Univariate Versus Multivariate Modeling of Panel Data: Model Specification and Goodness-of-Fit Testing

Juan Carlos Bou¹ and Albert Satorra²,³

Abstract
Two approaches are commonly in use for analyzing panel data: the univariate, which arranges data in long format and estimates just one regression equation; and the multivariate, which arranges data in wide format, and simultaneously estimates a set of regression equations. Although technical articles relating the two approaches exist, they do not seem to have had an impact in organizational research. This article revisits the connection between the univariate and multivariate approaches, elucidating conditions under which they yield the same—or similar—results, and discusses their complementariness. The article is addressed to applied researchers. For those familiar only with the univariate approach, it contributes with conceptual simplicity on goodness-of-fit testing and a variety of tests for misspecification (Hausman test, heteroscedasticity, autocorrelation, etc.), and simplifies expanding the model to time-varying parameters, dynamics, measurement error, and so on. For all practitioners, the comparative and side-by-side analyses of the two approaches on two data sets—demonstration data and empirical data with missing values—contributes to broadening their perspective of panel data modeling and expanding their tools for analyses. Both univariate and multivariate analyses are performed in Stata and R.

Keywords
structural equation modeling, longitudinal data analysis, quantitative research, multiple regression, multivariate analysis

Panel data are widely used in social and behavioral sciences such as economics, organizational research, and applied psychology. They offer researchers theoretical and methodological advantages

¹Department of Business Administration and Marketing, Universitat Jaume I, Castellón, Spain
²Department of Economics and Business, Universitat Pompeu Fabra, and Barcelona GSE, Spain
³BI Norwegian Business School, Oslo, Norway

Corresponding Author:
Juan Carlos Bou, Department of Business Administration and Marketing, Universitat Jaume I, Avinguda Sos Baynat s/n, 12071 Castelló, Spain.
Email: bou@uji.es
over cross-sectional data: Panel data allow the analysis of dynamic effects (effects among variables over time; Bergh, 1993; Greve & Goldeng, 2004; Pitaru & Ployhart, 2010); provide a way to control for unmeasured stable variables (the so-called unobserved heterogeneity in the econometric literature); and can address some sources of endogeneity of regressors (Hamilton & Nickerson, 2003) that prevent estimates from being interpreted as causal effects (Antonakis, Bendahan, Jacquart, & Lalive, 2010).

In the panel data literature two apparently distinct modeling approaches are in use: the univariate approach, where observations are stacked across units and time (so a unit generates as many rows as time points observed in the panel) and just one regression equation is estimated; and the multivariate approach, where each unit generates a single row of data (and each time-varying variable generates a different column for each time point) and a set of regression equations—one for each time point—is estimated simultaneously.

The univariate approach to panel data is extensively described in classic econometric manuals (e.g., Arellano, 2003, chaps. 2-3, 7-8; Baltagi, 2013, chaps. 2-5; Greene, 2003, chap. 13; Hsiao, 2014, chaps. 3-4; Wooldridge, 2013, chap. 14). It is also present in behavioral sciences and organizational research (Bliese & Ployhart, 2002; Holcomb, Combs, Sirmon, & Sexton, 2010; Hom & Haynes, 2007; Pedhazur, 1982; Rabe-Hesketh & Skrondal, 2012, chaps. 3-6; Snijders & Bosker, 2012, chap. 12). It encompasses both static models such as regression with random intercept, random coefficients models, growth models, and also dynamic models with lagged effects of dependent and independent variables (e.g., autoregressive models, finite distributed lag models). The multivariate perspective has a long tradition in social and behavioral sciences (e.g., Chamberlain, 1982; Davidson, 1972; O. D. Duncan, 1969; Finkel, 1995, chaps. 2-3, 5-6; Finn, 1969; Jöreskog, 1978; Markus, 1979, chaps. 3; Montfort, Oud, & Satorra, 2007, chaps. 9-17; Rogan, Keselman, & Mendoza, 1979). It is implicit in a variety of models discussed in the literature such as repeated-measures ANOVA (Bergh, 1995; O’Brien & Kaiser, 1985), repeated-measures regression (Misangyi, LePine, Algina, & Goeddeke, 2006), autoregressive panel models (Kessler & Greenberg, 1981, chap. 7), cross-lagged regression models (Mayer, 1986; Rogosa, 1980), and latent growth curve (LGC) models (Bollen & Curran, 2006, chaps. 2-3, 5, 7; Chou, Bentler, & Pentz, 1998; T. E. Duncan & Duncan, 1995; McArdle & Epstein, 1987; Meredith & Tisak, 1990; among others).

Recent research has investigated the relation between the univariate and multivariate approaches to panel data (Allison, 2009; Allison & Bollen, 1997; Ejrnaes & Holm, 2006; Teachman, Duncan, Yeung, & Levy, 2001); more specifically, Bollen and Brand (2010) have made explicit the restrictions under which the multivariate formulation matches particular univariate panel data estimators. This literature, however, is a somewhat technical body of work addressed to users of structural equation modeling (SEM), and it is confined to social and political science methodologies; it does not seem to have had an impact in the organizational research arena. A literature review on some of the leading journals in organizational sciences (Academy of Management Journal, Journal of Applied Psychology, Journal of Management, and Strategic Management Journal) for the three years 2013, 2014, and 2015 returned 247 articles that explicitly involved the use of panel data. The vast majority of these articles, 216, used the univariate perspective for panel data analysis; specifically, they all estimated one regression equation based on the specification of a varying intercept and, in some cases, also lagged dependent and independent variables. Only 31 out of the 247 articles could be classified as using the multivariate perspective to panel data (cross-lagged models, LGC models, path analysis, and others). Clearly, the univariate perspective for panel data is dominant in organization research and related disciplines. None of the articles classified as univariate do not even mentioned the multivariate perspective as referred to by Bollen and Brand (2010) and others.¹

We believe researchers using the univariate approach could benefit from knowing about the tools for estimation and model search in panel data that are readily available in the multivariate setup and implemented in current software. For organizational researchers familiar with the multivariate models discussed in the psychology and management literatures (e.g., Bollen & Curran, 2006; Chan,
The present article discusses the two approaches to panel data side by side, focusing on the regression models with varying intercepts. In the comparison, we adopt a didactic, nontechnical approach, providing step-by-step guidance to applied researchers on how to estimate, compare, and modify the models in the univariate and multivariate approaches. Our aim is to help organizational researchers understand the comparative advantages and limitations of the two approaches, and to acquaint them with practical tools now available in the multivariate setup that they may not be aware of. We also feel that a more balanced view of the two approaches could help researchers widen their view of panel data modeling and broaden the range of tools for their analysis.

The article is structured as follows. Section 1 presents the long and wide format arrangement of panel data and their basic models. Section 2 presents the models: model conditions for the equality or divergence of the two approaches; and, by capitalizing on the multivariate approach, expansion of the family of models to include time-varying parameters, dynamics, and other extensions. Section 3 compares the two approaches using artificial data. In Section 4, through an illustration using empirical data, the two approaches tackle practical issues such as missing data and model specification. Section 5 concludes with a discussion.

Long and Wide Formats

In panel data, a set of variables are observed for each of \( i = 1, 2, \ldots, N \) units (individuals, firms) repetitively at \( t = 1, 2, \ldots, T \) time points. We have the time-varying variables (in our notation, \( y \) and the \( x_s \)), characteristics of the units that vary with time, and the time-invariant variables (the \( z_s \)), characteristics of the units that remain constant across time. Panel data can be arranged in two different formats. In the long (or stacked) format each row corresponds to the values of a unit recorded at one time point, so a unit generates as many rows as time points observed in the panel. In the wide format, each unit is represented by a single row and each time-varying variable generates a different column for each time point. Tables 1 and 2 illustrate the long and wide arrangements for panel data. The tables display the data of variables \( x, y \) and \( z \) observed at 4 time points for the units 1, 2, and 3000 of the data used in Illustration 1. Long and wide formats are two alternative representations of the same data (Tables 1 and 2 show the same numbers, but in different arrangements).

<table>
<thead>
<tr>
<th>Table 1. The Long Format.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>3000</td>
</tr>
<tr>
<td>3000</td>
</tr>
<tr>
<td>3000</td>
</tr>
<tr>
<td>3000</td>
</tr>
</tbody>
</table>
Software routines are available to convert wide format data to long format, and vice versa (in *Stata* and *R*, the command `reshape`).

In the long format, increasing $T$ increases the number of rows but not the number of columns. In the wide format, increasing $T$ increases the number of columns but not the number of rows. An advantage of the wide format is that it allows the researcher more flexible specifications of the temporal interdependence in the model. However, in some research contexts, such as when dealing with a large number of time points, and/or the timing of the measurement occasions varies across units, the wide format may be impractical because of the considerable increase in the number of columns (Singer & Willett, 2003, p. 21). It should also be mentioned that some temporal interdependencies, such as an autoregressive error term, can also be specified in the long format data (see, for example, Rabe-Hesketh & Skrondal, 2012, chap. 6).

An essential feature of the long format for panel data is the likely association (statistical dependence) among rows that belong to the same unit (e.g., the same firm). Therefore, the long format panel data are likely to produce what is called positive *intraclass correlation (ICC)*, which is the proportion of total variance of a variable that can be explained by group (unit) membership. One estimate of ICC is the ICC(1) (Bliese, 2000, pp. 354-355). Positive values of the ICC(1) indicate that, on average, scores of rows belonging to the same group are more similar than scores between groups (Kenny & Judd, 1986). In the wide format, however, since different rows correspond to the different units, they are assumed to be independent, and ICC(1) manifests as a correlation between the variables that store the different time observations.

These two formats give rise to the two approaches to panel data discussed in this article: univariate for the long format data; multivariate for the wide format. The basic models for these two approaches are discussed next.

**Models**

**Univariate Approach**

The basic model for the univariate approach is the following regression equation:

$$y_{it} = \alpha + \beta x_{it} + \gamma z_{i} + u_{i} + e_{it},$$

where, for unit $i$ at time $t$, $y_{it}$, $x_{it}$, and $z_{i}$ are the values of the dependent variable $y$ and the independent variables $x$ and $z$, respectively; $\beta$ and $\gamma$ are the regression coefficients; $\alpha$ is a constant intercept; $u_{i}$ is a centered unobserved variable varying only with the unit $i$, with variance $\sigma_{u}^{2}$; and $e_{it}$ is an equation disturbance term (that varies with unit and time), centered, with variance $\sigma_{e}^{2}$, and independent of $x$, $z$ and $u_{i}$. Considering $u_{i} \equiv \alpha + u_{i}$, equation (1) can be viewed as a regression equation with an intercept $\alpha_{i}$ that varies with the unit $i$. Even though equation (1) is a multiple regression model, it

<table>
<thead>
<tr>
<th>Unit</th>
<th>$x_{1}$</th>
<th>$y_{1}$</th>
<th>$x_{2}$</th>
<th>$y_{2}$</th>
<th>$x_{3}$</th>
<th>$y_{3}$</th>
<th>$x_{4}$</th>
<th>$y_{4}$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.34</td>
<td>-4.53</td>
<td>-0.85</td>
<td>-1.96</td>
<td>0.05</td>
<td>-2.17</td>
<td>1.48</td>
<td>3.65</td>
<td>-1.42</td>
</tr>
<tr>
<td>2</td>
<td>-0.85</td>
<td>0.42</td>
<td>0.61</td>
<td>3.41</td>
<td>0.06</td>
<td>4.39</td>
<td>0.83</td>
<td>6.86</td>
<td>0.22</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>1.00</td>
<td>1.79</td>
<td>0.06</td>
<td>1.27</td>
<td>0.25</td>
<td>3.14</td>
<td>-0.64</td>
<td>-3.42</td>
<td>0.82</td>
</tr>
</tbody>
</table>

**Table 2.** The Wide Format.
is univariate in nature since it expresses the variation of a single variable, the dependent variable $y$, conditional to the values of independent variables.

