-fit test of the model to the results of the univariate approach. For each of the four models in the second half of the table, the chi-square goodness-of-fit tests show a resounding rejection of the model. This is an important issue for the comparison of the univariate and multivariate perspectives. Note that no information on the goodness-of-fit of the model is given in regular econometric WE and MEE; the F-test that is usually shown is simply a test of the overall significance of the regression coefficients; this F-test gives no information on the validity of the model assumptions implicit in the univariate formulation (e.g., the TI and other assumptions).

In the multivariate model, chi-square goodness-of-fit comparison of nested models produces a variety of misspecification tests. One of these chi-square difference tests will be the equivalent to the Hausman test. Let H0 be the multivariate model with the covariances $\phi(X, \eta)$ set to zero, and H1 the multivariate model with those covariances set free. The difference of the chi-square goodness-of-fit test of the models H0 and H1 is the multivariate model equivalent of the Hausman test.

With our data the chi-square values of H0 and H1 are respectively 1303.88 (df =467) and 1267.72 (df = 463). The difference of these two values is $\triangle \chi^2 = 36.17$, with a difference of degrees of freedom $\triangle df = 4$ (4 is the number of time varying covariates in the model). This has an associated p-value of < 0.05 that rejects the null hypothesis of no correlation among unobserved heterogeneity and covariates. The value of this new test statistic is

---

[13]The `Mplus` code for the multivariate specifications of WE and MEE is displayed in Appendix 2.

close to the value of the Hausman test obtained in the univariate analysis, recall it was 34.70 with $df = 4$. [14] Other chi-square statistics that could match the plethora of specification tests in the univariate approach could easily be reproduced in the multivariate approach by using the simple principle of chi-square difference testing of nested models. For example, we could a chi-square different test for autocorrelation, or heteroscedasticity, or both, etc.

4.4.2   Multivariate dynamic panel data models

The multivariate approach encompasses a wider class of models than the simple regression with varying intercept. We illustrate this by fitting to the firms' profitability data the dynamic panel data models discussed in Section 3.2.2 and represented in Figures 2 and 3. In addition to the autoregressive structure on the Y_t, the models to be fitted include time-invariant covariates and correlation of the unobserved heterogeneity with the time-varying variables, permitting heteroscedastic error variances and regression coefficients varying over time. The results for two models are shown in Table 6[15]: the model where the Zs have a direct effect on the Y_t (in the table, 'Dynamic (Z direct)'), and the one where Z impacts on Y_t only indirectly throughout η_i (in the table, 'Dynamic (Z indirect)').

Table 6 shows a significant autoregressive coefficient β for the Y_t, a dynamic effect that was ignored in the previous analysis. The estimate of β is 0.183 (z-value is 11.90), so we should conclude that there is a highly significant dynamic component in the model. With regard to the other parameters of the model, the average (across time) of the regression coefficients shown in Table 6 are similar in values to the regression estimates of the static model shown in Table 5 for the corresponding parameters, [16] except for the regression estimate of `size`, which was highly significant for the static model and now is not significant. Thus, substantive interpretation of results may change across model specifications. The chi-square goodness-of-fit test statistics are χ^2 = 518.07 (df = 355) for 'Dynamic (Z direct)' and χ^2 = 529.76 (df = 364) for 'Dynamic (Z indirect)'. In both cases the goodness-of-fit would reject the model, despite the fit being much better than the one reported in Table 5) for the corresponding static model (χ^2 = 1192.49, df = 427 for SEM (with Z)). These are chi-square goodness of fit test based on normal theory. The values of the (robust) scaled chi-square goodness-of-fit test of Satorra and Bentler (1994) are χ^2 = 383.93, df =

---

[14]In the computation of this equivalence to the Hausman test, the covariances $\phi(X, \eta)$ in the H1 model are restricted to be equal over time. The test can be computed with Z either included or not in the model. Another version of the test could be based on the chi-square difference test of the model with $\phi(X, \eta)$ completely free in H1.

[15]The Mplus code for the analysis is displayed in Appendix 2.C.

[16]The table reproduces only averages across time for estimates of effects of covariates for the purpose of comparing the two tables

355 (p-value = 0.14) for the 'Dynamic (Z direct)', and $\chi^2 = 395.44$, df = 364 (p-value = 0.12) for the 'Dynamic (Z indirect)', an acceptable fit for both models. We simply note that the multivariate approach has an advantage over the univariate approach for assessing model modification, since it makes the restrictions on the model visible; Lagrange-Multiplier Tests (LMT) for overparameterizing restrictions are easily developed (these LMT are standard in SEM analysis).

In the analysis reported in Table 6 the regression parameters γ_{1t} and γ_{2t} (and the intercept) vary across years. The table shows the average of those values, that is the 'over-time averaged' effects, which were computed using the 'supplementary parameter approach' (SPA) of Bou & Satorra (2010). Standard SEM software such as Mplus or EQS have provision for estimating these nonlinear functions of the model parameters, so we are able to compute standard errors of the averaged effects using the above mentioned SPA. Table 6 also shows that the unobserved heterogeneity has slightly smaller variance than in the previous static FE and RE models. This result indicates that the estimated permanent differences across firms partially disappear when the autoregressive structure $\mathtt{roa}_{t-1}$ is introduced into the model.

