A closer look at random and fixed effects panel regression in structural equation modeling using `lavaan`

Henrik Kenneth Andersen

Chemnitz University of Technology
Institute of Sociology
henrik.andersen@soziologie.tu-chemnitz.de

30 July, 2021

Abstract

This article provides an in-depth look at random and fixed effects panel regression in the structural equation modeling (SEM) framework, as well as their application in the `lavaan` package for `R`. It is meant as a applied guide for researchers, covering the underlying model specification, syntax, and summary output.

This is an Accepted Manuscript of an article published by Taylor & Francis in *Structural Equation Modeling: A Multidisciplinary Journal*, available online: https://doi.org/10.1080/10705511.2021.1963255.

Introduction

Several years ago, Curran and Bauer (2011) reflected positively on the growing use of panel studies in empirical social research. Some of the strengths of panel data are well-known, e.g., the ability to establish temporal precedence, increased statistical power and the reduction of potential alternative models. However, the greatest strength of panel data is that they allow for a more rigorous testing of substantive theories. Panel data, i.e., repeated measures of the same observed units (people, schools, firms, countries, etc.), allow researchers to decompose the error term into a part that stays constant within units and the part that changes over time. The part that does not change over time can be seen as the combined effect of all time-invariant influences (e.g., sex, nationality, personality traits, intelligence) on the dependent variable (Bollen and Brand 2010). Random effects (RE) and fixed effects (FE) regression involve accounting for these often unobserved time-invariant influences via a number of methods.

In the case of RE models, it is assumed that the stable characteristics are unrelated to the other model covariates. So, the stable characteristics, also known as 'individual effects' (Hsiao 2014, the term I prefer), 'unobserved effects' (Wooldridge 2002) or 'unit effects' (Skrondal and Rabe-Hesketh 2004; Zyphur et al. 2020a, 2020b), will not affect point estimates but their existence violates important regression assumptions. FE models are useful when the individual effects are assumed to be related to one or more of the model covariates. In that case, these unobserved stable characteristics will act as confounders and lead to biased and inconsistent point estimates. In controlling for these individual effects, FE regression thus accounts for a likely and common source of bias.

Structural equation modeling (SEM) is a popular regression framework. One of its main strengths is its flexibility (Hoyle 2015b, 142). Not only can complex causal structures with multiple dependent variables be tested simultaneously, but in longitudinal (and, more generally, hierarchical) studies, both time-varying and invariant predictors can be included, and effects can easily be allowed to vary over time. Thus researchers can allow for and study effects that increase or fade over time, or that appear only in specific periods. Beyond that, with the use of latent variables, SEM provides a way to deal with measurement error and get closer to the theoretical constructs of interest.

The article focuses on the `lavaan` (Rosseel 2012) package for `R` (R Core Team 2017). While `Mplus` (L. Muthén and Muthén 1998--2017) is arguably the most robust SEM software currently available (in terms of features like alignment and latent variable interactions, for example), the `lavaan` package has many benefits. First, like `R` it is open source and completely free. For researchers new to SEM, there is no financial barrier to entry. Second, the implementation of `lavaan` in the larger `R` environment is an enormous advantage. Instead of poring over reams of plain text, copying out coefficients by hand, every part of the `lavaan` output is available as an object. This means that all aspects of the model, from fit indices, to coefficients and standard errors, to the model matrices, can be accessed and easily integrated into tables and plots. Furthermore, `R` can be used for a great deal of applications. It can be used to manage and manipulate as well as simulate data, perform symbolic algebra, run more traditional analyses (e.g., multiple regression, logistic regression, principal component analysis), etc. Once one is comfortable using `R`, there is (at the time of writing) little need to switch between different software for data preparation and analysis.

There are a number of articles describing the basic concept of panel regression, including RE and FE regression in SEM (e.g., P. D. Allison 2009; P. Allison 2011; Bollen and Brand 2010; Teachman et al. 2001). This article is intended as a practical guide for researchers looking for help specifying panel regression models in SEM. It assumes some basic knowledge of SEM (e.g., identification, the basic logic of estimation, model fit, etc.).[1]

The article proceeds by giving a short review of panel regression models (Sections to and) while touching on the implementation in SEM (Section). Then the `lavaan` package is discussed in Section and an annotated syntax for specifying RE and FE models in SEM is shown in detail in Section . An example using simulated data is given in Section to show and discuss the model output and results. Section provides a brief summary and points to some extended panel SEMs that build on the basic RE and FE framework (including latent

[1] For an introduction to SEM, Bollen (1989) is a classic for good reason, and Ferron and Hess (2007) lay out exactly how SEM works in one easy to follow article.

growth curves and dynamic models). Finally, an online section touches on further topics, including some drawbacks to panel SEM and their potential remedies, comparability with non-SEM methods, and extensions to the basic models (relaxing assumptions, dealing with measurement error).

For experienced SEM-users with little knowledge of static panel regression, the entire article may be of some use. For SEM-users already familiar with panel regression in SEM, the online section may be of the most interest. For those new to SEM but familiar with panel regression, Sections and (as well as the online section) are likely most relevant.