The varying $u_i$ represents the combined effect on $y$ of all unobserved unit-specific characteristics that are constant (or fixed) over time (Halaby, 2004). In empirical applications it is reasonable to assume that units are heterogeneous with respect to these unobserved characteristics. Suppressing the term $u_i$ in equation (1) is equivalent to merging $u_i$ with the disturbance term $e_{it}$. In this case, when $u_i$ correlates with the independent variables $x$ and $z$, we have what is called “omitted variable bias” (Antonakis et al., 2010) and the OLS estimates of $\beta$ and $\gamma$ fail to be consistent (an estimator is said to be consistent when it converges to the true value when sample size increases, and this is the basic criterion for an estimator to be considered useful).

The traditional econometrics of panel data (see, e.g., Wooldridge, 2013, chap. 14) use two type of estimators for the parameters in equation (1), the so-called “fixed effect” (FE) and “random effect” (RE). Before discussing these estimators, it is useful to clarify the terms “fixed” and “random” as used in mixed-effect regression (or hierarchical linear models, HLM) versus their use in the econometric literature on panel data. In mixed-effects regression (and HLM), a parameter is said to be “fixed” when it remains constant at all levels of the hierarchy, for example, the $\beta$ and $\gamma$ of our model; it is said to be “random” when it varies at a certain level of the hierarchy, for example, the intercept $a_i \equiv \alpha + u_i$ that varies with the units. In the econometric literature, however, FE and RE simply denote two type of estimators of $\beta$ in a model where $a_i$ varies across units; so, in both cases, using mixed-effects regression terminology, the intercept is not a fixed parameter! In the present article, we assume $a_i$ is an unobserved random variable varying across units. Two different models will emerge depending on whether $a_i$ is allowed to correlate or not (at unit level) with the time varying variables.

**The Fixed-Effect (FE) Estimator**

Under the presence of the varying intercept $a_i$, one simple procedure to estimate the regression coefficient $\beta$ of (1) consists in eliminating the between-unit variation and then applying OLS to the resulting within-unit variation. The specification (1) implies the between-unit regression

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \gamma z_i + u_i + \bar{e}_i \tag{2}$$

where the bar over a symbol denotes a cluster mean; thus, subtracting (2) from (1) gives

$$y^*_it = \beta x^*_it + e^*_it \tag{3}$$

where $y^*_it = (y_{it} - \bar{y}_i)$, $x^*_it = (x_{it} - \bar{x}_i)$ and $e^*_it = (e_{it} - \bar{e}_i)$. The transformed regression (3) has the same regression coefficient $\beta$ as in (1) and is free of the problem of a varying intercept. The OLS estimator of $\beta$ in (3) is known as the fixed effects (FE) estimator (also known as the within estimator). Adjustment for the degrees of freedom consumed by computing the averages is required in the OLS formulas for standard errors and test statistics (see Wooldridge, 2013, p. 484). Statistical software such as Stata (Stata Corp., 2015), for example, uses the `xtreg` command with the option `fe` to produce the FE estimator. One important property of the FE is that it is a consistent estimator of $\beta$ (Allison, 2009, p. 7). Simple differentiation of the data has removed all possible influences of the (stable) unit characteristics on the dependent variable, rendering $x^*$ exogenous with respect to $e^*_it$ (Antonakis et al., 2010).

There are however some disadvantages in the FE estimator. First, it does not provide estimators for $\gamma$, since the time-invariant variable $z$ has been eliminated from the equation (Antonakis et al., 2010). Second, as the between-unit variation has been eliminated by the transformation, there is a loss of efficiency (i.e., precision of the estimates), thus decreasing the likelihood of detecting a significant effect (Bliese & Ployhart, 2002), unless the within-unit variation dominates in the analysis. Finally, as in any regression analysis, consistency of the FE estimator of $\beta$ requires the
correct specification of the regression equation (with no omitted variables, lack of interaction terms, etc.) and—this is essential in our panel data setup—having invariance across time of all the regression parameters. Violation of the latter assumption is exemplified in Illustration 1 below.

**Random-Effect (RE) Estimator**

An alternative approach to estimate (1) is to assume the \( a_i \)s are realizations of a random variable of mean \( a \) and variance \( \sigma^2_a \) independent of the variables \( x \) and \( z \). This is a particular case of the so-called mixed-effects regression where \( \beta \) and \( \gamma \) are fixed parameters and the intercept \( a_i \) is a random parameter. Mixed-effects regression models have a long tradition in biometrics; estimation methods have been developed based on maximum likelihood (ML) or iterative generalized least squares (Croissant & Millo, 2008; Laird & Ware, 1982).

Despite the fact that \( \beta \) and \( \gamma \) are fixed parameters (the only random parameter is \( a_i \)), in econometrics the estimates of \( \beta \) and \( \gamma \) arising from mixed-effect regression are denoted “random effects” (RE) estimators. The `xtreg` command in Stata uses two implementations of an RE estimator: the feasible generalized least squares (GLS) estimator with quasi-groups mean centered data, based on estimated variance components, and the ML estimator of a random intercept model. These estimators are produced with the options `re` and `mle` respectively. In what follows, by (univariate) RE estimator we mean the latter estimator. This is the estimator that will be compared to a (multivariate) RE estimator discussed in the section on the multivariate approach.

When the random intercept \( a_i \) is independent of \( x \) and \( z \), the RE estimator is consistent for \( \beta \) and \( \gamma \), and the RE estimator of \( \beta \) is more efficient (i.e., it has smaller variance) than the FE estimator. The efficiency gain of the RE increases with the increase of the between-unit variation (Halaby, 2004, p. 520). A problem with the RE estimator, however, is that it may suffer from bias (i.e., its expected value deviates from the true value) when \( a_i \) correlates with the independent variables (Antonakis et al., 2010); in that case, the FE estimator would be preferred since it provides a consistent estimator for \( \beta \). The FE estimator of \( \beta \) (and its corresponding standard error) can also be obtained when the RE estimator is applied to a model where the cluster mean \( \bar{x}_i \) has been added as an independent variable (Mundlak, 1978); this is called the correlated random effect approach (see Wooldridge, 2013, sec. 14.3).

**Multivariate Approach**

The alternative to the univariate approach of (1) is the following set of \( T \) simultaneous regression equations:

\[
y_{it} = \alpha_t + \beta_t x_{it} + \gamma_t z_{it} + u_t + \varepsilon_{it}, \quad t = 1, \ldots, T, \tag{5}
\]

where the variables \( x_{it}, z_{it}, u_t \) and \( \varepsilon_{it} \) are the same as in equation (1), and \( \alpha_t, \beta_t \) and \( \gamma_t \) are the parameters \( \alpha, \beta \) and \( \gamma \) of equation (1), now possibly varying with \( t \). The disturbance term \( \varepsilon_{it} \) has variance \( \sigma^2_{\varepsilon}(t) \), allowed to be heteroscedastic (varying with \( t \)) and, unless explicitly stated otherwise, it is assumed to be uncorrelated across time. Note that the equivalent to the random intercept \( a_i \) of equation (1) is \( \alpha_t \equiv \alpha_t + u_t \), whose mean \( \alpha_t \) now possibly varies with \( t \).

Time-varying coefficients could be reproduced in the univariate approach by introducing time-dummy variables and their interactions with the \( x \)s (see Allison, 2009, p. 19), as well as heteroscedasticity and autocorrelated error terms (see Rabe-Hesketh & Skrondal, 2012, chaps. 3-6; Snijders & Bosker, 2012, chap. 12). In the illustration sections we will comment further on the univariate approach with time varying parameters.

To complete the specification, we need to introduce the parameters of covariances among \( x_1, \ldots, x_T, z, u_t \). It is assumed that \( u_t \) is uncorrelated with the time-invariant variables (the \( z \)s). This assumption of zero correlation is needed to have an identified model (a free correlation would
confound $u_i$ with $z$), though equivalent models (Bentler & Satorra, 2010) are obtained when fixing the covariance between $u_i$ and $z$ to any arbitrary value. An even more general multivariate model than (5) could be specified. For example, Bollen and Brand (2010) propose a model specification that replaces $u_i$ with $\lambda_i u_i$ where the $\lambda_i$ accounts for heterogeneity across time of the random intercept’s impact on the dependent variables. As in regular factor analysis, one $\lambda_i$ needs to be set to 1 for identification of the model. When $x$ and $z$ are absent from the model, then (5) is simply a “confirmatory” single-factor model (Jöreskog, 1969).

As noted already in Bollen and Brand (2010), the specification (5) is a particular structural equation model. A useful language to represent an SEM is the path diagram. Figure 1 shows the path diagram representation of (5). The unobserved unit effect is represented by the variable $u_i$ in the circle. The variables in the squares are the time-varying variables $y$ and $x$ and the time-invariant $z$. Arrows in the figure correspond to regression coefficients, and double-headed arrows (both solid and dashed) correspond to covariances among variables. The path diagram has set free the correlations among the unobserved $u$ and the time-varying independent variables $x$s, denoted in the path model by $\rho(x, u)$, with the correlation between $u$ and $z$ set to zero. Disturbance terms are represented by the $\varepsilon$s with single-headed arrows pointing to $y$s. Finally, the triangle is the constant to 1 and serves to represent the intercepts of the equations (arrows emanating from the constant to 1 correspond to the $\alpha$s). Figure 1 shows that the loadings of $u$ to $y$s are set to 1.

Maximum likelihood (ML) estimation (and weighted least squares, for different weight matrices) of (5), can routinely be obtained using regular software for SEM analyses, such as EQS (Bentler, 2006), Mplus (Muthén & Muthén, 1998-2012), LISREL (Jöreskog & Sörbom, 2006), sem in Stata, the lavaan package of R (Rosseel, 2012), and others. Nowadays, these software packages also offer the option of ML analysis with missing data (Arbuckle, 1996), which can deal with MAR-type informative missingness (Rubin, 1976; we use this estimation method in Illustration 2 to cope with unbalance in the panel and missing values, as will be clarified below). See Bollen (1989) for a general overview of SEM methods; and, for example, Satorra and Bentler (1990, 1994) for technical
details of estimation, model testing and model modification in SEM under general conditions on the
distribution of the data.