——————— Table 6 around here ———————

5   Illustration using simulated data

To further illustrate the conceptual issues of this paper, we now apply the univariate and multivariate panel data analyses to simulated data. We consider data generated according to the path diagram of Figure 1 with particular values on the parameters that deviate slightly from the TI assumption mentioned in Section 3.2. The deviations considered are realistic in application and will serve to assess the impact of model misspecification in the analysis. The advantage of the simulated data (when compared to empirical data) is that we know precisely the true values of the parameters, so we can assess the validity of the different approaches. The true value of the parameters used in the data generating process are shown in the column 'Pop. Value' in Table 8, alongside estimation results for the multivariate model. For the sake of simplicity we only consider variables Y, X and Z, and four time points. We use a relatively large sample, $n = 3000$, to avoid clouding of the results by small sample size artifacts. [17]

Table 7 parallels with simulated data the results of Table 5 obtained with empirical data. Now there are just two covariates, one time-variant (X) and the other time-invariant

---

[17]The data analyzed is accessible at the following web site of one of the authors xxxx.

(Z). The model for the simulated data introduces a time effect on the level of Y with year 1 as the year of reference (the dummies `Dyear2, Dyear3, Dyear4` for years 2, 3 and 4, in the table). As in Table 5, Table 7 clearly shows that estimates obtained by the univariate approaches WE (no Z), ME (no Z) and MEE (with Z) are replicated by the multivariate approaches SEM (no Z), SEM (no Z) and SEM (with Z), respectively. Note that the estimate of the intercept μ_1 in the SEM approach equates to the estimate α in the univariate models, and that the differences of $\mu_t - \mu_1, t = 2, 3, 4$ in SEM coincide precisely with the estimates of the dummies. The estimates of the variances of the varying intercepts are also comparable across univariate and multivariate approaches, with the multivariate approach also producing an s.e. for this estimate of variance, as well as for the estimate of the ICC. [18] An important point to note is that the information on the chi-square goodness-of-fit test is available in the multivariate approach, but is unavailable in the univariate models (first half of the table). The chi-square values are very large and lead to clear rejection of the models. Note that rejection of the SEM models of the second half of the table implies rejection of the results obtained for the univariate models (i.e., the same results obtained in the first half of the table). We know the true value of the parameters, which do not satisfy the TI assumption, so we are not surprised to see rejection since that hypothesis is tested by the chi-square goodness-of-fit of the SEM approach. Note that for simplicity of this illustration, we only used one sample replicate. Given the high values of the chi-square goodness-of-fit test, it would be surprising to find a sample replicate where the models fit.

The large values of the chi-square goodness-of-fit statistics suggest the time invariance assumption TI implicitly assumed in the models of Table 7 should be relaxed. Table 8 shows estimation results without imposing the TI assumption, that is, we use the model of Figure 1 with parameters specific to each year. The analysis was carried out using `Mplus` (the code for this analysis is shown in Appendix 2C). Comparing estimates with the true value of the parameters used to generate the data set (the column 'Pop. Values' of the table), shows that that all the true values are within the corresponding 95% confidence intervals (i.e., estimate $\pm 2 \times$ s.e., regardless of whether one uses the normal theory or the asymptotic robust s.e.). We note that the normal-theory and robust s.e. (in parenthesis) are all very similar in value except for the estimate of $\sigma^2_{\eta_i}$ despite the fact that the data is non-normal since η_i was simulated from a highly skewed non-normal distribution (a χ^2_1). This, however, coincides with the theory of asymptotic robustness (AR) which guarantees correctness of inferences for all the parameters except for the estimates of the variance of the non-normal constituents of the model, and regardless of the assumption that η_i is fixed or random. The same theory of AR guarantees the asymptotic chi-squaredness of

---

[18]The latter is estimated using Bou and Satorra's (2010) SPA.

the goodness-of-fit test statistic; that is, the normal-theory based chi-square goodness-of-fit test is a valid test statistic to judge the adequacy of the model. We observe a chi-square value equal to 18.704 that has an associated p-value of 0.13 implying, thus, that the model is not rejected (at the usual 5% level, not even at the 10% level). The conclusion of the test coincides with what we know to be true about model since we have simulated data. Note the contrast in chi-square values when we compare the model with the TI assumption (the univariate regression models) and the model with TI relaxed.

————————— Table 7 and 8 around here —————————

6 Conclusion

Two perspectives for analyzing panel data have been discussed and compared. One perspective uses univariate regression and arranges the data in LF; the other perspective uses multivariate regression and arranges the data in WF. The univariate approach is very popular in econometric analysis of panel data; the second approach uses structural equation models (SEM) and is also common in the behavioral sciences. In an attempt to disentangle the commonalities and differences of the two perspectives, two data sets were analyzed with both methods and the results compared. One illustration used empirical data on firms' profitability, the second used simulated data.

In contrast to previous attempts to compare the econometric univariate regression with SEM (Ejrnaes and Hold, 2006; and Bollen and Brand, 2010), we have been able to pinpoint a fundamental assumption, what we called the TI assumption, under which the univariate and multivariate approaches produce the same estimates. We have argued that the TI assumption is likely to be violated in applications and that information on the validity of TI is lacking in the univariate perspective. The multivariate approach, however, produces information –in the form of a chi-square goodness-of-fit test of the model– on the validity of TI. We have shown that the multivariate perspective allows us to relax the TI assumption and assess fundamental hypotheses such as the un-correlation of the varying intercept and covariates. A chi-square difference test of two multivariate models has been shown to be equivalent to the Hausman test for panel data.

The multivariate approach has been shown to encompass panel data models that allow for dynamics, error in variables, heteroscedasticity of disturbance terms, and other deviations from the TI assumption.