Random and Fixed Effects Panel models

It is typically the case that the values for a given unit on a variable at one point in time will tend to tell us something about that unit's values on the same variable at another point in time. There are two main explanations for this 'empirical regularity' (Heckman 1981; Hsiao 2014, 261; Brüderl and Ludwig 2015; Bianconcini and Bollen 2018). First, it could be that an experience at one point in time has an effect on future experiences. In other words, experiencing an event could change the probability of the same or a similar event taking place in the future. For example, if employment increases wages, then the incentive to continue working should increase over time, thereby increasing the likelihood that someone who was employed at one point in time will continue to be employed in the future (Heckman 1981). When past experiences impact future events, the empirical regularity is referred to as 'state dependence.' So-called 'dynamic' panel models that include the lagged dependent variable in the equation for the current dependent variable, like autoregressive and autoregressive cross-lagged models, are examples of panel models that account for state dependence (Zyphur et al. 2020a, 2020b). The second explanation is that correlations over time are due to stable characteristics, like sex, place of birth, motivation, ability or personality.[2] In other words, stable unit-specific characteristics might predispose individuals to experience events with a certain likelihood over time. For example, stable characteristics like sex or motivation could be part of the reason why some individuals tend to be continuously employed, while others experience spells of unemployment, or are habitually unemployed. This second source of empirical regularity is referred to as individual heterogeneity or, more generally, unobserved heterogeneity (Wooldridge 2012). So-called 'static' panel models (that do not include the lagged dependent variable in the equation for the current one) focus on this second source of empirical regularity (Zyphur et al. 2020a, 2020b). The random and fixed effects models that will be discussed here are examples of models that attempt to account for unobserved heterogeneity as a source of empirical regularity.

A review of static panel models and their implementation in SEM

Let us begin with a simplified cross-sectional model with a single covariate

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,\ i = 1, \ldots, N$$

where y_i is the dependent variable for unit i, β_0 is the intercept, x_i and z_i are scalar covariates that are considered random variables in most social science applications with observational data, β_1 and β_2 are the associated regression coefficients and ε_i is the normal idiosyncratic error. We can think of the intercept, β_0, as 'absorbing' the expectation of the error, $\mathbb{E}(\varepsilon_i)$, so that it is safe to assume that $\mathbb{E}(\varepsilon_i) = 0$.[3]

We typically assume the following (Wooldridge 2002; B. O. Muthén, Muthén, and Asparouhov 2016; Bollen and Brand 2010):

[2] This article will tend to refer to individuals as the unit of analysis, but others, such as families, schools and countries, are of course possible.

[3] Take a simplified equation without an intercept: $y_i = \beta_1 x_i + \varepsilon_i$. We can add and subtract the expectation of ε_i without changing the equation: $y_i = \beta_1 x_i + \varepsilon_i + \mathbb{E}(\varepsilon_i) - \mathbb{E}(\varepsilon_i)$. Now, we can define $\beta_0 = \mathbb{E}(\varepsilon_i)$ and $\varepsilon_i^* = \varepsilon_i - \mathbb{E}(\varepsilon_i)$ for $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i^*$ where $\mathbb{E}(\varepsilon_i^*) = 0$, see Wooldridge (2009), p. 25, 61; or https://stats.stackexchange.com/questions/231549/expected-value-of-error-term-equals-zero-formal-proof. Throughout, this is taken for granted and the notation with the star is dropped.

1. Linearity: the model can be written as $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$
2. Variation in covariates, no perfect multicollinearity
3. Zero conditional mean: $\mathbb{E}(\varepsilon_i|x_i, z_i) = \mathbb{E}(\varepsilon_i) = 0$, which implies $\text{Cov}(\varepsilon_i, x_i) = \text{Cov}(\varepsilon_i, z_i) = 0$ and $\mathbb{E}(y_i|x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i$
4. $\text{Var}(\varepsilon_i|x_i, z_i) = \text{Var}(\varepsilon_i) = \sigma^2_\varepsilon$ and $\text{Cov}(\varepsilon_i, \varepsilon_j|x_i, z_i, x_j, z_j) = \text{Cov}(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$ which implies homoskedasticity (constant variance) and uncorrelated errors
5. Normally distributed errors: $\varepsilon_i|x_i, z_i \sim N(0, \sigma^2_\varepsilon)$, for significance testing.

In a cross-sectional setting, we could estimate the parameters of interest, β_0, β_1, β_2 by ordinary least squares (OLS) or maximum likelihood (ML), where it is widely known that if the assumption of normally distributed errors holds, the OLS estimator is the ML estimator.[4]

Now, assume we are dealing with repeated measures of the same units over time so we can write

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \varepsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{1}$$

where the subscript it stands for unit i at time t (P. D. Allison 2009). Now, x_{it} is referred to as a 'time-varying' covariate because it changes across i and t, while z_i is a 'time-invariant' covariate because it changes over i but not t. Time-varying covariates encompass characteristics that can change over time, like mood, stress level, income, number of children, etc. Time-invariant covariates do not change over time, or at least not usually over the period of observation. These can be things like sex, nationality, upbringing, etc. β_0 is the overall intercept, absorbing the expected value of ε_{it}. Sometimes we will want to allow the intercept to vary over time. These are sometimes referred to as 'occasion effects' (Zyphur et al. 2020b) and can be modeled by including time dummies. We will ignore these for now, and allow the overall intercept to absorb the expectation of the error over i and t.

Now, there is nothing practical stopping us from stacking all NT observations on top of one another and estimating the model using OLS or ML. This is called the 'pooled' OLS (POLS) estimator with an analogous ML estimator. The problem with this approach is that it ignores the panel nature of the data. As such, the assumption of uncorrelated errors is likely violated within units. In other words, there are likely unit-specific characteristics that do not change over time affecting the dependent variable, many of which will remain *unobserved*. We can call the sum of all these effects on the dependent variable *individual effects* or *unobserved heterogeneity* and write them as α_i with the *composite error*: $\varepsilon_{it} = \alpha_i + \nu_{it}$. Then Equation (1) becomes

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \alpha_i + \nu_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T. \tag{2}$$

The individual effects α_i are treated as a random variable, even if the use of the Greek letter makes it look like another parameter to be estimated. This is the common notation used in the literature on panel regression in SEM, see for example P. Allison (2011); Bollen and Brand (2010). When the number of cross-sectional units is large, Wooldridge (2002) (p. 247 and p. 252) suggests *always* treating the unobserved heterogeneity as a random variable rather than a fixed unknown parameter since little is lost in doing so. Notice this equation still contains an overall intercept, β_0. This will absorb the expected value of α_i as well.