**A Restricted Multivariate Panel Data Model**

A more restricted specification of (5) is when we impose the following TI assumption:

**Time Invariance (TI):** *Invariance across time of the intercept, regression parameters and
variance of the disturbance terms:* \( \alpha = \alpha, \beta = \beta, \gamma = \gamma, \sigma_{e}^{2}(t) = \sigma_{e}^{2}. \)

Bollen and Brand (2010; see also Ejrnaes & Holm, 2006; Teachman et al., 2001) argue that the TI
assumption and the specification of the correlation between \( u_i \) and the independent variables \( x \)—
correlation set free, or restricted to be zero—make FE and RE estimates of parameters obtained
under specifications (1) and (5) (asymptotically) equal. Illustration 1 below explores this equiva-
lence in the context of a specific data set.

In some research scenarios, the restrictions imposed by the TI assumption may be questionable,
such as when the effect of an independent variable on the dependent variable increases (or decreases)
over time. For instance, in the organizational research context it is plausible that the impact of yearly
R&D investments on firm profitability differs for recession and expansion periods; or that the impact
of advertising investments decreases as the message filters through to current customers. Similarly,
homoscedastic error terms is not a suitable assumption when the dependent variable tends to become
either more or less variable over time, as in the case, for example, of analyzing the performance of
employees in a new job, which initially shows a high degree of variability but diminishes over time
as their expertise grows (Bliese & Ployhart, 2002), or in the study of firm profitability, whose
variation may diminish over time when, as consequence of competitive pressures, less profitable
firms exit the market. Therefore, the TI assumption is essential for the validity of inferences of the
univariate approach to panel data. Violation of this assumption can introduce severe bias in the FE
and RE parameter estimates. Tools for assessing the validity of the restrictions are important in
practice and will be further discussed below.

**Multivariate FE and RE Estimates**

The TI assumption does not fully determine the equivalence between estimates derived from the
univariate and multivariate approaches. We also need to specify the free or restricted nature of the
covariances between \( u \) and the independent variables \( x \)s (recall that the covariance of \( u \) with the \( z \)s
have already been set to zero). In the multivariate specification (5) we distinguish two cases:

(A): covariances of \( u \) with each \( x_1, \ldots, x_T \) set free

(B): covariances of \( u \) with each \( x_1, \ldots, x_T \) set to zero

Illustration 1 below (which uses demonstration balanced panel data) will show that the multi-
ivariate approach under hypotheses (A) or (B) exactly matches the (univariate) FE estimate of \( \beta \), or
the (univariate) RE estimate of \( \beta \) and \( \gamma \), respectively. This exact match also applies to the corre-
sponding standard error. \( ^{11} \) So inferences are asymptotically equivalent for the univariate and multi-
ivariate approaches for FE and RE estimates. In the present article, multivariate FE and RE estimates
refer to the estimates obtained under the specification (5), conditions (A) and (B), respectively, and
the TI assumption. A path diagram representation of both options is obtained when all the para-
eters in Figure 1 are set time-invariant and the covariance parameters \( \rho(x, u) \) are set free (FE) or
restricted to zero (RE).

Since the specification (B) is nested within (A), a SEM chi-square difference test for the two
models should serve the same purpose as the classic Hausman test (Hausman, 1978) common in the
univariate approach (Allison & Bollen, 1997; Bollen & Brand, 2010; Teachman et al., 2001).
A significant value of the test statistics will reject using the RE estimator (assumption B) in favor of the FE estimator (assumption A).

Both specifications (A) and (B) imply SEMs with a large number of overidentifying restrictions that may induce model misfit, not only the time-invariance constraints implied by TI, but the lack of dynamic components of the model (i.e., serial correlated disturbances, lagged dependent variables, etc.). The path diagram of Figure 1 clearly shows the missing paths between variables across different years; in SEM, these missing paths correspond to parameters that are set to zero.

**Dynamic Panel Data Models**

The multivariate specification (5) can be expanded very simply to the so-called dynamic panel data models where the lagged variable $y_{t-1}$ is included in the regression equation. The model is expressed as the following set of $T - 1$ simultaneous equations:

$$y_{it} = \alpha_t + \beta_t x_{it} + \gamma_t z_{it} + \delta_t y_{it-1} + u_t + e_{it}, \quad t = 2, \ldots, T,$$

where, as in the previous multivariate model (5), we allow time-varying parameters and covariances among the varying intercept and the time-varying variables $x$s and $y$ at time 1. The corresponding path diagram representation of this model is shown in Figure 2. Note the one-way arrows connecting the $y$s (dynamics in the model), and that one wave of data is lost in the analysis. A formal presentation of this model and the assumptions for estimating it can be found in Bollen and Brand (2010). This specification is a SEM that can be estimated using standard software.

This multivariate formulation considerably expands the type of dynamic panel models accessible to practitioners. Estimating models with autoregressive errors is often recommended in the univariate perspective (Bliese & Ployhart, 2002; Short, Ketchen, Bennett, & Du Toit, 2006). Halaby (2004) makes, however, a distinction between the dynamics introduced as an autoregressive structure on the $y$s, or simply as correlation among error terms. We could also have measurement error on the dependent variable, lagged effects of independent variables, multiple indicators, and so on. These models can be easily estimated in the same SEM setup. See Finkel (1995, chap. 5) for more examples of dynamic panel data models expressed in the multivariate approach, and Bou and Satorra

![Figure 2. General multivariate model for dynamic panel data.](image-url)
(2007, 2009, 2010) for specific analyses of dynamic panel data models with measurement error in different data settings (multiple-group, hierarchical data, and finite mixtures).

The univariate approach also deals with dynamic panel data models; the presence, however, of both a varying intercept and lagged dependent variables pose endogeneity problems that go beyond the methods described above for the FE and RE estimates (Antonakis et al., 2010). A great deal of econometric research has been devoted to dynamic panel data models. Estimation methods based on instrumental variables have been developed following Anderson and Hsiao (1982) with estimators differing on the transformation of the data used and the choice of instruments (see Arellano, 2003, chaps. 5-7, for full details of these methods).

**Other Extensions of the Multivariate Approach**

A variety of related models can be proposed in the multivariate framework, depending whether the model includes time-varying independent variables or not, and whether the independent variables are observable or latent. For example, the classic first-order autoregressive model (Kessler & Greenberg, 1981, chap. 7) can be specified as a special case of the dynamic multivariate model in equation (6) in which the independent variables are not included in the model. Similarly, other dynamic models allowing lagged values of both the dependent and independent variables and reciprocal effects, such as the cross-lagged model, have been estimated in the same multivariate framework (see Allison, 2009, chap. 6). These models, framed within the tradition of the analysis of interindividual differences (Baltes & Nesselroade, 1979), provide a better understanding of the relationships between repeatedly measured variables (and constructs) over time.

Other well-established panel data models used in repeated-measures designs to study intraindividual differences over time can be fitted using the multivariate panel data approach. This is the case of univariate and multivariate repeated-measures ANOVA (Misangyi et al., 2006; O’Brien & Kaiser, 1985) used for testing changes (i.e., growth) in means over time (Bergh, 1995), and LGC models (T. E. Duncan & Duncan, 1995; McArdle & Epstein, 1987; Meredith & Tisak, 1990; Rogosa & Willett, 1985) that allow change trajectories that vary across individuals. In particular, a standard LGC model (i.e., with no independent variables) can be specified by adding to the multivariate model in Figure 1 a second latent variable with time varying loadings (the random slope component) that accounts for unit-specific variation across time of the mean of $y$. The loadings of the random slope variable can be specified as either fixed values (defining linear or polynomial growth trajectories), or free parameters to empirically define the functional form of the changes in the mean of $y$. Similarly, repeated-measures ANOVA can be considered as a particular case of the LGC model where the random slopes are not allowed to vary across individuals (i.e., setting to zero the variance and the covariances of the random slope latent variable; McArdle, 2012).

**Illustration 1: Demonstration Data**

To aid the comparison of the univariate and multivariate approaches, we first present an illustration with demonstration—i.e., artificial—data. This involves one-sample panel data with certainty on the model and population values of the parameters. This type of data allows estimates to be compared with each other and with respect to the true values, and uncovers the impact of misspecification on the values of various test statistics. For simplicity, only three variables, the $y, x$ and $z$, and four time points are considered. The data was produced by equation (5) with the population values of parameters shown in Table 3. Note that the regression coefficient $\beta$ is not constant across time. This violation of the TI assumption will allow model misspecification to be illustrated. A reasonably large sample, $N = 3000$, was chosen to avoid cluttering with small sample size artifacts. The Stata syntax and seed number that
generates the one-sample data is shown in Appendix A. The code for all the univariate and multivariate analyses reported in this section is shown in Appendix A (code in Stata) and Appendix B (code in R).

Table 3 shows estimates, standard errors, and 95% confidence intervals (CIs) for all the parameters of the true model (a model that does not impose the TI assumption). Since we are analyzing a (true) correct model, it is not surprising that the 95% CI of each of the parameters contains its true population value. Furthermore, the chi-square goodness-of-fit test equals 14.65, \(df = 13\), \(p = .33\); so, as expected, the model is not rejected by the test. Our intention, however, with this illustration is to see the estimates and test statistics when the model is misspecified.

### Comparison of Univariate and Multivariate Approaches

Table 4 shows the estimation results (estimates, \(SE\), and CI) for the univariate (first panel) and multivariate (second panel) representations of FE (left-hand side of the table) and RE (right-hand side). The results were obtained with Stata: the univariate used `xtreg, fe` (for FE) and `xtreg, mle`, or `mixed with ml` (for RE); the multivariate used `sem`, estimation method `ml`. Identical numbers were obtained when using other SEM software such as EQS and Mplus.\(^\text{13}\)