With regard to the classic panel data choice between 'fixed' and 'random' specification

(i.e., WE versus MEE), we have shown that the key choice is whether parameters of correlations among the covariates and the varying intercept are fixed to zero or just estimated in a multivariate model approach. The illustration with empirical data showed that where the Hausman test gave a chi-square value of 34.87 (df = 4), the parallel chi-square difference test of two nested multivariate models gave a chi-square value of 36.17 (df = 4), very similar values despite the different formulae for the two statistics. The representation as chi-square difference testing of nested multivariate models of a variety of specification tests in the univariate approach, such as the test for homoscedasticity, autocorrelation, and others, may add conceptual simplicity for practitioners (in the multivariate perspective, whether the 'return' of η_i on Y_t is constant is also open to testing). We have shown that the varying intercept is just one example of a latent factor in multivariate analysis, and that its assumption of being 'fixed' or 'random' has no consequences for the inferences

A specific practical consequence of our paper is the diminishing role of the 'FE' approach (that is, the WE). The classic econometric argument for using the 'FE' approach is that it guarantees consistent estimates despite possible correlation of the unobserved individual effect (η_i) and covariates. We have seen that there is an alternative in the multivariate approach where the random intercept η_i is a factor (in the classic factor-analysis tradition) with possible non-zero correlation with covariates (see, specifically, the models SEM (no Z) and SEM (with Z) of Tables 5 and 7). The consistency claimed by the 'FE' approach is guaranteed in the multivariate approach when covariances among η_i and covariates are set as free parameters to be estimated. Recall that in the SEM approach the assumption of fixed versus random of the 'factor' η_i has no consequences on inferences of parameters of interest (as was concluded by the theory of asymptotic robustness).

A clear advantage of the multivariate perspective with η_i assumed random is that Zs do not need to be excluded from the model (as in the classic 'FE' approach) and that general methods of inference are readily available in standard SEM software (robust s.e. and test statistics, correction for clustering, weighted data, etc.; see, Muthén and Satorra, 1995). The multivariate approach facilitates also the generalization of the model to accommodate for measurement error and dynamics among variables, as exemplified in Subsection 4.4.2.

With regard to the empirical illustration, the multivariate perspective indicated a poor fit of both the WE and MEE regression models, and suggested introducing a dynamic component in the specification, as undertaken in Subsection 4.4.2. A classic limitation of the FE econometric approach is that the time constant covariates are excluded from the analysis, a limitation that is overcome with the multivariate approach that can encompass time-invariant covariates provided it is assumed these covariates are uncorrelated with the varying intercept. This can be seen in Tables 5 and 7 where cells are not available for

the WE (with Z).

We considered a data sample with complete data for all the firms in all the years of observation. If every unit is observed completely in every year, the panel is said to be *balanced*; otherwise, the panel is said to be *unbalanced*. For the unbalanced case, completeness of the WF can be maintained by introducing missing values into the rows (or individual cells) on unobserved items. Software for SEMs has now options for missing data which could be used for unbalanced panel data. The discussion of missing data, however, goes beyond the purpose of the present paper.

To conclude, this comparison of the univariate and multivariate perspectives for panel data should raise awareness among researchers of the commonalities of two widely used approaches that nowadays may be perceived as different. The multivariate perspective should bring conceptual simplicity to a variety of panel data models. The illustrations discussed should facilitate implementation of the methods for practitioners.

7   References

Aaker, D. A., & Jacobson, R. (1987). The role of risk in explaining differences in profitability. *Academy of Management Journal, 30*, 277-296.

Anderson, T. W. (1987). Multivariate linear relations. In T. Pukkila & S. Puntanen (eds). *Proceedings of the Second International Conference in Statistics*, pp 9-36. Tampere, Finland: University of Tampere.

Anderson, T. W. (1989). Linear latent variable models and covariance structures. *Journal of Econometrics, 41*, 91-119.

Anderson, T. W.,& Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics, 18*, 47-82.

Arellano, M. (2003) *Panel Data Econometrics*. Oxford, U.K.: Oxford University Press.

Andras, T. L., & Srinivasan, S. S. (2003). Advertising intensity and R&D intensity: Differences across industries and their impact on firm's performance. *International Journal of Business and Economics 2*, 167-176.

Bai, J. (2013). Fixed-effects dynamic panel models, a factor analytical method. *Econometrica, 81*, 285-314.

Balestra, P. & Nerlove, M. (1966). Pooling cross-section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica, 3*, 585-612.

Baltagi, B. H. (2008). *Econometric analysis of panel data*. Chichester: John Wiley and Sons.

Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review, 89*, 634-647.

Bentler, P.M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: John Wiley and Sons.

Bollen, K. A. & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 89*, 1-34.

Bou, J. C., & Satorra, A. (2007). The persistence of abnormal returns at industry and firm levels: Evidence from Spain. *Strategic Management Journal, 28*, 707-722.

Bou, J. C., & Satorra, A. (2009a). Multivariate Panel Data Modelling of Business Data: A SEM Application to the Study of Determinants of Firms' Profitability. *Unpublished manuscript*.

Bou, J. C., & Satorra, A. (2009b). Patterns of persistence of abnormal returns: A finite mixture distribution approach. In K. van Montfort, J. Oud, & A. Satorra (eds). *Longitudinal models in the behavioral and related sciences*. New Jersey: Lawrence Erlbaum Associates, Inc. Publishers.

Bou, J. C., & Satorra, A. (2010). A multigroup structural equation approach: A demonstration by testing variation of firm profitability across EU samples. *Organizational Research Methods, 13*, 738-766.

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies, 47*, 239-253.

Brown J. S., & Eisenhardt, K. M. (1995). Product development: Past research, present findings and future directions. *Academy of Management Review, 20*, 343-378.

Buzzell, R. D., Gale, B. T., & Sultan, R. G. (1975). Market share - A key to profitability. *Harvard Business Review, 53*, 97-106.

Capon, N., Farley, J. U. & Hoenig, S. (1990). Determinants of financial performance: A meta-analysis. *Management Science, 36*, 1143-1159.