Now, because

$$\mathbb{E}(\varepsilon_{it}\varepsilon_{js}) = \mathbb{E}[(\alpha_i + \nu_{it})(\alpha_j + \nu_{js})]$$

[4] In this section, we will focus on the typical least squares-based treatment of RE and FE models because the link between the ML-based multilevel or hierarchical models and multilevel and panel SEM is already arguably fairly strong, see for example Hox (2010), in which SEM is made reference to constantly, and the final two chapters are dedicated to SEM-based approaches to multilevel modeling.

the assumption about uncorrelated errors becomes

$$\mathbb{E}(\varepsilon_{it}\varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2, & i = j,\ t = s \\ \sigma_\alpha^2, & i = j,\ t \neq s \\ 0, & \text{otherwise} \end{cases}$$

where $\sigma_\varepsilon^2 = \sigma_\alpha^2 + \sigma_\nu^2$, which is the sum of the variances of the individual effects and idiosyncratic errors. This follows if we assume the individual effects are unrelated to the idiosyncratic errors at all points in time while retaining the independent errors assumption from above as it pertains to the idiosyncratic errors (Hsiao 2014, 40). This means that the variance of the errors can no longer be described "up to a multiplicative constant" (Wooldridge 2002, 153), i.e., $\sigma_\varepsilon^2 \boldsymbol{I}$ (where $\boldsymbol{I}$ is the identity matrix with ones down the diagonal and zeros everywhere else), and the pooled model is no longer appropriate.

Let us write Equation (2) in matrix notation

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\iota}_T\alpha_i + \boldsymbol{\nu}_i,\ i = 1, \dots, N \end{aligned} \tag{3}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{iT})^\intercal$, $\boldsymbol{X}_i = (\boldsymbol{\iota}_T, \boldsymbol{x}_i, \boldsymbol{\iota}_T z_i)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\intercal$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{iT})^\intercal$, $\boldsymbol{\iota}_T$ is a vector of T ones (sometimes written as $\mathbf{1}$), and write the covariance matrix of the errors as

$$\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\intercal) = \boldsymbol{\Omega}_i = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\varepsilon^2 \end{bmatrix}. \tag{4}$$

The full covariance matrix over all i takes a block-diagonal form with $\boldsymbol{\Omega}_i$ down the diagonal and $\mathbf{0}$ everywhere else (Wooldridge 2002, 259; Schmidheiny 2019). Notice the homoskedasticity assumption here now applies across time, as well.

Random effects

In a typical random effects (RE) model, a transformation to Equation (3) is carried out such that the covariance matrix of the errors becomes the identity matrix, i.e.,

$$\begin{aligned} \boldsymbol{\Omega}_i^{-1/2}\boldsymbol{y}_i &= (\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{X}_i)\boldsymbol{\beta} + \boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\varepsilon}_i, \\ \boldsymbol{y}_i^* &= \boldsymbol{X}_i^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^* \end{aligned} \tag{5}$$

(Wooldridge 2002, 155). This is akin to dividing a random variable by its standard deviation in order to standardize it with a variance of one. In this case, we have

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_i^*\boldsymbol{\varepsilon}_i^{*\intercal}) &= \mathbb{E}[(\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\varepsilon_i})(\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\varepsilon}_i)^\intercal] \\ &= \mathbb{E}(\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\intercal\boldsymbol{\Omega}_i^{-1/2\intercal}) \\ &= \boldsymbol{I}_T \end{aligned}$$

(Wooldridge 2002, 155). The variances of σ_ε^2 and σ_α^2 are not known, but they can be estimated by the sample variances, see Wooldridge (2002) p. 260 f.; Schmidheiny (2019). With the transformation, Equation (5) can be estimated by OLS which is referred to as the (feasible) generalized least squares (GLS) estimator Brüderl and Ludwig (2015).

Notice the model puts the unobserved characteristics in α_i in the composite error term. Consistency and unbiasedness thus hinge on the assumption that the model covariates are unrelated to the *composite error*, i.e., $\mathbb{E}(\varepsilon_{it}|\boldsymbol{X}_i) = \mathbb{E}(\varepsilon_{it}) = 0$ or $\mathbb{E}(\nu_{it}|\boldsymbol{X}_i) = \mathbb{E}(\nu_{it}) = 0$ *and* $\mathbb{E}(\alpha_i|\boldsymbol{X}_i) = \mathbb{E}(\alpha_i) = 0$:[5]

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{y}_i) \\
&= \boldsymbol{\beta} + (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{\varepsilon}_i) \\
&= \boldsymbol{\beta} + (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} (\boldsymbol{\iota}_T \alpha_i + \boldsymbol{\nu}_i)) \\
&= \boldsymbol{\beta} + (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{\iota}_T \alpha_i + \sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{\nu}_i). \qquad (6)
\end{aligned}$$

Equation (6) shows that consistency and unbiasedness depend on the relationship between the model covariates and the individual effects and idiosyncratic errors (take the probability limit and expectation to see that consistency requires the covariances to be zero, while unbiasedness requires conditional independence).

Notice, as well, that the RE model requires the assumption that the composite errors at *each point in time* are unrelated to the model covariates at *all points in time*.[6] This is the *strict exogeneity* assumption, and it comes from the transformation in Equation (5), where

$$\mathbb{E}(\boldsymbol{X}_i^\intercal \boldsymbol{\Omega}_i^{-1} \boldsymbol{\varepsilon}_i) = \mathbf{0}$$

involves multiplying $\boldsymbol{X}_i$ (which holds x_{it} at all points in time) by the errors at each point in time. For this to be zero, it is therefore not enough to assume contemporaneous exogeneity, which is sufficient for pooled OLS (Wooldridge 2002, 146).