---

**Table 3. Population Values and SEM Analysis: Demonstration Data.**

<table>
<thead>
<tr>
<th>Population Value</th>
<th>Estimate</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t = 1):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x)</td>
<td>2.000</td>
<td>2.002</td>
<td>0.023</td>
</tr>
<tr>
<td>(z)</td>
<td>3.000</td>
<td>2.996</td>
<td>0.021</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.000</td>
<td>0.987</td>
<td>0.019</td>
</tr>
<tr>
<td>(\sigma^2_{\epsilon})</td>
<td>1.000</td>
<td>1.016</td>
<td>0.028</td>
</tr>
<tr>
<td>(t = 2):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x)</td>
<td>0.000</td>
<td>0.050</td>
<td>0.037</td>
</tr>
<tr>
<td>(z)</td>
<td>3.000</td>
<td>2.938</td>
<td>0.035</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.000</td>
<td>0.954</td>
<td>0.018</td>
</tr>
<tr>
<td>(\sigma^2_{\epsilon})</td>
<td>1.000</td>
<td>0.979</td>
<td>0.027</td>
</tr>
<tr>
<td>(t = 3):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x)</td>
<td>2.000</td>
<td>2.022</td>
<td>0.022</td>
</tr>
<tr>
<td>(z)</td>
<td>3.000</td>
<td>3.021</td>
<td>0.019</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.000</td>
<td>0.996</td>
<td>0.018</td>
</tr>
<tr>
<td>(\sigma^2_{\epsilon})</td>
<td>1.000</td>
<td>0.996</td>
<td>0.018</td>
</tr>
<tr>
<td>(t = 4):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x)</td>
<td>4.000</td>
<td>4.020</td>
<td>0.021</td>
</tr>
<tr>
<td>(z)</td>
<td>3.000</td>
<td>2.969</td>
<td>0.018</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.000</td>
<td>0.978</td>
<td>0.018</td>
</tr>
<tr>
<td>(\sigma^2_{\epsilon})</td>
<td>1.000</td>
<td>0.948</td>
<td>0.026</td>
</tr>
<tr>
<td>(\sigma^2_u)</td>
<td>1.000</td>
<td>0.962</td>
<td>0.025</td>
</tr>
<tr>
<td>(\text{cov}(u, x_1))</td>
<td>0.1</td>
<td>0.131</td>
<td>0.011</td>
</tr>
<tr>
<td>(\text{cov}(u, x_2))</td>
<td>0.5</td>
<td>0.472</td>
<td>0.007</td>
</tr>
<tr>
<td>(\text{cov}(u, x_3))</td>
<td>0.2</td>
<td>0.187</td>
<td>0.012</td>
</tr>
<tr>
<td>(\text{cov}(u, x_4))</td>
<td>0.8</td>
<td>0.785</td>
<td>0.012</td>
</tr>
<tr>
<td>(\chi^2)</td>
<td></td>
<td>14.65</td>
<td></td>
</tr>
<tr>
<td>(df)</td>
<td></td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>(p) value</td>
<td></td>
<td>.33</td>
<td></td>
</tr>
</tbody>
</table>

Bou and Satorra
We first discuss the results for FE estimates. The univariate and multivariate FE estimates of $\beta$ are identical, with the same standard error, in the three decimal places reported. The multivariate FE also produces an estimator (and standard error) for $\gamma$, the regression coefficient of $z$, which was not available in the univariate FE. An estimate of $\gamma$, however, could also have been obtained by using the correlated random effect approach discussed in the RE estimator section; in our data, this estimate is $2.101$ ($\text{SE} = 0.021$), a value that departs from the multivariate FE estimate reported in the table. We note equality of the univariate and multivariate FE estimates of $s^2_e$; the multivariate FE also providing a standard error that could be useful to assess the significance of this variance. There is identity between the univariate and multivariate FE estimates of the mean of the random intercept $a$. Note, however, there is no equality among the univariate and multivariate FE estimates of the variance $s^2_u$ of the varying intercept. This deviation should be expected, since they are two different types of estimates: in the univariate FE, the estimate of $s^2_u$ is a between variance computed from the “observed” residuals (thus it has no standard error); in multivariate FE, the value for $s^2_u$ is a regular ML estimation (thus equipped with standard error); the 95% CI for $s^2_u$, obtained with multivariate FE, is a useful tool for assessing the presence of panel effects in the data. In the univariate approach, alternative tests for the presence of a significant variance of $a$ are the Breusch and Pagan (1980) test, if the model is estimated using the GLS estimator, or a likelihood-ratio test if the model is estimated with ML (see Antonakis et al., 2010, p. 1093). Finally, the multivariate approach provides a chi-square goodness-of-fit test for the model, the $\chi^2 = 10531.35$ ($df = 25$, $p = .00$), which indicates (strong!) rejection of the model. This warning of model misspecification is absent in the univariate FE. While there is no equivalent overall test in the univariate approach, several diagnostics of

<table>
<thead>
<tr>
<th></th>
<th>FE</th>
<th>Coef.</th>
<th>SE</th>
<th>95% CI</th>
<th>RE</th>
<th>Coef.</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>2.079</td>
<td>0.023</td>
<td></td>
<td>[2.034, 2.124]</td>
<td>2.503</td>
<td>0.022</td>
<td></td>
<td>[2.461, 2.545]</td>
</tr>
<tr>
<td>$z$</td>
<td>2.563</td>
<td>0.013</td>
<td></td>
<td>[2.538, 2.588]</td>
<td>2.415</td>
<td>0.022</td>
<td></td>
<td>[2.371, 2.459]</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.999</td>
<td>0.016</td>
<td></td>
<td>[0.967, 1.031]</td>
<td>0.973</td>
<td>0.021</td>
<td></td>
<td>[0.932, 1.014]</td>
</tr>
<tr>
<td>$\sigma^2_x$</td>
<td>8.082</td>
<td></td>
<td></td>
<td></td>
<td>0.489</td>
<td>0.040</td>
<td></td>
<td>[0.417, 0.574]</td>
</tr>
<tr>
<td>$\sigma^2_z$</td>
<td>3.160</td>
<td></td>
<td></td>
<td></td>
<td>3.278</td>
<td>0.050</td>
<td></td>
<td>[3.181, 3.378]</td>
</tr>
</tbody>
</table>

Table 4. Results for the Univariate and Multivariate Analyses (FE and RE): Demonstration Data.
possible model misspecification can be applied in the postestimation stage, as discussed in the following subsection.

For the multivariate RE, we now observe equality of estimates (and the same standard error) for all the parameters of the model: all regression coefficients, mean and variance of the varying intercept, and the variance of the disturbance term. Here also, the multivariate RE provides a chi-square goodness-of-fit test for the model: \( \chi^2 = 15436.41 \) (\( df = 29, \ p = .00 \)), an indication that the model is rejected. Note that RE increases by 4 the df of FE; this corresponds to 4 zero constraints (the correlations of \( u_t \) with \( x_1, x_2, x_3, \) and \( x_4 \)) added to the model (option (B) of the subsection “Multivariate FE and RE Estimates”).

Comparing the population values of parameters (see second column of Table 3) with the 95\% CI of the multivariate FE and RE estimates, we see that none of the population values is covered by the CI, in clear contrast to the results shown in Table 3 when a true model was used. We know TI does not hold in our data, since the population values of \( \beta \) are 2, 0, 2 and 4 for \( t = 1, 2, 3, 4 \). The models underlying the FE and RE estimates force all the parameters to be equal, obtaining a common estimate for \( \beta \) equal to 2.079 (\( SE = 0.023 \)) in the case of FE, and 2.503 (\( SE = 0.022 \)) in the case of RE, both numbers biased with respect to the true values. Note that bias is also observed for the estimates of \( \gamma \), a parameter that is correctly restricted invariant across time, but whose population value also falls outside the 95\% CI of the FE and RE estimate for both univariate and multivariate approaches. We see that misspecification of the model, attested by the chi-square goodness-of-fit test, may substantially deteriorate the quality of estimates. Thus, prior to attempting substantive interpretation of results, a model diagnostic is desirable. Misfit and model modification for the models underlying FE and RE estimates are discussed next.

**Diagnostics for Model Modification**

The chi-square goodness-of-fit test shown in the multivariate approach (both for FE and RE) alerts the researcher to problems of misspecification of the model, likely prompting her/him to search for its sources. In SEM, the “Modification Indices” (MI; Sörbom, 1989; these are Lagrange Multiplier test statistics) in combination with the “Expected Parameter Change” (EPC; Saris, Satorra, & Sörbom, 1987) are commonly used to detect restrictions causing model misfit. Inspection of the MIs of the model for FE point clearly to the parameter \( \beta \) at times 2 and 4. The MI value for \( \beta \) at \( t = 4 \) was \( \chi^2 = 6198.56 \) (\( df = 1, \ p = .00 \)), EPC = 1.937 (i.e., an estimated shift of 1.937, from the current estimate 2.079, if the restriction were released) strongly suggests the need to relax the constraint of \( \beta \) at \( t = 4 \) (note that the true value for \( \beta \) is 4). A model that sets free all the \( \beta \)s yields \( \chi^2 = 27.54 \) (\( df = 22, \ p = .19 \)); so, the model now fits the data. The chi-square difference of the two nested models (\( \Delta \chi^2 = 10531.35 - 27.54 = 10503.81 \), \( df = 3, \ p = .00 \)) is highly significant, indicating the need to release the time-invariance constraint on \( \beta \)s.

The univariate approach provides \( F \) or Wald tests for the overall significance of the regression coefficients (as well as tests of significance of individual parameters). However, these tests give no information on the validity of the TI assumption or other sources of misspecification. Even though the univariate approach lacks overall testing of goodness-of-fit, regression-based diagnostic tools for misspecification could be applied to assess deviations from TI. An anonymous reviewer suggested using the added variable plot; Figure 3 shows this plot for the regression of \( y \) on \( x \) and \( z \), with line fits for each time point added to the classic plot. This plot clearly shows the time-variation of the regression coefficient (slopes) for \( x \) (when controlling for \( z \)), a pattern of variation coincident with the true one. Adaptation to panel data analysis of classic regression diagnostics tools is an interesting avenue that could supplement the results of the MIs and other tests with graphical information.

As mentioned above, regression coefficients that change over time can also be contemplated in the univariate approach. A model with time-varying \( \beta \) (i.e., introducing interaction terms of \( x \) with
the time-dummy variables) gives an estimate of $\beta$ of 2.003 ($SE = 0.021$), and the following estimates for the interaction terms (standard errors in parentheses): $-2.000 (0.028)$, $0.022 (0.028)$, and $2.013 (0.028)$ for time periods 2, 3, and 4, respectively (the first time period is taken as a reference category). These values correctly reproduce the population values shown in Table 3.

**Illustration 2: Empirical Data**

We now compare the univariate and multivariate approaches in a real-data setup where we confront important practical issues such as missing data and uncertainty on the true model. The dependent variable $y$ is firms’ profitability, the analysis aiming to specify a model that gives a reasonable description of the relation of profitability with firms’ characteristics such as investment in R&D, advertising, size, and so on. First, we briefly describe the variables and the panel.

**Variables**

The data come 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 survey involves annual data for 10 consecutive years, 1993 to 2002, for a sample of $N = 2,515$ firms.

The dependent variable $y$ is Return on Assets ($\text{roa}$; the ratio of annual net income to total assets), a measure of firm profitability widely used in organizational research (for exhaustive details on this variable, see, e.g., McGahan & Porter, 1997; Rumelt, 1991; Schmalensee, 1985). The following time-varying independent variables ($x$s) are used: Capacity Utilization ($\text{cu}$; ratio of the utilized productive capacity to the total installed capacity of the firm); R&D Intensity ($\text{r&d}$; the firms’ annual R&D expenditures divided by annual sales); Advertising Intensity ($\text{adv}$; ratio of the firms’ total advertising expenses to sales); Size of the firm ($\text{size}$; log of number of employees). Just one time-invariant independent variable ($z$) is included in the data analyzed: the year the firm was founded ($\text{founded}$) taking 2002 as the reference year. Previous studies (e.g., Capon, Farley, & Hoenig, 1990) found that these variables have a significant impact on firms’
profitability. See also classic references that report a positive influence of independent variables on \( \text{roa} \) such as Aaker and Jacobson (1987), Farris and Reibstein (1979), Andras and Srinivasan (2003), Phillips, Chang, and Buzzell (1983), Kirner, Kinkel, and Jaeger (2009), Thornhill (2006), Brown and Eisenhardt (1995), Sinclair, Keppler, and Cohen (2000), and Buzzell, Gale, and Sultan (1975).