Chamberlain, G, (1982). Multivariate regression models for panel data. *Journal of Econometrics, 18*, 5-46.

Ejrnaes, M. & Hold, A. (2006). Comparing Fixed Effecs and Covariance Structure Estima-

tors for Panel data. *Sociological Methods & Research, 35*, 35-61.

Farris, P. W., & Reibstein, D. J. (1979). How prices, ad expenditures, and profits are linked. *Harvard Business Review, 57*, 173-184.

Greene, W. H. (2003). *Econometric Analysis*. Upper Saddle River: Prentice Hall.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 41*, 733-750.

Hoechle, D. (2013). Robust standard errors for panel regressions with cross-sectional dependence. *The Stata Journal,7*, 281-312.

Hsiao, C. (2003). *Analysis of Panel Data*. New York: Cambridge University Press.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum-likelihood factor analysis. *Psychometrika, 34*, 183-202.

Jöreskog, K. G. (1978). An econometric model for multivariate panel data. *Annals de L'INSEE, 30-31.*

Jöreskog, K. G. & Sörbom, D. (1978). *LISREL IV. Estimation of linear structural equation systems by maximum likelihood methods: A FORTRAN IV program.* Chicago: International Educational Services.

Jöreskog, K. G. & Sörbom, D. (2006). *LISREL 8.8 for windows [Computer Sortware].* Lincolnwood, IL: Scientific Software Internacional, Inc.

Kirner, E., Kinkel, S., & Jaeger, A. (2009). Innovation paths and the innovation performance of low-technology firms - An empirical analysis of German industry *Research Policy, 38*, 447-458.

Laird, N. M. & Ware, J. H. (1982). Random-effects model for longitudinal data. *Biometrics, 38*, 963-974.

McGahan A. M., & Porter M. E. (1997). How much does industry matter, really? *Strategic Management Journal, 18*, 15-30.

McGahan A. M., & Porter M. E. (1999). The persistence of shocks to profitability. *Review of Economics and Statistics, 81*, 143-153.

Montfort, K., Oud, J. H., & Satorra, A. (eds). (2007). *Longitudinal models in the behavioral and related sciences*. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46*, 69-85.

Muthén, L. K. & Muthén, B. O. (1998-2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén

Muthén, L. K. and A. Satorra (1995), Complex Sample Data in Structural Equation Modeling. *Sociological Methodology , 25*. 267-316

Papadopoulos, S. & Amemiya, Y. (2005). Correlated samples with fixed and non normal latent variables. *The Annals of Statistics, 33*, 2732-2757.

Phillips, L. W., Chang, D. R., & Buzzell, R. D. (1983). Product quality, cost position and business performance: A test of some key hypotheses. *Journal of Marketing, 47*, 26-43.

Rumelt, R. P. (1991). How much does industry matter? *Strategic Management Journal, 12*, 167-185.

Satorra, A. (2002). Asymptotic robustness in multiple group linear-latent variable models. *Econometric Theory, 18*, 297-312.

Satorra, A. & Bentler, P.M. (1990). Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations, *Computational Statistics & Data Analysis , 10*, 235-249.

Satorra, A. & Bentler, P.M. (1994) Corrections to test statistics and standard errors in covariance structure analysis. In *Latent variable Analysis in Developmental Research* (p. 285-305) (A. van Eye and C.C. Clogg, Editors) , Sage Publications.

Schmalensee R. (1985). Do markets differ much? *American Economic Review, 12*, 341-351.

Sinclair, G., Keppler, S. & Cohen, W. (2000). What's experience got to do with it? Sources of cost reduction in a larger specialty chemical producer. *Management Science, 46*, 28-45.

Stoel, R. D., van den Wittenboer, G. & Hox, J. (2004). Including time-invariant covariates in the latent growth curve model. *Structural Equation Modeling, 11*, 155-167.

Thornhill, S. (2006). Knowledge, innovation and firm performance in high- and low-technology regimens. *Journal of Business Venturing, 21*, 687-703.

Wallace, T.D. and A. Hussain (1969). The Use of Error Components Modelsin Combining Cross Section With Time Series Data, *Econometrica, 37*, 55-68

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

Tables and Figures

Table 1: Long format, LF

Individual	Time	X	Y	Z
1	1	x_{11}	y_{11}	z_1
1	2	x_{12}	y_{12}	z_1
1	3	x_{13}	y_{13}	z_1
1	4	x_{14}	y_{14}	z_1
$\vdots$				
i	1	x_{i1}	y_{i1}	z_i
i	2	x_{i2}	y_{i2}	z_i
i	3	x_{i3}	y_{i3}	z_i
i	4	x_{i4}	y_{i4}	z_i
$\vdots$				
N	1	x_{N1}	y_{N1}	z_N
N	2	x_{N2}	y_{N2}	z_N
N	3	x_{N3}	y_{N3}	z_N
N	4	x_{N4}	y_{N4}	z_N

Table 2: Wide format, WF

| Individual | Time varying | | | | | | | | Time Invariant |
| | Time 1 | | Time 2 | | Time 3 | | Time 4 | | |
	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4	Z
1	x_{11}	y_{11}	x_{12}	y_{12}	x_{13}	y_{13}	x_{14}	y_{14}	z_1
$\vdots$									
i	x_{i1}	y_{i1}	x_{i2}	y_{i2}	x_{i3}	y_{i3}	x_{i4}	y_{i4}	z_i
$\vdots$									
N	x_{N1}	y_{N1}	x_{N2}	y_{N2}	x_{N3}	y_{N3}	x_{N4}	y_{N4}	z_N