Fixed effects

The critical assumption in the RE model is $\mathbb{E}(\varepsilon_{it}|\boldsymbol{X}_i) = 0$, which implies $\mathbb{E}(\nu_{it}|\boldsymbol{X}_i) = 0$ and $\mathbb{E}(\alpha_i|\boldsymbol{X}_i) = 0$. This means that any of the stable characteristics contained in α_i are assumed mean independent of the model covariate(s), though typically, it is enough to look at the covariance between α_i and the model covariate(s), which accounts for any linear relationship, at least (see, for example, Bollen and Brand 2010, 6; Wooldridge 2002, 252). As an example, if we were interested in the effect of social isolation (x_t) on well-being (y_t), this would mean that any and all (potentially unobserved) personality characteristics (α), like extroversion or neuroticism, affecting someone's usual well-being level cannot be correlated with social isolation (Seifert and Andersen 2020). This is obviously implausible because extroverted people tend to engage more socially and neurotic people less so. In this case, the assumption of unrelated effects is likely violated and the RE model will return biased and inconsistent coefficient estimates for the effect of social isolation on well-being.

The conventional FE model relaxes this assumption by transforming Equation (2) such that the individual effects are eliminated. This usually entails *demeaning* the equation by subtracting the respective person- or unit-means from each of the variables

[5] In applied settings, it is normally enough to establish that the correlation between the individual effects and the model covariates are zero (Bollen and Brand 2010; Wooldridge 2002, 254). Also note that assuming the unconditional expectations are zero is unproblematic as long as an intercept is included (Wooldridge 2002, 257).

[6] Contemporaneous exogeneity would be $\mathbb{E}(\varepsilon_{it}|\boldsymbol{x}_{it}) = 0$, where, in this case, $\boldsymbol{x}_{it} = (1, x_{it}, z_i)$ (Wooldridge 2002).

$$
\begin{aligned}
y_{it} - \bar{y}_i &= (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)\boldsymbol{\beta} + \alpha_i - \bar{\alpha}_i + \nu_{it} - \bar{\nu}_{it} \\
\ddot{y}_{it} &= \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{\nu}_{it}, \\
\ddot{\boldsymbol{y}}_i &= \ddot{\boldsymbol{X}}_i\boldsymbol{\beta} + \ddot{\boldsymbol{\nu}}_i
\end{aligned}
$$

where the variables with the dots above them represent the demeaned versions and $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ and the person-means of the other variables are calculated analogously. Notice the average of something that does not change is that thing itself, so α_i, the individual effects, as well as any time-invariant variables in $\boldsymbol{x}_{it}$ are eliminated by the transformation (Brüderl and Ludwig 2015; Wooldridge 2002; Angrist and Pischke 2009). Now, the strict exogeneity assumption applies only to the idiosyncratic errors $\ddot{\nu}_{it}$. That is, the idiosyncratic error at each point in time is assumed to be uncorrelated with the model covariate(s) at all points in time.

Random and fixed effects in SEM

In this section, I explain the implementation of RE and FE models in SEM in a non-technical way. For a more detailed explanation, see the online section.

We can use path models like the one in Figure 1 to help us convey the model and its assumptions quickly and intuitively. Figure 1 shows a four-wave RE-SEM like the one discussed above. The observed variables are conventionally shown as squares or rectangles. Latent variables, like α_i, are shown as circles or ellipses. Effects and factor loadings are shown as directional arrows, coming out of the independent variable and pointing towards the dependent variable. Covariances or correlations are shown as bidirectional arrows between variables. The errors, here ν_{it}, are not enclosed in anything. They point to the time-varying dependent variable, y_{it}, with an implicit coefficient of one. The circular arrows around the errors indicate an endogenous error variance to be estimated. The fact that each is labelled σ^2_ν signifies that the variance has been constrained to be equal over time. Sometimes, intercepts are shown as a triangle labelled 1, with arrows pointing to each of the variables for which an intercept or mean should be estimated. For the sake of legibility, this is not shown here, although we allow an intercept to be estimated for each y_{it}, as well as unconditional means for each x_{it} and z_i. The mean of α_i is again set to zero for identification purposes (we can only estimate as many intercepts and means as there are observed variables). The unconditional mean of α_i will be absorbed into the time-varying intercepts κ_{y_t}.

[Figure 1 about here.]

For the RE and FE models, we switch from the long format setup (where each model variable is $NT \times 1$) to wide format, i.e., T individual $N \times 1$ columns for each variable. Then, we write T individual equations:

$$
\begin{aligned}
y_{i1} &= \kappa_{y_1} + \beta x_{i1} + \gamma z_i + \alpha_i + \nu_{i1}, \; i = 1, \ldots, N \\
y_{i2} &= \kappa_{y_2} + \beta x_{i2} + \gamma z_i + \alpha_i + \nu_{i2}, \; i = 1, \ldots, N \\
&\vdots \\
y_{iT} &= \kappa_{y_T} + \beta x_{iT} + \gamma z_i + \alpha_i + \nu_{iT}, \; i = 1, \ldots, N.
\end{aligned} \tag{7}
$$

where we have a single time-varying covariate, x_{it}, and a time-invariant covariate, z_i, along with the unobserved individual effect, α_i, that captures the sum of all other time-invariant *unobserved* variables (Bollen and Brand 2010; Wooldridge 2002). Note that the change of notation compared to the base text (from β_1 to β, β_2 to γ) is inconsequential.

In this setup, with a separate equation per timepoint, there is nothing stopping us from allowing each of the coefficients (even the implicit coefficient in front of α_i) to be estimated freely. In fact, while it is the

standard practice in the long-format least squares models discussed above, when we write the model like this, it may even seem arbitrary to constrain the coefficients to be time-invariant across equations. I retain these constraints for now, in order to establish a baseline equivalence with the least squares-based models. In the supplementary section, models will be discussed that do allow for time-varying coefficients. Something that does separate the SEMs from the conventional models are the intercepts. With T individual equations, it is straightforward to have a freely estimated intercept, κ_{y_t}, per timepoint. Note as well that we could easily include other time-varying and invariant coefficients and turn x_{it} and z_i into vectors.