### Missing Data

The data for this illustration deviate from the balanced panel data setup of Illustration 1, where every unit had complete information on all the variables for all time points. Panel data are said to be balanced when all the units are observed in all the time periods. In our data, only 560 firms (22% of the firms) have complete information for all the variables in all the years; that is, 78% of the firms have some form of missingness, either because they are not in the panel in some years (unbalanced panel), or because they miss one or more variables in some years (missing data); that is, in addition to deviation from a balanced panel, we still suffer from missingness in some variables. An important point in discussing consequences of missing data is the assumption of missing at random (MAR). This is the case when conditioning on the observed data the probability of missingness is unrelated with the value actually missed. See Rubin’s (1976) seminal discussion on missing data. We address the analysis under MAR.

The simplest approach to missing data is listwise deletion: It discards any row of the data matrix that has an empty slot (i.e., a missing value in some of the columns). In our wide format data, applying listwise deletion leaves a data matrix with only 560 rows (the firms with complete data in all the years); that is, we lost 78% of the firms. In our long format data, we have 15,567 rows (firms by year with information in some of the variables), and listwise deletion leaves a data matrix with 13,980 rows (the remaining 1,587 rows have missing data in some variables in some years). So, in our data, the information retained with the listwise deletion varies for the wide and long format data sets. In the demonstration data of Illustration 1, where panel data was balanced with no missingness, this issue did not arise.

Fortunately, nearly all current software for SEM analysis now has the estimation option of ML with missing data (Arbuckle, 1996; Graham, 2003). This estimator is invoked in \textit{sem} of Stata by \texttt{mlmv}, the option to be used for the multivariate approach reported in the present illustration. This estimate ensures consistency under MAR and it retains and processes all the data of the initial 2,515 rows of the wide format data. In fact it uses more data than when applying listwise deletion to the long format data (the only option available in \texttt{xtreg} and \texttt{mixed} of Stata) that will necessarily discard the mentioned 1,587 rows that have missingness in some variables. It will be interesting the comparison of the results obtained from the univariate (listwise to long format data) and multivariate (ML for missing data) in this context of empirical data.

### Comparison of Univariate and Multivariate Approaches

Table 5 shows the estimation results for the univariate (first panel) and multivariate (second panel) representations of the FE (left-hand side) and RE estimates (right-hand side). Estimates, standard errors, and 95% CIs are displayed in the table. All the numbers in this table were produced using Stata: in the univariate approach using \texttt{xtreg, fe} for FE, and \texttt{xtreg, mle} (or \texttt{mixed with ml}) for the RE; the multivariate approach using \texttt{sem}, estimation method \texttt{mlmv}.

We first discuss the results for FE estimates. The univariate and multivariate FE estimates of the regression coefficients for the time-varying variables (the \( \beta \) s) are very similar, with also similar standard error (the largest difference on estimates occurs for \( \beta \) of variable \textit{size}, an estimate, however, that has a large standard error in both the univariate and multivariate approaches). The univariate and multivariate FE estimates of \( \sigma^2_e \) and the mean of the random
intercept are also very similar. As was noted in the illustration with demonstration data, univariate and multivariate FE estimates of the variance of the varying intercept differ due to the estimators being based on different principles. The chi-square goodness-of-fit test \( \chi^2 = 1,658.81 \) (\( df = 427, p = .00 \)), pointed to rejection of the model, a warning that is not present in the univariate approach.

The univariate and multivariate RE results also show very close estimates and standard error, now for all the parameters of the model: regression coefficients, mean and variances of the varying intercept, and the variance of the disturbance term. Clearly, the z-values of the estimates lead to the same conclusion for the univariate and multivariate approaches. Now the chi-square goodness-of-fit test is \( \chi^2 = 1,815.39 \) (\( df = 467, p = .00 \)), which also indicates rejection of the model. Note that the number of \( df \) of the RE estimator has increased by 40, as we have restricted to zero the 40 parameters corresponding to the covariances of \( u_t \) with the time-varying independent variables (4 variables times 10 years).

In addition, FE and RE estimates give nearly the same results except for the significance of the variable advertising \( \text{adv} \), which is highly significant in the case of FE but nonsignificant in the case

### Table 5. Results for the Univariate and Multivariate Analyses (FE and RE): Empirical Data.

<table>
<thead>
<tr>
<th></th>
<th>FE</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>SE</td>
</tr>
<tr>
<td><strong>cu</strong></td>
<td>0.072</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>r&amp;d</strong></td>
<td>-0.290</td>
<td>0.063</td>
</tr>
<tr>
<td><strong>adv</strong></td>
<td>-0.356</td>
<td>0.074</td>
</tr>
<tr>
<td><strong>size</strong></td>
<td>-0.809</td>
<td>0.359</td>
</tr>
<tr>
<td><strong>founded</strong></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>z</strong></td>
<td>7.250</td>
<td>1.654</td>
</tr>
<tr>
<td><strong>σ&lt;sup&gt;2&lt;/sup&gt;u</strong></td>
<td>78.194</td>
<td>1.564</td>
</tr>
<tr>
<td><strong>σ&lt;sup&gt;2&lt;/sup&gt;e</strong></td>
<td>84.670</td>
<td>1.143</td>
</tr>
</tbody>
</table>

### Table 5. Results for the Univariate and Multivariate Analyses (FE and RE): Empirical Data.

<table>
<thead>
<tr>
<th></th>
<th>FE</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>SE</td>
</tr>
<tr>
<td><strong>cu</strong></td>
<td>0.084</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>r&amp;d</strong></td>
<td>-0.268</td>
<td>0.062</td>
</tr>
<tr>
<td><strong>adv</strong></td>
<td>-0.361</td>
<td>0.075</td>
</tr>
<tr>
<td><strong>size</strong></td>
<td>-1.290</td>
<td>0.343</td>
</tr>
<tr>
<td><strong>foundation</strong></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>z</strong></td>
<td>8.225</td>
<td>1.576</td>
</tr>
<tr>
<td><strong>σ&lt;sup&gt;2&lt;/sup&gt;u</strong></td>
<td>40.250</td>
<td>2.272</td>
</tr>
<tr>
<td><strong>σ&lt;sup&gt;2&lt;/sup&gt;e</strong></td>
<td>88.059</td>
<td>1.137</td>
</tr>
</tbody>
</table>

<sup>1</sup>Ratio between estimate and SE.
of RE. We note that the standard error of the RE estimates are generally smaller than those of the FE estimates. For example, in the case of regression coefficient size, the standard error of the RE estimate is three times smaller than for the FE estimate. The gross misfit of the model, however, indicated by the chi-square goodness-of-fit tests, raises suspicions on any substantive interpretation of both FE and RE estimates. Clearly, model modification is needed.

Diagnostics for Model Modification

In the univariate setup, the differences observed between the FE and the RE estimates prompt the researcher to perform a Hausman test to choose between both estimators. In this data, the Hausman test equals $\chi^2 = 37.08$, $df = 4$, $p = .00$, and suggests basing the inferences on the FE estimator. A cautious researcher, however, would apply other tests for misspecification before proceeding with the substantive interpretation of the FE estimates. Specifically, the Wald test for heteroscedasticity in the FE model rejects the hypothesis of homoscedasticity of error variances (Wald test is $\chi^2 = 6.4 \times 10^{36}$, $df = 2268$, $p = .00$). Similarly, the Wooldridge test for autocorrelation (Wooldridge, 2002) rejects the hypothesis of zero first-order serial correlation, $F(1, 1867) = 45.32$, $p = .00$. A residuals plot, not shown due to space restrictions, indicates that the assumption of normality of the error term is also violated. Similarly, other regression diagnostic plots, such as the added variable plot used with the demonstration data, could also be applied here, though the number of these plots increases with the number of time-invariant variables, and no clear conclusion is obtained. Overall, these tests would lead us to modify the model by considering not only dropping TI constraints, but also allowing for heteroscedasticity and autocorrelation on the residuals. Models with complex error structures (heteroscedasticity and autocorrelation) can be estimated using the `mixed` command of Stata. See Snijders and Bosker (2012, chap. 6) for strategies on model specification and testing in the context of the univariate HLM, and Bliedle and Ployhart (2002) and Holcomb et al. (2010) on the more specific context of growth-curve models and linear models with random effects at all levels.

In the multivariate approach, the test statistics for model search are based on the simple principle of the overall goodness-of-fit test, and the chi-square difference test for nested models. The MIs are just handy tools that approximate chi-square difference testing for restricted models that are nested to the fitted one. The difference between the chi-square goodness-of-fit test for the model with covariances $\rho(x, u)$ set free, and the other with the same covariances set to zero, is our multivariate version of the Hausman test. The chi-square values for the two models are 1815.39 ($df = 467$) and 1769.11 ($df = 463$). This yields the chi-square difference $\Delta \chi^2 = 46.28$ with a difference of degrees of freedom $\Delta df = 4$ (4 is the number of time-varying independent variables in the model), and $p = .00$. The test rejects the null hypothesis of zero correlation among the varying intercept and time varying independent variables, the same conclusion obtained when using the classic Hausman test for this data in the univariate approach.

The multivariate RE and FE are both rejected by the chi-square goodness-of-fit test (see the $\chi^2$ values in Table 5). Table 6 shows a “‘bottom-up” strategy (from more to fewer restrictions) for

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Restrictions (Within TI)</th>
<th>$\chi^2(df)$</th>
<th>$\Delta\chi^2(df)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: FE</td>
<td>$\alpha_i = \alpha; \beta_i = \beta; \gamma_t = \gamma; \lambda_t = \lambda; \sigma^2_t(t) = \sigma^2_e$</td>
<td>1658.805* (427)</td>
<td>156.58* (40)</td>
</tr>
<tr>
<td>M1: $\alpha_i = \alpha$</td>
<td>$\beta_i = \beta; \gamma_t = \gamma; \lambda_t = \lambda; \sigma^2_t(t) = \sigma^2_e$</td>
<td>1509.082* (418)</td>
<td>149.72* (9)</td>
</tr>
<tr>
<td>M2: $\beta_i = \beta$</td>
<td>$\gamma_t = \gamma; \lambda_t = \lambda; \sigma^2_t(t) = \sigma^2_e$</td>
<td>1423.567* (382)</td>
<td>85.52* (36)</td>
</tr>
<tr>
<td>M3: $\gamma_t = \gamma$</td>
<td>$\lambda_t = \lambda; \sigma^2_t(t) = \sigma^2_e$</td>
<td>1414.970* (373)</td>
<td>8.60 (9)</td>
</tr>
<tr>
<td>M4: $\lambda_t = \lambda$</td>
<td>$\gamma_t = \gamma; \sigma^2_t(t) = \sigma^2_e$</td>
<td>1184.477* (373)</td>
<td>239.09* (9)</td>
</tr>
<tr>
<td>M5: $\sigma^2_t(t) = \sigma^2_e$</td>
<td>$\gamma_t$</td>
<td>865.513* (364)</td>
<td>318.96* (9)</td>
</tr>
</tbody>
</table>
model modification within TI. The first column of the table shows the sequence of nested models considered, with an indication of which set of TI restrictions is relaxed in each step. The second column specifies the restrictions associated to each model. The chi-square goodness-of-fit test for each model and the chi-square difference test in each step are presented in the last two columns of the table. In parenthesis we show the corresponding df of the test, an asterisk denoting the test is significant at the 5% level. The sequence starts with a model that imposes TI in full. Subsets of restrictions of TI are subsequently released according to whether the corresponding chi-square difference test is significant. The choice of the sets of parameter to be freed in each step is left to the research context and the researcher. In panel data, however, the choice made in Table 5 seems reasonable: setting free first the means, then the regression coefficients, and so on, as shown in the table. It also left to the researcher to decide on the number of parameters to be freed in each step; for example, some of the β coefficients could be invariant across time while others could vary. In other research contexts, a similar table could be used as a “top-down” strategy, starting with a less restricted model and adding restrictions as we move up.