Table 3: LF of variables `roa`, `cu` and `age`

firm	time	roa	cu	age
1	1993	6.23	60	29
2	1993	10.39	72	38
3	1993	-3.90	50	45
⋮	⋮	⋮	⋮	⋮
560	1993	1.52	70	10
1	1994	9.81	95	29
2	1994	9.17	76	38
3	1994	2.67	80	45
⋮	⋮	⋮	⋮	⋮
560	1994	4.67	75	10
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
1	2002	8.82	91	29
2	2002	-2.78	67	38
3	2002	4.37	85	45
⋮	⋮	⋮	⋮	⋮
560	2002	0.89	70	10

Table 4: WF for variables `roa`, `cu` and `age`

firm	roa1993	roa1994	...	roa2002	cu1993	cu1994	...	cu2002	age
1	6.23	9.81	...	8.82	60	95	...	91	29
2	10.39	9.17	...	-2.78	72	76	...	67	38
3	-3.90	2.67	...	4.37	50	80	...	85	45
⋮	⋮	⋮	...	⋮	⋮	⋮	...	⋮	⋮
560	1.52	4.67	...	0.89	70	75	...	70	10

Table 5: Univariate (WE and MEE) and multivariate (SEM) analyses: empirical data

Univariate Approach (Stata):

	WE (no Z)			WE (with Z)[†]			MEE (no Z)			MEE(with Z)		
	Coef.	s.e.	z[‡]	Coef.	s.e.	z[‡]	Coef.	s.e.	z	Coef.	s.e.	z
cu	0.069	0.013	5.55				0.071	0.011	6.51	0.071	0.011	6.51
r&d	-0.118	0.089	-1.33				-0.035	0.071	-0.50	-0.035	0.071	-0.50
adv	-0.466	0.109	-4.29				-0.014	0.064	-0.22	-0.014	0.064	-0.22
size	-2.042	0.504	-4.05				-0.894	0.152	-5.89	-0.894	0.152	-5.89
age	–	–	–				–	–	–	-0.020	0.011	-1.84
α	13.034	2.324	5.61				7.730	1.073	7.21	7.730	1.073	7.21
$\sigma^2_{\alpha_j}$	34.340						19.620			19.450		
$\sigma^2_{\epsilon_{it}}$	70.060						70.390			70.060		
rho (ICC)	0.330						0.220			0.220		

Multivariate Approach (Mplus):

	SEM (no Z)			SEM (with Z)			SEM (no Z)			SEM(with Z)		
	$\rho(X,\eta) \neq 0$			$\rho(X,\eta) \neq 0$ & $\rho(Z,\eta) = 0$			$\rho(X,\eta) = 0$			$\rho(X,\eta) = 0$ & $\rho(Z,\eta) = 0$		
	Coef.	s.e.	z[‡]	Coef.	s.e.	z	Coef.	s.e.	z	Coef.	s.e.	z
cu	0.069	0.013	5.55	0.069	0.013	5.55	0.072	0.013	5.71	0.071	0.012	5.65
r&d	-0.118	0.088	-1.33	-0.118	0.088	-1.33	-0.041	0.062	-0.66	-0.036	0.062	-0.57
adv	-0.466	0.109	-4.29	-0.466	0.109	-4.29	-0.036	0.057	-0.64	-0.015	0.060	-0.25
size	-2.042	0.504	-4.05	-2.042	0.504	-4.05	-0.973	0.145	-6.72	-0.895	0.148	-6.04
age	–	–	–	0.024	0.017	1.42	–	–	–	-0.020	0.011	-1.84
μ	13.034	2.334	5.58	12.298	2.053	5.99	7.408	1.061	6.98	7.732	1.074	7.20
$\sigma^2_{\eta_j}$	27.271	4.003	6.81	27.012	3.770	7.17	19.899	1.633	12.19	19.622	1.619	12.12
$\sigma^2_{\epsilon_{it}}$	69.978	1.394	50.20	69.978	1.394	50.20	70.276	1.402	50.14	70.309	1.402	50.13
ICC	0.280	0.030	9.34	0.279	0.028	9.780	0.220	0.020	14.28	0.220	0.020	14.13
χ^2	1183.3			1192.49			1294.87			1303.88		
d.f.	418			427			458			467		
p-value	0.00			0.00			0.00			0.00		

[†] Analysis not available in the WE approach
[‡] Ratio between estimate and s.e.

Table 6: Results for the dynamic panel data models: empirical data

	Dynamic (Z direct)			Dynamic (Z indirect)		
	Coef.*	s.e.*	z	Coef.	s.e.**	z
roa on:						
roa_{t-1}	0.183	0.015	11.90	0.184	0.015	11.92
cu	0.070	0.013	5.51	0.070	0.013	5.50
r&d	-0.117	0.096	-1.23	-0.115	0.096	-1.20
adv	-0.386	0.108	-3.57	-0.366	0.108	-3.40
size	-0.616	0.496	-1.24	-0.763	0.491	-1.55
age	0.003	0.015	-0.179	–	–	–
μ	5.180	2.047	2.53	5.629	2.038	2.76
η on:						
age	–	–	–	-0.013	0.009	-1.43
$\sigma^2(\eta_i)$	13.894	1.724	8.06	13.741	1.752	7.84
$\sigma^2(U_{it})$	68.072	1.462	46.55	68.255	1.467	46.53
$\bar{\chi}^{2\dagger}$	383.93			395.44		
df	355			364		
p-value	0.14			0.13		

* Averaged effects of estimates across the years, except for $\sigma^2(\eta_i)$

** s.e. computed using Bou and Satorra's (2010) supplementary parameter approach

† This is the robust Satorra and Bentler's (1994) scaled chi-square goodness of fit test

Table 7: Univariate (WE and MER) and multivariate (SEM) results: simulated data.