Now, the assumption of constant effects over time is conveyed by the lack of a subscript t on the coefficients, β and γ. The latent individual effects are also assumed constant and constrained to 1 at each timepoint. Again, these can easily be relaxed but are retained for now for the sake of establishing baseline equivalence with the least squares-based models. Strict exogeneity in the RE model means the idiosyncratic errors, as well as the individual effects are uncorrelated with the model covariates at each point in time. This is conveyed by the lack of any two-headed arrows between them. Notice the idiosyncratic errors, ν_{it}, are specific to a given timepoint, while the individual effects, α_i, is common to all timepoints (P. D. Allison 2009). This is how we translate the structure seen in $\mathbf{\Omega}_i$ in Equation (4). Once we have converted the data to wide-format, the conditional covariance between any two columns of the dependent variable, $\text{Cov}(y_{it}, y_{is}|x_{it}, x_{is}, z_i)$, is just the conditional variance of the individual effects, $\text{Var}(\alpha_i|x_{it}, x_{is}, z_i)$. In the RE case, it is assumed that α_i is independent of all the other model covariates, so $\text{Var}(\alpha_i|x_{it}, x_{is}, z_i) = \text{Var}(\alpha_i)$. In other words, by regressing y_{it} onto the latent variable α_i at each point in time, we capture the unobserved heterogeneity; the common variance in y_{it} due to unobserved time-invariant characteristics. Finally, homoskedasticity of the idiosyncratic errors over time is conveyed by labelling each of the idiosyncratic error variances (the circular arrows around ν_{it}) the same over time. It can be difficult to create path models that effectively convey all the important information of a model, but they at least have the potential to express such information in a fairly intuitive and non-technical way.

For the FE model, Equation (7) from the RE setup pertains here as well. We relax the assumption of unrelated effects by allowing the individual effects, α_i, to correlate with the time-varying model covariates at each point in time, rather than fixing them to zero. The assumption of $\text{Cov}(\alpha_i, z_i) = 0$ must be retained for identification. Again, we can represent the FE-SEM in a non-technical way by altering the path diagram to include the covariances between each x_{it} and α_i, see the new two-headed arrows in Figure 2.

[Figure 2 about here.]

Random and fixed effects in lavaan

The basic panel models discussed in this tutorial consist of three to four main parts. First, we want to create a latent variable representing the individual effects. Second, we regress the time-varying dependent variable on the time-varying and time-invariant covariates. Third, we specify the correlations depending on our assumptions. If we believe the individual effects are unrelated to the time-varying covariates (we must assume they are unrelated to the time-invariant ones, otherwise the model is not identifiable), then we apply the RE assumptions and constrain the covariances between the individual effects and the time-varying covariates to zero. If we believe the individual effects are indeed related to the model covariates (or more realistic assumption in many circumstances), then we must specify these correlations between the individual effects and the time-varying covariates. Finally, in an optional step, we can constrain the residual variances to be equal over time, if we want or need to, potentially in order to save degrees of freedom. The following section will explain these steps in `lavaan` in detail.

The `lavaan` package in `R`

The package `lavaan` needs to be installed once with `install.packages("lavaan")`. This can be entered directly into the console or at the top of the script. To be able to use the package, we need to load it for every new `R` session:

```
library(lavaan)
```

For users unfamiliar with R, SEM analyses can be carried out with almost no knowledge of the language. Typically, someone unfamiliar with R would prepare their data using some other statistical software, and then save the intended dataset as a .csv, .xlsx, .dta, .sav, etc. file. The user must then import the data, preferably as a dataframe, and the rest occurs using the lavaan syntax.[7]

To use lavaan, we create an R object using the assignment operator <-, see the model syntax example below. Here, the object has been called fe_sem because in the following example we will be assuming an FE model is appropriate. The object can be named anything that complies with naming conventions in R (e.g., the object name must start with a letter or dot, underscores and dots can be used to separate words, etc.). The model syntax is enclosed in quotes, either single '' or double "". This means that the model syntax is essentially a string that the lavaan package interprets in a second step. Once the model has been specified, we use the sem() function to 'fit' the model. Notice a second object is made out of the fitted lavaan object. Here the fitted lavaan object has been named fe_sem.fit.

To reiterate, we specify the SEM by writing the model syntax as a string and saving it as an object. Then, in a second step, we run the sem() function on that object. The sem() function requires at least two arguments: model, i.e., the model object (here: fe_sem), and data, i.e., the dataframe or covariance matrix (along with the mean vector, if desired). That is, at a bare minimum, we must tell lavaan how the model is specified and where the data is. There are a number of other optional arguments that can be included. If they are not, the defaults of the sem() wrapper are used.[8] For this example, even though maximum likelihood is the default estimator, estimator = "ML" has been included as an optional argument to emphasize the fact. The online section provides some guidance on dealing with nonnormal endogenous variables and missing values, both of which can be addressed with optional arguments in the sem() function call. Finally, we add the argument meanstructure = TRUE. If we leave this out, the model is essentially fit to centered data so that each variable has a mean of zero. When we turn the mean structure on, the default behaviour is to estimate an intercept or mean (depending on whether the variable is exogenous or endogenous) per equation.

Model syntax

Again, specifying the most basic random or fixed effects model, like the one shown in Bollen and Brand (2010) and described in Equation (2) involves three to four components. First, we define the latent individual effects variable using the =~ 'measured by' or 'manifested by' (Rosseel 2012) operator while at the same time constraining the factor loadings at each timepoint to one with 1* (see line 3 in the model code below). I will call the latent variable alpha. Constraining all of the factor loadings to one reflects our implicit assumption that the combined effect of the unit-specific unobserved factors is constant over time (Zyphur et al. 2020a, 660). This is the default behaviour of traditional least squares-based approaches to RE and FE that use the stacked long-format data.