For our data, in each step the chi-square goodness-of-fit test is significant (indicating the model in that step does not fit the data) and also the chi-square differences are all significant except when testing for the time invariance of γ. Therefore, the strategy suggests keeping γ invariant in the final model. Model search within the family of TI restrictions, however, leaves us with a chi-square goodness-of-fit of \( \chi^2 = 865.513 \) (\( df = 364, p = .00 \)), still a nonfitting model. Other modifications of the model therefore seem to be required. Given the type of data, allowing for dynamics in the model is likely to be the course of action.

As noted by one reviewer, this sequence of testing could also be undertaken in the univariate using the `mixed` of Stata. Minus two times the differences of log likelihoods of two nested models is a likelihood ratio test statistic (see Bliese & Ployhart, 2002; and Snijders & Bosker, 2012, chap. 6), that can be used to contrast alternative models and to aid decisions about including specific terms in the model. This approach, however, would be limited to the class of models allowed by `mixed`. In particular, following the strategy proposed by Bliese and Ployhart (2002) to build and test mixed models, the univariate approach would lead to a model with across time heteroscedastic error terms, which is different from the dynamic model proposed using the “bottom-up” strategy described below. This, therefore, is an example where the univariate and multivariate approaches lead to different final models.

Comparison of models could also be undertaken using the AIC (Akaike, 1974) and BIC (Raftery, 1995) criteria instead of chi-square values. These information criteria provide a rational trade-off between fit and complexity and can be useful to compare nonnested models (Rabe-Hesketh & Skrondal, 2012, sec. 6.5).

### A Dynamic Model

Inspection of the MIs of the last model fitted in Table 6 points to the significance of autoregressive effects in the dependent variable. This suggests enlarging the model within the family of dynamic panel data models. In particular, we fit the dynamic model represented by equation (6), shown in the path diagram of Figure 2, except that (as a result of the model modification strategy of previous section) we set the parameter γ invariant over time and we also allow interaction between time and the varying intercept (loadings \( \lambda, s \) varying in t).

The chi-square goodness-of-fit test for this model is \( \chi^2 = 479.406 \) (\( df = 286, p = .00 \)). We see a large decrease in the chi-square value, compared with the decrease in df, but the test still rejects the model. The scaled chi-square goodness-of-fit test (correction for nonnormality, Satorra & Bentler, 1994) gives a \( \chi^2 = 415.644 \) (\( df = 286, p = .00 \)), values that still fall in the rejection side. In a further
search for model improvement, inspection of MIs shows significant values for second-order auto-regressive effects in some years, with small EPCs (expected parameter change) in comparison with the first-order effects already introduced. Significant MIs and small EPCs have been argued as indications of nonsubstantial misspecification in the model (Saris, Satorra, & Van der Veld, 2009); this leads us to halt our search for a better fitting model. We now describe the results of this final model.

Table 7 displays estimates and standard error for all the parameters of the dynamic model. Note that in addition to the dynamics, the TI assumption is not in place, so the table displays parameter estimates varying across time. The table shows a significant autoregressive coefficient $d_{rot}$ in 8 out of the 10 years, a dynamic effect that was ignored (i.e., set to zero) by the previous static models. With regard to the other model parameters, most of the regression coefficients have the same sign as in the estimates obtained under the static model (see Table 5). Regression coefficients are positive (and significant in 8 out of the 10 years) in the case of $cu$, and negative in the case of both $r&d$ and $adv$ (significant values in 7 and 6 years, respectively). The exception, however, is the regression coefficient of $size$, which was estimated as negative and significant in the static model ($/C0_{1}: 396, z-value = -3.84$) and now is still negative, but not significant. The results derived from this dynamic model show a substantial difference with regard to the impact of $size$ on $roa$. Note also that the estimates of $\lambda$, change over time, and that the varying intercept has a smaller variance (although still significant) as compared with the previous FE and RE models; that is, introduction of the autoregressive structure $roa_{t-1}$ decreases the variance of the unobserved permanent differences across firms. In summary, our search strategy, which started with a very poor fitting static model, suggests a dynamic panel data model that shows time variation in fundamental parameters and an autoregressive structure on the dependent variable $roa$.

<table>
<thead>
<tr>
<th>Year</th>
<th>$roa_{t-1} \rightarrow roa_t$</th>
<th>$cu \rightarrow roa_t$</th>
<th>$r&amp;d \rightarrow roa_t$</th>
<th>$adv \rightarrow roa_t$</th>
<th>$size \rightarrow roa_t$</th>
<th>$u$</th>
<th>$\sigma^2_u$</th>
<th>$\sigma^2_{\epsilon(t)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>0.209* (0.043)</td>
<td>0.140* (0.024)</td>
<td>-0.353* (0.144)</td>
<td>-0.249 (0.145)</td>
<td>-0.278 (0.511)</td>
<td>1.000 (0.159)</td>
<td>26.373* (8.127)</td>
<td>106.497* (10.220)</td>
</tr>
<tr>
<td>1994</td>
<td>0.277* (0.058)</td>
<td>0.051* (0.022)</td>
<td>-0.268* (0.124)</td>
<td>-0.143 (0.122)</td>
<td>-0.283 (0.334)</td>
<td>0.547* (0.198)</td>
<td>26.373* (8.127)</td>
<td>101.831* (11.221)</td>
</tr>
<tr>
<td>1995</td>
<td>0.282* (0.084)</td>
<td>0.092* (0.023)</td>
<td>0.022 (0.150)</td>
<td>-0.289* (0.131)</td>
<td>0.429 (0.418)</td>
<td>0.752* (0.145)</td>
<td>26.373* (8.127)</td>
<td>89.488* (9.943)</td>
</tr>
<tr>
<td>1996</td>
<td>0.219* (0.055)</td>
<td>0.034 (0.023)</td>
<td>-0.366* (0.112)</td>
<td>-0.249* (0.146)</td>
<td>0.418 (0.532)</td>
<td>0.714* (0.221)</td>
<td>26.373* (8.127)</td>
<td>82.022* (9.733)</td>
</tr>
<tr>
<td>1997</td>
<td>0.118* (0.049)</td>
<td>0.097* (0.024)</td>
<td>-0.244 (0.141)</td>
<td>-0.396* (0.181)</td>
<td>0.532 (0.584)</td>
<td>1.095* (0.231)</td>
<td>26.373* (8.127)</td>
<td>64.609* (8.237)</td>
</tr>
<tr>
<td>1998</td>
<td>0.075 (0.087)</td>
<td>0.099* (0.026)</td>
<td>-0.501* (0.167)</td>
<td>-0.331 (0.129)</td>
<td>0.032 (0.410)</td>
<td>1.108* (0.221)</td>
<td>26.373* (8.127)</td>
<td>64.075* (9.662)</td>
</tr>
<tr>
<td>1999</td>
<td>0.099 (0.066)</td>
<td>0.071* (0.021)</td>
<td>0.167 (0.139)</td>
<td>-0.306* (0.121)</td>
<td>0.032 (0.369)</td>
<td>0.957* (0.283)</td>
<td>26.373* (8.127)</td>
<td>76.959* (11.207)</td>
</tr>
<tr>
<td>2000</td>
<td>0.240* (0.060)</td>
<td>0.051* (0.022)</td>
<td>-0.181 (0.139)</td>
<td>-0.319* (0.202)</td>
<td>0.032 (0.397)</td>
<td>1.108* (0.221)</td>
<td>26.373* (8.127)</td>
<td>82.109* (11.157)</td>
</tr>
<tr>
<td>2001</td>
<td>0.148* (0.053)</td>
<td>0.039* (0.019)</td>
<td>-0.419* (0.123)</td>
<td>-0.238* (0.112)</td>
<td>0.032 (0.397)</td>
<td>0.722* (0.156)</td>
<td>26.373* (8.127)</td>
<td>62.503* (7.530)</td>
</tr>
</tbody>
</table>

*Significant at the 5% level.
Discussion and Conclusions

Two perspectives for analyzing panel data have been discussed and illustrated with demonstration and empirical data: one perspective arranges the data in long format and uses univariate regression methods; the other perspective arranges the data in wide format and uses a set of regression equations. Although few technical articles have compared the two approaches (Allison, 2009; Allison & Bollen, 1997; Bollen & Brand, 2010; Ejrnaes & Holm, 2006; Teachman et al., 2001), we find a body of work in organization research using the univariate approach to panel data which reveals a lack of awareness that a similar analysis could be undertaken using the multivariate approach. By reviewing the two approaches side by side, and by working with the illustrations, we provide an overview of the practice of panel data modeling that we feel would be useful to the wide audience of practitioners, especially in the field of organizational research.

Illustration 1, using demonstration data, first allows the researcher to exemplify the exact equality of results of the univariate and multivariate approaches for the classic FE and RE estimates. Second, it provides evidence of the severe consequences on bias of estimates (not only on misspecified parameters, but also on other correctly specified ones) of ignoring the violation of what has been called the TI assumption, an assumption that may be violated in practice and, thus, as with other assumptions (i.e., random slopes), should be empirically tested (Bliese & Ployhart, 2002). Third, it allows tools for model modification to be assessed by comparing the ones obtained using the univariate approach (e.g., regression diagnostic tools, added variable plots, deviance tests, etc.) with the ones provided by the multivariate perspective (chi-square goodness-of-fit test, chi-square difference tests, MI, EPC, etc.). Classic test statistics used in the univariate approach to panel data can be replicated in the multivariate approach. For example, the Hausman test for panel data can be framed as a chi-square difference test. Other common tests in the econometrics of panel data (e.g., tests for heteroscedasticity, autocorrelation, etc.) can be performed with the chi-square difference test common in SEM analysis. This may add conceptual simplicity to a plethora of apparently conceptually different tests presently in use in the econometric approach to panel data.