Univariate Approach (Stata):

	WE (no Z)			WE (with Z)[†]			MEE (no Z)			MEE(with Z)		
	Coef.	s.e.	$z^{\ddagger}$	Coef.	s.e.	z	Coef.	s.e.	z	Coef.	s.e.	z
x	2.067	0.023	89.42				2.215	0.023	97.21	2.096	0.022	95.41
z	–	–	–				–	–	–	2.541	0.043	59.33
α	2.489	0.033	76.07				2.490	0.066	37.74	2.543	0.512	49.71
Dyear2[†]	1.072	0.046	23.17				1.075	0.047	22.71	1.072	0.046	23.17
Dyear3	1.990	0.046	43.02				1.993	0.047	42.11	1.990	0.046	43.02
Dyear4	2.982	0.046	64.45				2.981	0.047	62.98	2.982	0.046	64.45
$\sigma^2_{\alpha_j}$	12.11						9.24			4.62		
$\sigma^2_{\epsilon_{it}}$	3.20						3.20			3.20		
rho (ICC)	0.79						0.74			0.59		

Multivariate Approach (Mplus):

	SEM (no Z) $\rho(X,\eta) \neq 0$			SEM (with Z) $\rho(X,\eta) \neq 0 \,\&\, \rho(Z,\eta) = 0$			SEM (no Z) $\rho(X,\eta) = 0$			SEM(with Z) $\rho(X,\eta) = 0 \,\&\, \rho(Z,\eta) = 0$		
	Coef.	s.e.	$z^{\ddagger}$	Coef.	s.e.	z	Coef.	s.e.	z	Coef.	s.e.	z
x	2.067	0.023	89.44	2.067	0.015	140.87	2.194	0.023	96.64	2.096	0.022	95.40
z	–	–	–	2.551	0.039	66.03	–	–	–	2.541	0.043	59.27
μ_1	2.489	0.070	35.76	2.543	0.046	55.85	2.490	0.069	36.11	2.543	0.051	49.68
μ_2	3.560	0.070	51.21	3.615	0.054	66.44	3.564	0.069	51.69	3.616	0.051	70.63
μ_3	4.479	0.070	64.42	4.533	0.046	99.47	4.483	0.069	65.01	4.534	0.051	88.57
μ_4	5.470	0.070	78.69	5.525	0.058	95.53	5.471	0.069	79.34	5.525	0.051	107.93
$\sigma^2_{\eta_j}$	11.291	0.316	35.75	4.655	0.381	12.21	11.043	0.309	35.72	4.649	0.141	32.90
$\sigma^2_{\epsilon_{it}}$	3.209	0.048	67.08	3.209	0.019	168.88	3.220	0.048	66.87	3.209	0.048	67.07
ICC	0.78	0.01	140.36	0.592	0.020	29.62	0.77	0.01	136.75	0.59	0.01	70.00
χ^2	10325.3			8540.1			11695.6			10444.6		
d.f.	19			22			23			26		
p-value	0.00			0.00			0.00			0.00		

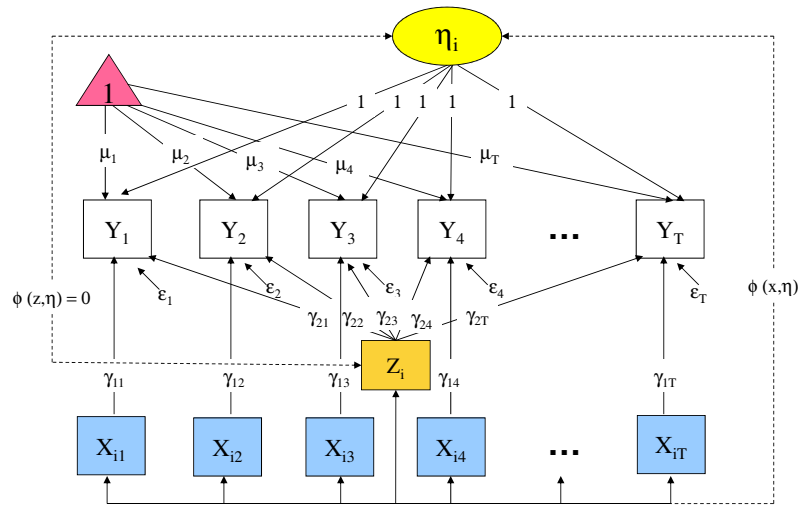[‡] Ratio between estimate and s.e.
[†] Dummy of year

Figure 1: Multivariate model representation of univariate panel data models. Depending on whether $\phi(X, \eta) = 0$ or are free parameters we reproduce the MEE and WE analyses respectively.

Table 8: Multivariate approach: general SEM model for simulated data.

|	Pop. Value	Estimate	s.e.	z
$t = 1$:				
x	2.000	2.002	0.024 (0.024)*	83.43
z	3.000	2.978	0.045 (0.043)	65.48
μ	2.600	2.552	0.045 (0.045)	56.75
$\sigma^2(U_{it})$	1.000	1.056	0.036 (0.037)	29.50
$t = 2$:				
x	0.000	-0.060	0.038 (0.038)	-1.61
z	3.000	3.043	0.053 (0.052)	56.93
μ	3.600	3.561	0.045 (0.045)	79.41
$\sigma^2(U_{it})$	1.000	1.024	0.035 (0.034)	29.18
$t = 3$:				
x	2.000	1.998	0.022 (0.023)	90.43
z	3.000	2.931	0.045 (0.042)	65.86
μ	4.600	4.539	0.045 (0.045)	101.37
$\sigma^2(U_{it})$	1.000	1.005	0.035 (0.035)	29.02
$t = 4$:				
x	4.000	4.024	0.022 (0.021)	183.69
z	3.000	2.964	0.044 (0.042)	66.92
μ	5.600	5.539	0.045 (0.044)	123.80
$\sigma^2(U_{it})$	1.000	0.996	0.034 (0.034)	28.89
$\sigma^2(\eta_i)$	5.120	5.006	0.136 (0.383)	36.84
χ^2		18.704		
d.f.		13		
p-value		0.13		