Second, we regress the dependent variable on the independent variable using the ~ regression operator (see lines 5–8). With stacked, long-format data, only one regression coefficient per covariate is estimated over all observed timepoints. To have our model mimicking this behaviour, we need to constrain the the estimated coefficients to equal over time. We do so by adding the same label to the regression coefficient at every time point. We will use the labels beta and gamma (though we could use any letter or string of characters for labels) and have them act as equality constraints for the regression coefficient of interest, β and γ.

[7] There are many online tutorials for importing data in various formats, see for example one of the many posts on stackoverflow (https://stackoverflow.com/search?q=r+import+data).

[8] The main defaults of the sem() wrapper are: intercepts of the latent variables set to zero, the first factor loading in factor models is set to one, the residual variances and variances of exogenous latent variables are freely estimated, exogenous latent variables are set to covary. Further details can be found at https://rdrr.io/cran/lavaan/man/sem.html and https://cran.r-project.org/web/packages/lavaan/lavaan.pdf, or by entering lavOptions() into the console to get a full list of defaults. An explanation of the optional arguments can be found by entering ?lavOptions in the console. There are other 'wrappers' with slightly different default options, like cfa() for example, see the lavaan tutorial website at https://lavaan.ugent.be/tutorial/cfa.html.

The key to a FE model, as opposed to an RE model are our assumptions about the relatedness of our time-varying covariates and the individual effects. For an FE model, we want to partial out any potential covariance between the independent variable and the individual effects. This accounts for any linear relationship between x_{it} and the unit-specific characteristics influencing the dependent variable. Further, allowing unrestricted covariances between the independent variable itself over time will not affect how the coefficients are estimated, but will improve model fit (see lines 10–14).[9] To mimic the behaviour of a conventional FE model, we allow the independent variable to be correlated with the individual effects and itself over time. Covariances (including covariances between a variable and itself, i.e., variances) are specified using the `~~` operator. For the RE model, we would constrain the covariances between `alpha` and the time-varying and invariant covariates with `alpha ~~ 0*x1 + 0*x2 + 0*x3 + 0*x4 + 0*z`.

The last component of our code involves the variances of the residuals (lines 16–19). This component is optional, but we can constrain the residual variances to be equal over time to again mimic the behaviour of a conventional RE/FE model using least squares-based methods on stacked data. Here, again, we use labels to make equality constraints. Because y_{it} is endogenous, the `~~` operator specifies the variances of *residuals*, i.e., ν_{it}.

```
1  fe_sem <- '
2  # Define individual effects variable
3  alpha =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
4  # Regressions, constrain coefficient to be equal over time
5  y1 ~ beta*x1 + gamma*z
6  y2 ~ beta*x2 + gamma*z
7  y3 ~ beta*x3 + gamma*z
8  y4 ~ beta*x4 + gamma*z
9  # Correlations, individual effects covary with the covariate
10 alpha ~~ x1 + x2 + x3 + x4 + 0*z
11 x1 ~~ x2 + x3 + x4 + z
12 x2 ~~ x3 + x4 + z
13 x3 ~~ x4 + z
14 x4 ~~ z
15 # Constrain residual variances to be equal over time
16 y1 ~~ nu*y1
17 y2 ~~ nu*y2
18 y3 ~~ nu*y3
19 y4 ~~ nu*y4
20 '
21 fe_sem.fit <- sem(model = fe_sem,
22                   meanstructure = TRUE,
23                   data = dfw,
24                   estimator = "ML")
```

A simulated example

To demonstrate the application of panel regression in SEM, a dataset can be simulated that embodies the FE assumptions. The code for the simulation can be found in the online supplementary materials. The simulated data was constructed such that the time-varying covariate of y_{it} is correlated with two separate time-invariant variables. One of these time-invariant variables will be considered observed; the other will be

[9] If we do not specify the correlations between the covariate over time, we are telling the model we believe them to be zero. This is unrealistic in most cases and discrepancy between the model-implied covariances of zero and the likely non-zero observed covariances will be a source of misfit. I.e., the χ^2 statistic will be larger than it needs to be (there is not usually a good reason to fix covariate correlations to zero) potentially leading to otherwise well-fitting models being rejected by the test statistic.

treated as unobserved. This means that approaches that fail to account for this confounding influence, such as POLS or RE, will be biased.

The equations for the data generating process (DGP) can be described as:

$$x_{it} = \beta_{x_t,z} z_i + \beta_{x_t,\alpha} \alpha_i + \delta_{it}$$
$$y_{it} = \beta_{y_t,x_t} x_{it} + \beta_{y_t,z} z_i + \beta_{y_t,\alpha} \alpha_i + \nu_{it}$$

where $\alpha_i \sim N(2, 1^2)$, $z_i \sim N(1.5, 3^2)$, $\delta_{it} \sim N(3, 2^2)$, $\nu_{it} \sim N(3, 1^2)$ and the means and variances were chosen arbitrarily. Correlations between x_{it} and both z_i and α_i are induced by making x_{it} dependent on each of them. Note that the mean of ν_{it} is not zero, but that it will be absorbed into the intercept for y_{it}. The coefficients were also chosen arbitrarily and set to $\beta_{x_t,z} = 0.5$, $\beta_{x_t,\alpha} = 0.85$, $\beta_{y_t,x_t} = 0.30$, $\beta_{y_t,z} = 0.45$ and $\beta_{y_t,\alpha} = 0.75$. For the following example, a sample size of 1,000, observed over four waves, was chosen.