Illustration 2 uses an empirical data set and allows us to compare the two approaches in a setup where we have missing data and uncertainty about which is the true model. Capitalizing on diagnostic and model modification tools available in the multivariate approach using SEM, we provide a step-by-step guide on how to analyze panel data to achieve a better fitting model. The “bottom-up” strategy of model modification proposes a sequence of models (from the most to the least restrictive) that releases the restrictions implied by the TI assumption, and helps applied researchers in the search for a model that significantly improves the fit to the data. Although the choice of the sets of parameters to be freed usually depends on the research context and substantive knowledge, we propose a sequence of models that we believe is reasonable in the case of panel data analysis. In any case, all modifications should make sense theoretically (Green & Thompson, 2010). The multivariate approach would also simplify the expansion of the model to include measurement error, multiple indicators, lagged regressors, random slopes, and other aspects of SEM. In Illustration 2 we showed how the search for a fitted model, and the need to introduce an autoregressive structure on the dependent variable, led us to enlarge the model within the family of dynamic panel data models. In the multivariate approach this extension can be accomplished in the same frame of models and estimation methods, unlike the univariate approach which requires the use of instrumental variables adapted to this problem (hence, adding conceptual and practical complexities, e.g., Halaby, 2004, p. 541). The article also discusses how the general multivariate panel data model encompasses typical models in organizational research such as repeated-measures ANOVA models and LGC models. A comparison of the models discussed in the article and extensions is shown in Table 8.

The comparison of the univariate and multivariate approaches to panel data also brings up front practical differences regarding the treatment of missing data that has not received much attention to
Table 8. Comparison of Panel Data Models Discussed in the Paper and Extensions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data Format</th>
<th>Unobserved Heterogeneity (Z)</th>
<th>TI Assumption</th>
<th>Type of Independent Variables</th>
<th>Correlation Between Independent Variables and Z</th>
<th>Error Structure</th>
<th>Measurement Error Allowed</th>
<th>Dynamic Components Estimators</th>
<th>Intraunit Time Variation</th>
<th>Software Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate regression</td>
<td>Wide</td>
<td>Random</td>
<td>Optional: time-invariant</td>
<td>Optional: time-invariant</td>
<td>FE: Fixed effect estimator</td>
<td>RE: ML</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>SEM</td>
</tr>
<tr>
<td>Multivariate regression</td>
<td>Wide</td>
<td>Random</td>
<td>Optional: time-invariant</td>
<td>Optional: time-invariant</td>
<td>FE: Fixed effect estimator</td>
<td>RE: ML</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>SEM</td>
</tr>
<tr>
<td>LGC models</td>
<td>Wide</td>
<td>Random</td>
<td>No</td>
<td>Time-invariant</td>
<td>FE: Fixed effect estimator</td>
<td>RE: ML</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>SEM</td>
</tr>
<tr>
<td>Multivariate repeated measures ANOVA</td>
<td>Wide</td>
<td>Not present</td>
<td>No</td>
<td>Dummy variables</td>
<td>N/A</td>
<td>RE: ML</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>SEM</td>
</tr>
</tbody>
</table>

Note: FE = fixed effect; i.i.d. = independent and identically distributed; LGC = latent growth curve; ML = maximum likelihood; OLS = ordinary least squares; RE = random effect; SEM = structural equation model.

a. E.g., SAS, SPSS, Stata, etc.
b. E.g., EQS, LISREL, Mplus, etc.
c. E.g., SAS, SPSS, Stata, SEM software, etc.
d. Except for the autoregressive latent curve model of Bollen and Curran (2006, chap. 7)
date. It has been argued that the increased number of columns needed to record the data in the wide format makes the multivariate approach unfeasible, and that the univariate approach, using the long format, is more appropriate to deal with panels in which missing data are present (Singer & Willett, 2003). The illustration with empirical data has demonstrated that nowadays missing data are no longer an issue since we can circumvent the problem with no trouble to the practitioner by using the estimation method of ML with missing data. The missing data issue adds further interest to the comparison of the univariate and multivariate approaches.

To conclude, our comparison of the univariate and multivariate perspectives for panel data should raise awareness among practitioners of the commonalities and differences of the two alternative ways of analyzing panel data. This comparison should also facilitate communication among researchers working with different methodologies. Improved understanding of how the two approaches are related to one another will help them to make better use of each approach. Although we argued that the multivariate perspective excels on conceptual simplicity and in facilitating diagnostic statistics for goodness-of-fit tests and model modification, this should not be understood as a recommendation for the widespread use of one approach over the other. Researchers will find it useful to add the other approach to their repertoire of analytic skills, and we encourage them to do so. We hope the concrete illustrations worked out in the article—with the computed code for the analyses available in the appendix, and the demonstration data replicable from the seed code provided—will help researchers apply the methods discussed in their own panel data settings.

Appendix A: Stata Code for the Analysis of Demonstration Data

**Generation of the Demonstration Data**

```stata
set seed 2016
matrix M = 0, 0, 0, 0, 0, 0
matrix V = (1, .4, .3, .2, .1, .4) .4, 1, .4, .3, .5, .8
 .3, .4, 1, .4, .2, .2 .3, .4, 1, .8, 0
 .1, .5, .2, .8, 1, 0 .4, .8, .2, 0, 0, 1)
matrix list M
matrix list V
drawnorm x1 x2 x3 x4 eta z, n(3000) cov(V) means(M)
gen y1 = 1 + 2*x1 + 3*z + eta + rnormal()
gen y2 = 1 + 0*x2 + 3*z + eta + rnormal()
gen y3 = 1 + 2*x3 + 3*z + eta + rnormal()
gen y4 = 1 + 4*x4 + 3*z + eta + rnormal()
gen i = _n
```

**Convert From Wide to Long Format and Declare the Data as a Panel**

```stata
reshape long x y, i(i) j(t)
xtset i t
```

**FE (Within Estimator) Model (Table 4)**

```stata
xtreg y x, fe
```

**RE (ML Estimator) Model (Using the xtreg and mixed Options; Table 4)**

```stata
xtreg y x z, mle
mixed y x z || i:, variance
```
Hausman Test
quietly xtreg y x, fe
estimates store fixed
quietly xtreg y x z, re
estimates store random
hausman fixed random

Multivariate Approach Without the TI Assumption (Table 3)
sem (u@1 x1 z -> y1) (u@1 x2 z -> y2) (u@1 x3 z -> y3) ///
(u@1 x4 z -> y4), latent(u) cov(_lexogenous*oexogenous) ///
cov(z*u@0)

Multivariate FE (Table 4)
sem (_cons@alpha u@1 x1@b z@g -> y1) (_cons@alpha u@1 x2@b z@g -> y2) ///
(_cons@alpha u@1 x3@b z@g -> y3) (_cons@alpha u@1 x4@b z@g -> y4), ///
latent(u) var(e._OEn@e) cov(_lex*oex) cov(z*u@0)

Multivariate RE (Table 4)
sem (_cons@alpha u@1 x1@b z@g -> y1) (_cons@alpha u@1 x2@b z@g -> y2) ///
(_cons@alpha u@1 x3@b z@g -> y3) (_cons@alpha u@1 x4@b z@g -> y4), ///
latent(u) var(e._OEn@e) cov(_lex*oex@0)

Univariate Model With Time-Varying Regression Parameters
tab t, gen(time)
gen xtime1 = x*time1 ! Generate the time-dummy variables
gen xtime2 = x*time2
gen xtime3 = x*time3
gen xtime4 = x*time4
xtreg y x xtime2-xtime4, fe
xtreg y x xtime2-xtime4, mle
mixed y x xtime2-xtime4 || i:, variance

Correlated Random Effect Model (Mundlak, 1978, Approach)
bysort i: egen mean_x = mean(x)
mixed y x z mean_x || i:, variance

Univariate Model With Heteroscedastic Disturbances
mixed y x z || i:, residuals(independent, by(t)) variance

Univariate Model With First Order Autocorrelated Disturbances
mixed y x z || i:, residuals(ar 1, t(t)) variance
Appendix B: R Code for the Analysis of Demonstration Data

**Generation of the Demonstration Data**

```r
set.seed(2017)
n <- 3000;
sdeta <- 1
muC <- c(1,1,1,1)
beta <- c(2,0,2,4)
gamma <- c(3,3,3,3)
library(MASS)
mu <- rep(0,6)
# X 1X 2X 3X 4.e t a Z
sigma <- matrix(c(
  1.0, 0.4, 0.3, 0.2, 0.1, 0.4,
  0.4, 1.0, 0.4, 0.3, 0.5, 0.8,
  0.3, 0.4, 1.0, 0.4, 0.2, 0.2,
  0.2, 0.3, 0.4, 1.0, 0.8, 0.0,
  0.1, 0.5, 0.2, 0.8, 1.0, 0.0,
  0.4, 0.8, 0.2, 0.0, 0.0, 1.0
), 6,6)
XX <- mvrnorm(n, mu, sigma)
X <- cbind(XX[,1], XX[,2], XX[,3], XX[,4])
colnames(X) <- c("X1," "X2," "X3," "X4")
eta <- XX[,5]
Z <- XX[,6]
Y <- cbind(Y1, Y2, Y3, Y4)
dataw <- as.data.frame(cbind(X, Z, Y))
attach(dataw)
```

**Packages Used**

```r
install.packages("reshape") # install the reshape package
install.packages("plm") # install the plm package
install.packages("nlme")
install.packages("lavaan")
```

**Convert From Wide to Long Format**

```r
library(reshape)
datal <- reshape(dataw, direction = "long," varying = c(1:4,6:9), sep = "")
names(datal) <- c("z," "t," "x," "y," "i")
attach(datal)
```
FE (Within Estimator) Model (Using the plm Package; Table 4)

```r
library(plm)
fixed <- plm(y ~ x, data = datal, index = c("i", "t"), model = "within")
summary(fixed)
```

RE (Feasible GLS) Model

```r
random <- plm(y ~ x + z, data = datal, index = c("i", "t"), model = "random")
summary(random)
```

Hausman Test

```r
fixed <- plm(y ~ x, index = c("i", "t"), model = "within", data = datal)
summary(fixed)
random <- plm(y ~ x + z, index = c("i", "t"), model = "random", data = datal)
summary(random)
phtest(fixed, random)
```

RE (ML Estimator) Model (Using the nlme Package; Table 4)

```r
library(nlme)
random <- lme(y ~ x + z, random = ~1|i, data = datal)
summary(random)
```

Univariate Model With Heteroscedastic Disturbances

```r
update(random, weights = varIdent(form = ~1|t))
```

Univariate Model With First Order Autocorrelated Disturbances

```r
update(random, correlation = corAR1())
```

Group-Mean Centering (Note 14)

```r
# function to compute the group-mean centering of variable
groupcentered <- function(x, group = i){
  gm <- aggregate(x, list(group), mean)
  xx <- x
  for (ii in 1:dim(gm)[1])
    xx[group == ii] = gm[ii,2]
  # gcx is group centered x
  gcx <- x-xx
  list(gmean = xx, gcentered = gcx)
}
# Computation of the group-mean x and the group-mean centered x
xx <- groupcentered(x, i)$gmean
gcx <- groupcentered(x, i)$gcentered
```

Correlated Random Effect Model (Mundlak, 1978, Approach)

```r
fixedM <- lme(y ~ x + z + xx, random = ~1|i, data = datal)
summary(fixedM)
```
Hybrid Model (Allison, 2009)

```r
fixedH <- lme(y ~ gcx + z + xx, random = ~1|i, data = datal)
summary(fixedH)
```