* In brackets are the robust s.e.
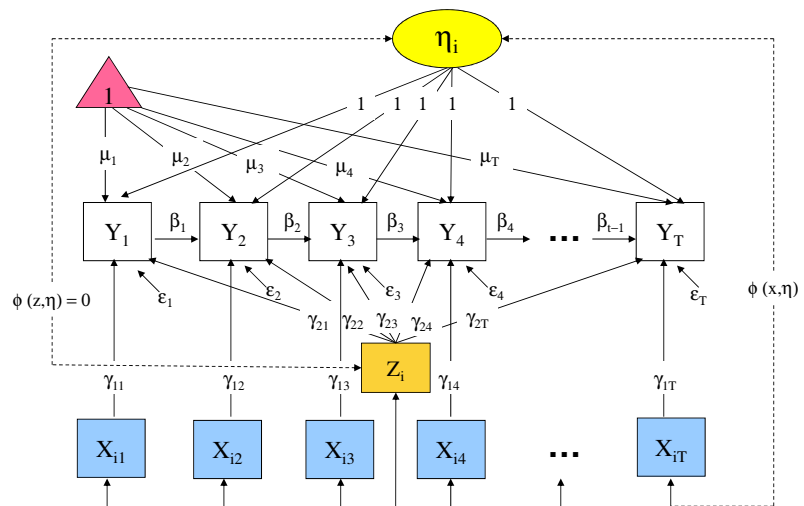
Figure 2: General Multivariate Model for Panel-Data: Dynamic (Z direct)
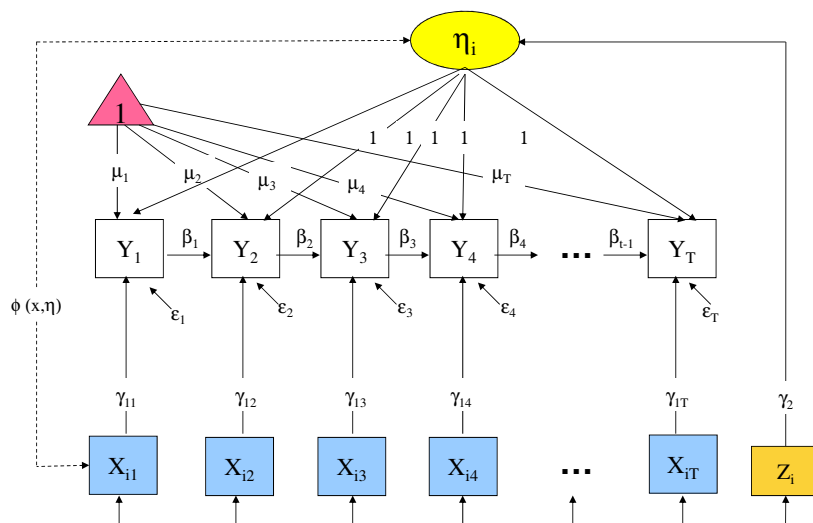
Figure 3: General Multivariate Model for Panel-Data: Dynamic (Z indirect)

Appendix

Appendix 1: `Stata` code for the univariate approach

1.A. WE approach

```
xtset firm
xtreg roa  cu r&d adv size, fe
```

1.B. MEE model

```
xtset firm
xtreg roa age cu r&d adv size, re
```

1.C. Hausman Test

```
quietly xtreg roa     cu r&d adv size, fe
estimates store fixed
quietly xtreg roa   age cu r&d adv size, re
estimates store random
hausman fixed random
```

Appendix 2: `Mplus` code for the multivariate approach

2.A. Multivariate WE

```
    TITLE: Multivariate fixed effect (FE) model
    DATA:
        FILE IS 'mydata.txt';
    VARIABLE:
       NAMES ARE firm roa1993-roa2002 age cu1993-cu2002
       r&d1993-r&d2002 adv1993-adv2002 size1993-size2002 ;
       USEVARIABLES ARE
       roa1993-roa2002 cu1993-cu2002 r&d1993-r&d2002
       adv1993-adv2002 size1993-size2002  ;
```

```
ANALYSIS:

      ESTIMATOR = ML;

MODEL:

  F BY roa1993-roa2002@1 ;

  roa1993-roa2002 PON cu1993-cu2002 (u)

                  r&d1993-r&d2002 (i)

                  adv1993-adv2002 (g)

                  size1993-size2002 (s);

  roa1993-roa2002 (resid);

  F (eta);

  [roa1993-roa2002] (mu);

  F WITH cu1993-cu2002 r&d1993-r&d2002 adv1993-adv2002

  size1993-size2002;

  roa1993-roa2002 WITH roa1993-roa2002@0;

MODEL CONSTRAINT:

  NEW (ICC);

 ICC = eta/(eta + resid);

 OUTPUT:  SAMPSTAT MODINDICES TECH1;
```