We can get a summary of the model with `summary()`. Optional arguments for `summary()` are, for example, `standardized = TRUE` for standardized coefficients, `fit.measures = TRUE` for further (e.g., comparative) fit indices. The first portion of the summary output gives an overview of some basic information and fit statistics. The maximum likelihood estimator is the default, so it did not have to be explicitly selected in the fitting function call. Other estimators are available, including generalized and unweighted least squares (`GLS` and `ULS`, respectively), robust standard errors maximum likelihood (`MLM`) and several others (see the lavaan online tutorial at https://lavaan.ugent.be/tutorial/est.html as well as the online section).

This part of the summary output also tells us that the analysis is based on 1,000 observations (missings would be shown here as well if there were any), and that the χ^2 statistic is 18.21 based on 22 degrees of freedom (45 observed (co)variances minus 1 error variance, 2 coefficients, 1 latent variable variance, 5 exogenous variable variances and 14 covariances for $45 - 23 = 22$ df). The p-value on the χ^2 statistic is not significant with $p = 0.693$ which suggests we should retain the model (given how the data was generated, it would be surprising if this were not the case).

```
summary(fe_sem.fit)
```

```
## lavaan 0.6-8 ended normally after 92 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        41
##   Number of equality constraints                     9
##
##   Number of observations                          1000
##
## Model Test User Model:
##
##   Test statistic                                18.210
##   Degrees of freedom                                22
##   P-value (Chi-square)                           0.693
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
...
```

Next, the summary output shows the measurement models for the latent variables, if any. In this case the

latent variable, `alpha`, is measured by each of the four observed dependent variables with factor loadings fixed to 1.0.

```
...
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   alpha =~
##     y1                1.000
##     y2                1.000
##     y3                1.000
##     y4                1.000
...
```

The regressions are shown next. Here, because we have constrained the regression coefficients to be equal over time (the equality constraint label `(beta)` and `(gamm)` is listed to the left of the estimates), the estimates of $\beta = 0.298$ (0.01) and $\gamma = 0.456$ (0.01) are both repeated four times. The corresponding z- and p-values show that the coefficients are, unsurprisingly based on the simulated DGP, significant.

```
...
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   y1 ~
##     x1      (beta)    0.298    0.010   31.112    0.000
##     z       (gamm)    0.456    0.010   43.476    0.000
##   y2 ~
##     x2      (beta)    0.298    0.010   31.112    0.000
##     z       (gamm)    0.456    0.010   43.476    0.000
##   y3 ~
##     x3      (beta)    0.298    0.010   31.112    0.000
##     z       (gamm)    0.456    0.010   43.476    0.000
##   y4 ~
##     x4      (beta)    0.298    0.010   31.112    0.000
##     z       (gamm)    0.456    0.010   43.476    0.000
...
```

Next, the covariance estimates are listed. First, the covariances between the latent individual effects and the independent variables are shown, followed by the covariances between the independent variables themselves.

One should always take care to double-check that there are no unintended covariances listed here. Like `Mplus`, the `lavaan` package estimates some covariances per default, without the user explicitly having to add them to the model syntax. For example, covariances between latent variables are estimated per default. If one does not wish for them to covary, it must be explicitly stated, e.g., with `f1 ~~ 0*f2`, assuming the latent variables are called `f1` and `f2`, or by overriding the default behaviour for the entire model by adding `orthogonal = TRUE` (which sets the correlation between all latent variables to zero) to the fitting call.[10]

```
...
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   alpha ~~
##     x1                0.680    0.071    9.640    0.000
##     x2                0.617    0.070    8.797    0.000
```

[10] This is at least the current behaviour of both the `cfa` and `sem` wrappers. In fact, both wrappers seem to be identical in terms of the default settings, see Rosseel et al. (2020).

```
##     x3                0.660    0.070    9.470    0.000
##     x4                0.700    0.066   10.547    0.000
##     z                 0.000
##   x1 ~~
##     x2                3.472    0.260   13.337    0.000
##     x3                3.528    0.256   13.791    0.000
##     x4                3.118    0.237   13.153    0.000
##     z                 4.919    0.311   15.814    0.000
##   x2 ~~
##     x3                3.536    0.263   13.459    0.000
##     x4                3.534    0.249   14.186    0.000
##   z ~~
##     x2                5.363    0.326   16.459    0.000
##   x3 ~~
##     x4                3.241    0.240   13.498    0.000
##   z ~~
##     x3                5.115    0.316   16.192    0.000
##     x4                4.739    0.295   16.065    0.000
...
```

Next the intercepts and means are shown. The intercepts of the endogenous variables are marked with a dot (`.`) beside the variable name. The variables without the dot are exogenous and so the estimates refer to unconditional means. x_{it} is treated as an exogenous variable in the model, but its unconditional mean can be derived from its equation,

$$\begin{aligned}\mathbb{E}(x_{it}) &= \beta_{x_t,z}\,\mathbb{E}(z_i) + \beta_{x_t,\alpha}\alpha_i + \mathbb{E}(\delta_{it}) \\ &= 0.5(1.5) + 0.85(2.0) + 3.0 = 5.45,\end{aligned}$$

where we simply plug in the values from the simulation description above. For y_{it}, since it is an endogenous variable, we are interested in the intercept. We do not have enough empirical information to estimate means for the individual effects and the idiosyncratic errors, so we let them get 'absorbed' into the intercepts. The expected value of y_{it} is

$$\begin{aligned}\mathbb{E}(y_{it}) &= \kappa_{y_t} + \beta_{y_t,x_t}\,\mathbb{E}(x_{it}) + \beta_{y_t,z}\,\mathbb{E}(z_i) \\ &= (0.75(2.0) + 3.0) + 0.30(5.45) + 0.45(1.5) = 6.81\end{aligned}$$

where, again, $\kappa_{y_t} = \mathbb{E}(\alpha_i) + \mathbb{E}(\nu_{it})$ 'absorbs' the expectations of α_i and ν_{it}. The equation can be rearranged in terms of κ_{y_t},

$$\begin{aligned}\kappa_{y_t} &= \mathbb{E}(y_{it}) - \beta_{y_t,x_t}\,\mathbb{E}(x_{it}) + \beta_{y_t,z}\,\mathbb{E}(z_i) \\ &= 6.81 - 0.30(5.45) + 0.45(1.5) = 4.5.\end{aligned}$$

which is roughly the intercept shown below in the output.[11]

```
...
## Intercepts:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .y1                4.469    0.062   71.781    0.000
##    .y2                4.520    0.062   72.818    0.000
```