Group-Mean Centering (Enders & Tofighi, 2007)

```r
fixedET <- lme(y ~ gcx + z, random = ~1|i, data = datal)
summary(fixedET)
```

Univariate Model With Time-Varying Regression Parameters

```r
attach(datal)
# Generate the time-dummy variables
d1 <- 1*(t = 1)
d2 <- 1*(t = 2)
d3 <- 1*(t = 3)
d4 <- 1*(t = 4)
# Generate the interaction terms
d1x <-d1*x
d2x <-d2*x
d3x <-d3*x
d4x <-d4*x
# Using the function groupcentered defined above:
d2xc <-groupcentered(d2x, i)$gmean
d3xc <-groupcentered(d3x, i)$gmean
d4xc <-groupcentered(d4x, i)$gmean

fixedtv <- lme(y ~ x + xx + d2x + d3x + d4x + d2xc + d3xc + d4xc + z,
random = ~1|i, data = datal)
summary(fixedtv)
```

Multivariate Approach Without the TI Assumption (Table 3)

```r
###reading the data of Demonstration 1
names(dataw)<- c("X1,""X2,""X3,""X4,""Z,""Y1,""Y2,""Y3,""Y4")
attach(dataw)
# calling lavaan
library(lavaan)
pmodel <-'
alpha =~ 1*Y1 + 1*Y2+1*Y3+1*Y4
Y1 =~ mu1*1+ be1*X1 + ga1*Z + 1*alpha
Y2 =~ mu2*1+ be2*X2 + ga2*Z + 1*alpha
Y3 =~ mu3*1+ be3*X3 + ga3*Z + 1*alpha
Y4 =~ mu4*1+ be4*X4 + ga4*Z + 1*alpha
alpha ~~ alpha
X1 ~~ X1+X2+X3+X4
X2 ~~ X2+X3+X4
X3 ~~ X3+X4
X4 ~~ X4
Z ~~ Z + X1+X2+X3+X4
```
Y1 ~ sig1*Y1
Y2 ~ sig2*Y2
Y3 ~ sig3*Y3
Y4 ~ sig4*Y4
  X1 ~ 1
  X2 ~ 1
  X3 ~ 1
  X4 ~ 1
  Z ~ 1
## fixed effect
  X1 ~ cor1*alpha
  X2 ~ cor2*alpha
  X3 ~ cor3*alpha
  X4 ~ cor4*alpha

fit <- lavaan(pmodel, data = dataw)
summary(fit)

**Multivariate FE (Table 4)**

Zc <- Z - mean(Z) # note: we center Z, see note in article

pmodelFE <- '  alpha = ~ 1*Y1 + 1*Y2+1*Y3+1*Y4
  Y1 ~ mu*1+ be*X1 + ga*Zc + 1*alpha
  Y2 ~ mu*1+ be*X2 + ga*Zc + 1*alpha
  Y3 ~ mu*1+ be*X3 + ga*Zc + 1*alpha
  Y4 ~ mu*1+ be*X4 + ga*Zc + 1*alpha
  alpha ~ ~ alpha
  X1 ~ X1+X2+X3+X4
  X2 ~ X2+X3+X4
  X3 ~ X3+X4
  X4 ~ X4
  Zc ~ Zc + X1+X2+X3+X4
  Y1 ~ sig*Y1
  Y2 ~ sig*Y2
  Y3 ~ sig*Y3
  Y4 ~ sig*Y4
  X1 ~ 1
  X2 ~ 1
  X3 ~ 1
  X4 ~ 1
  Zc ~ 1
## fixed effect
  X1 ~ cor1*alpha
  X2 ~ cor2*alpha
  X3 ~ cor3*alpha
  X4 ~ cor4*alpha

fit <- lavaan(pmodelFE, data = as.data.frame(cbind(dataw, Zc)))
summary(fit)
Multivariate RE (Table 4)

\[
\text{pmodelRE} <- \alpha = \sim 1*Y1 + 1*Y2 + 1*Y3 + 1*Y4 \\
Y1 \sim \mu*1 + \text{be}^*X1 + \text{ga}^*Z + 1*\alpha \\
Y2 \sim \mu*1 + \text{be}^*X2 + \text{ga}^*Z + 1*\alpha \\
Y3 \sim \mu*1 + \text{be}^*X3 + \text{ga}^*Z + 1*\alpha \\
Y4 \sim \mu*1 + \text{be}^*X4 + \text{ga}^*Z + 1*\alpha \\
\alpha \sim \alpha \\
X1 \sim X1 + X2 + X3 + X4 \\
X2 \sim X2 + X3 + X4 \\
X3 \sim X3 + X4 \\
X4 \sim X4 \\
Z \sim Z + X1 + X2 + X3 + X4 \\
Y1 \sim \text{sig}\text{*Y1} \\
Y2 \sim \text{sig}\text{*Y2} \\
Y3 \sim \text{sig}\text{*Y3} \\
Y4 \sim \text{sig}\text{*Y4} \\
X1 \sim 1 \\
X2 \sim 1 \\
X3 \sim 1 \\
X4 \sim 1 \\
Z \sim 1
\]

fit <- lavaan(pmodelRE, data = dataw)
summary(fit)


<table>
<thead>
<tr>
<th>Approach</th>
<th>Studies</th>
</tr>
</thead>
</table>

(continued)
<table>
<thead>
<tr>
<th>Approach Studies</th>
</tr>
</thead>
</table>
Table C. (continued)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Studies</th>
</tr>
</thead>
</table>

Multivariate


1Hausman test has been used to decide between fixed effect and random effect.
2Group-mean centered has been used.

References for Table C


Firth, B. M., Chen, G., Kirkman, B. L., & Kim, K. (2015). Newcomers abroad: Expatriate adaptation during early phases of international assignments. *Academy of Management Journal, 57*(1), 280-300.


Acknowledgments
We are thankful for the valuable suggestions and highly constructive comments from associate editor John Antonakis and three anonymous reviewers.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been supported by the Spanish MEC Grants ECO2015-66671-P and ECO2014-59885-P.

Notes
1. The literature review resulted from a manual revision of the articles obtained by Boolean searches (on full-text articles) using the keywords “panel,” “longitudinal,” “repeated measures,” “Hausman test,” “fixed effects,” “random effects,” “unobserved heterogeneity,” and using the following databases: Business Source Premier for Academy of Management Journal; PsycARTICLES for Journal of Applied Psychology; SAGE Complete A-Z List for Journal of Management; and Willey Online Library for Strategic Management Journal. The search provided an initial list of (418) articles, from which we discarded (171) articles that, in fact, did not involve panel data models (hazard models, survival analysis, cross-section regression, mediation models, etc. were excluded). The final list of articles surveyed is shown in Appendix C, with a classification on whether they use the univariate or the multivariate approach.

2. See LeBreton and Senter (2008) for more detailed discussion on ICC indices, especially in relation to measuring interrater reliability and interrater agreement.

3. For simplicity of the exposition, we consider the case of just one \( x \) (time-varying variable) and one \( z \) (time-invariant variable). Extending the case to several \( x \) and \( z \) is straightforward.

4. With cross-section data (i.e., when \( T = 1 \)), this is the only option available to the researcher, since lack of repeated observations of \( y \) across time confounds \( u_t \) with the disturbance term \( e_{it} \). Provided \( u_t \) is independent of the variables \( x \)s and \( z \)s, a model with a constant intercept would still yield consistent OLS estimators of the regression coefficients.

5. We also adhere to Mundlak (1978) in using the term fixed to refer to inferences conditional on the values of those variables in the sample, leaving the researcher to decide whether inferences are made conditionally on the sample at hand, or just refer to the population from where the sample is taken.

6. There are alternative estimation approaches to the fixed effect estimator, such as the first-difference estimator, in which the data are transformed by taking first differences of all the variables, and the least squares dummy variable (LSDV) model, in which a dummy variable is used for each unit (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Greene, 2003, chap. 13; Pedhazur, 1982, chap. 14). Since these estimators are equivalent, in the illustrations we report only the within estimator.
7. This is a non-iterative GLS whose estimate of the error variance and serial correlation is resolved using the ICC(1) of the pooled OLS residuals (that is, variance components analysis is applied to the pooled OLS residuals to estimate an ICC from which we infer the error structure used by the feasible GLS). A basic assumption for consistency of the estimate of the ICC from pooled OLS residuals is the zero correlation between the varying intercept and the independent variables at each time point (see Wooldridge, 2013, sec. 14.2, for details).

8. The Stata mixed-regression command mixed with the option ml produces an identical estimator to the xtreg, mle, but adds more flexibility in modeling the disturbance covariance matrix (e.g., allowing for heteroscedasticity and autocorrelation) and also in model comparison (likelihood ratio tests).

9. An alternative is Allison’s (2009, pp. 23-25) hybrid model where, in addition to adding the \( \bar{x}_i \), the time-varying variables are expressed in deviation from their cluster mean. Therefore, cluster mean centering of the \( x \)s (but not grand mean centering) solves the problem of inconsistency of the RE estimate of \( \beta \) when \( z_i \) correlates with the independent variables (see Enders & Tofighi, 2007, for further discussion of centering in mixed models). Both approaches also produce an estimate of \( \gamma \); however, this estimate “should be interpreted with caution” (Wooldridge, 2013, p. 498). See Schunck (2013) for a discussion of the two approaches and the limitations of the estimate for \( \gamma \) when \( z_i \) correlates with the independent variables.

10. Using the univariate approach regression coefficients may even be assumed to be random, an extension that is also possible with the multivariate approach, but not discussed in the present article.

11. The same FE estimate (and standard error) would be obtained by the multivariate approach when \( z \) is absent from the model, and also when the covariance between \( u \) and \( z \) is fixed to any arbitrary value.

12. In this specification, the arrows emanating from the constant to 1 to \( y_1, \ldots, y_T \) (means) should be replaced by two arrows, one pointing to \( u_i \) and the other pointing to the added random slope component.

13. The code for the analysis with this alternative software is available from the authors upon request.

14. For this equality to hold, we used a centered \( z_i \); if \( z \) is not centered, then the estimate of \( \alpha \) in the univariate and multivariate approaches would differ by the mean of \( z \) multiplied by the estimate of \( \gamma \).

15. See Appendix A for the Stata code of univariate models with time-varying regression coefficient and heteroscedastic and autocorrelated disturbances. See also Singer and Willett (2003, chap. 7) and Rabe-Hesketh and Skrondal (2012, chap. 6) for applications using Stata.

16. We computed this test statistic using the difference of chi-square model fit values for two nested models, one with the correlations set free but equal over time, and the other with correlations set to zero. Another version of the test statistic could be based on the chi-square difference where the less restricted model sets the correlations \( \rho(x, u) \) completely free (i.e., not equal across time). This would lead to a chi-square value of 168.59 with 49 degrees of freedom.

17. See in Bollen and Brand (2010, Table 1) a backward search strategy of fitting models from less to more restrictive.

References


Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structure analysis (LMACS) and multiple indicator latent growth modeling (MLGMO). *Organizational Research Methods, 1*, 421-483.