2.B. Multivariate MEE

```
TITLE: Multivariate random effect (RE) model

DATA:

      FILE IS 'mydata.txt';

VARIABLE:

   NAMES ARE firm roa1993-roa2002 age cu1993-cu2002

   r&d1993-r&d2002 adv1993-adv2002 size1993-size2002 ;

   USEVARIABLES ARE

   roa1993-roa2002 age cu1993-cu2002 r&d1993-r&d2002

   adv1993-adv2002 size1993-size2002  ;

ANALYSIS:

      ESTIMATOR = ML;
```

```
MODEL:

  F BY roa1993-roa2002@1 ;

  roa1993-roa2002 PON cu1993-cu2002 (u)

                 r&d1993-r&d2002 (i)

                 adv1993-adv2002 (g)

                 size1993-size2002 (s);

  roa1993-roa2002 ON  age (a) ;

  roa1993-roa2002 (resid);

  F (eta);

  [roa1993-roa2002] (mu);

  F WITH cu1993-cu2002@0  r&d1993-r&d2002@0 adv1993-adv2002@0

  size1993-size2002@0 age@0;

  roa1993-roa2002 WITH roa1993-roa2002@0;

MODEL CONSTRAINT:

  NEW (ICC);

 ICC = eta/(eta + resid);

 OUTPUT:  SAMPSTAT MODINDICES TECH1;
```

2.C. Multivariate dynamic panel data model (Z Direct)

```
TITLE: Multivariate dynamic panel data model

DATA:

      FILE IS 'mydata.txt';

VARIABLE:

   NAMES ARE firm roa1993-roa2002 age cu1993-cu2002

   r&d1993-r&d2002 adv1993-adv2002 size1993-size2002 ;

   USEVARIABLES ARE

   roa1993-roa2002 age cu1993-cu2002 r&d1993-r&d2002

   adv1993-adv2002 size1993-size2002  ;

ANALYSIS:

      ESTIMATOR = ML;

MODEL:
```

```
          F BY roa1993-roa2002@1 ;

          roa1993-roa2002 PON cu1993-cu2002 (u1-u10)

                               r&d1993-r&d2002 (i1-i10)

                               adv1993-adv2002 (g1-g10)

                               size1993-size2002 (s1-s10);


          roa1993-roa2002 ON  age (a1-a10) ;

          roa1994-roa2002 PON  roa1993-roa2001  (r1-r9) ;

          roa1993-roa2002 (v1-v10);

          F (eta);

          [roa1993-roa2002] (mu1-mu10);


          F WITH cu1993-cu2002 r&d1993-r&d2002 adv1993-adv2002

          size1993-size2002 ;

          F WITH age@0;

          roa1993-roa2002 WITH roa1993-roa2002@0;

MODEL CONSTRAINT:

NEW (Mean_cu Mean_r&d Mean_adv Mean_size Mean_age Mean_roa Mean_int);

Mean_uc = (u1 + u2 + u3 + u4 + u5 + u6 + u7 + u8 + u9 + u10)/10;

Mean_idv = (i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10)/10;

Mean_gpv = (g1 + g2 + g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10)/10;

Mean_size = (s1 + s2 + s3 + s4 + s5 + s6 + s7 + s8 + s9 + s10)/10;

Mean_age = (a1 + a2 + a3 + a4 + a5 + a6 + a7 + a8 + a9 + a10)/10;

Mean_Roa = (r1 + r2 + r3 + r4 + r5 + r6 + r7 + r8 + r9)/9;

Mean_int = (mu1 + mu2 + mu3 + mu4 + mu5 + mu6 + mu7 + mu8

            + mu9 + mu10)/10;

 OUTPUT:  SAMPSTAT TECH1;
```

2.C. Multivariate dynamic panel data model (Z Indirect)

```
TITLE: Multivariate dynamic panel data model

DATA:

     FILE IS 'mydata.txt';
```

```
VARIABLE:

   NAMES ARE firm roa1993-roa2002 age cu1993-cu2002

   r&d1993-r&d2002 adv1993-adv2002 size1993-size2002 ;

   USEVARIABLES ARE

   roa1993-roa2002 age cu1993-cu2002 r&d1993-r&d2002

   adv1993-adv2002 size1993-size2002  ;

ANALYSIS:

      ESTIMATOR = ML;

MODEL:

  F BY roa1993-roa2002@1 ;

  roa1993-roa2002 PON cu1993-cu2002 (u1-u10)

                      r&d1993-r&d2002 (i1-i10)

                      adv1993-adv2002 (g1-g10)

                      size1993-size2002 (s1-s10);


  roa1993-roa2002 ON  age (a1-a10) ;

  roa1994-roa2002 PON  roa1993-roa2001  (r1-r9) ;

  roa1993-roa2002 (v1-v10);

  F (eta);

  [roa1993-roa2002] (mu1-mu10);


  F WITH cu1993-cu2002 r&d1993-r&d2002 adv1993-adv2002

  size1993-size2002 ;

  F ON  age;

  roa1993-roa2002 WITH roa1993-roa2002@0;

MODEL CONSTRAINT:

NEW (Mean_cu Mean_r&d Mean_adv Mean_size Mean_age Mean_roa Mean_int);

Mean_uc = (u1 + u2 + u3 + u4 + u5 + u6 + u7 + u8 + u9 + u10)/10;

Mean_idv = (i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10)/10;

Mean_gpv = (g1 + g2 + g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10)/10;

Mean_size = (s1 + s2 + s3 + s4 + s5 + s6 + s7 + s8 + s9 + s10)/10;

Mean_age = (a1 + a2 + a3 + a4 + a5 + a6 + a7 + a8 + a9 + a10)/10;

Mean_Roa = (r1 + r2 + r3 + r4 + r5 + r6 + r7 + r8 + r9)/9;
```

```
Mean_int = (mu1 + mu2 + mu3 + mu4 + mu5 + mu6 + mu7 + mu8

         + mu9 + mu10)/10;

OUTPUT:  SAMPSTAT TECH1;
```