[11] The values in the output will differ slightly from the theoretical values due to sampling error.

```
##    .y3                4.496    0.061   73.230    0.000
##    .y4                4.546    0.062   72.789    0.000
##     x1                5.438    0.085   63.970    0.000
##     z                 1.431    0.100   14.286    0.000
##     x2                5.412    0.088   61.642    0.000
##     x3                5.315    0.086   62.067    0.000
##     x4                5.467    0.080   68.186    0.000
##     alpha             0.000
...
```

Finally, the variance estimates are listed. Here, we see that in order to mimic the behaviour of a traditional FE model, the error variances over time were specified to be equal using the equality constraint `(nu)`. Notice again the `.` beside `y1`, `y2`, etc.: this indicates that the listed variance refers to an endogenous variable, and that it is thus an error variance. In this case, these refer to the variances of ν_t. After that, the variances of the exogenous variables, both observed and unobserved are listed.

```
...
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .y1        (nu)    1.039    0.027   38.730    0.000
##    .y2        (nu)    1.039    0.027   38.730    0.000
##    .y3        (nu)    1.039    0.027   38.730    0.000
##    .y4        (nu)    1.039    0.027   38.730    0.000
##     x1                7.227    0.323   22.361    0.000
##     z                10.040    0.449   22.361    0.000
##     x2                7.709    0.345   22.361    0.000
##     x3                7.332    0.328   22.361    0.000
##     x4                6.429    0.288   22.361    0.000
##     alpha             0.611    0.042   14.711    0.000
```

Testing against a random effects model

The RE model differs from the FE model in that it assumes there is no correlation between the model covariates and the individual effects. We can test this assumption by running an RE model and then comparing it with the FE one. If the model fit for the RE model is not substantially worse, it would be an indication that the could retain the assumption $\text{Cov}(x_{it}, \alpha_i) = 0$ and choose a RE specification. Because the RE model is *nested* within the FE model, i.e., we can get the RE model by fixing some of the parameters in the FE model to zero Bollen and Brand (2010), we can perform a likelihood ratio test and see if the difference in model fit is significant, or not.

The RE model is achieved by simply fixing the covariances in line 10 in the code above to zero to represent the assumption that $\text{Cov}(x_{it}, \alpha_i) = 0$.

```
1  re_sem <- '
2  # Define individual effects variable
3  alpha =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
4  # Regressions, constrain coefficient to be equal over time
5  y1 ~ beta*x1 + gamma*z
6  y2 ~ beta*x2 + gamma*z
7  y3 ~ beta*x3 + gamma*z
8  y4 ~ beta*x4 + gamma*z
9  # Correlations, individual effects do not covary with the covariate
10 alpha ~~ 0*x1 + 0*x2 + 0*x3 + 0*x4 + 0*z
```

```
x1 ~~ x2 + x3 + x4 + z
x2 ~~ x3 + x4 + z
x3 ~~ x4 + z
x4 ~~ z
# Constrain residual variances to be equal over time
y1 ~~ nu*y1
y2 ~~ nu*y2
y3 ~~ nu*y3
y4 ~~ nu*y4
'
re_sem.fit <- sem(model = re_sem,
                  meanstructure = TRUE,
                  data = dfw,
                  estimator = "ML")
```

The RE model shows a χ^2 statistic of 299.997 based on 26 degrees of freedom. Contrast this with the χ^2 statistic of the FE model of only 18.210 (22). The difference in degrees of freedom comes from the four covariances that are constrained to zero in the RE model and estimated freely in the FE one.

The estimated coefficient for x_{it} is $\hat{\beta}_{RE} = 0.363^{***}$ (0.009), which is substantially larger than the effect of the FE model, which was $\hat{\beta}_{FE} = 0.298^{***}$ (0.010). This is because we have a kind of omitted variable bias stemming from the incorrect assumption that $\text{Cov}(x_{it}, \alpha_i) = 0$.

As mentioned, because the RE model is nested within the FE one, we can perform a likelihood ratio test to determine if the difference in model fit is significant, or not. In `R`, we can use the `anova()` function and supply it with the fitted models from above to perform a likelihood ratio test:

```
anova(fe_sem.fit, re_sem.fit)
```

```
## Chi-Squared Difference Test
##
##            Df   AIC   BIC   Chisq Chisq diff Df diff Pr(>Chisq)
## fe_sem.fit 22 34895 35052  18.209
## re_sem.fit 26 35169 35307 299.997     281.79       4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table shows a comparison of the nested models, in descending order according to degrees of freedom. The RE model does not estimate the correlations between the individual effects and the covariates, so it is more parsimonious and thus listed at the bottom. The `Chisq` column shows the χ^2 statistic for both models and the `Chisq diff` column calculates the difference between the two. Obviously, according to the simulated data generating process (DGP), the correlation between the individual effects and x_{it} is not zero, so fixing these to zero leads to a substantial amount of misfit. The last column puts the χ^2 difference in relation to the difference in degrees of freedom and gives a p-value that we can use to decide on whether to discard the null hypothesis that the models fit equally well. Here, the change in χ^2 is highly significant, so the more general FE model should be retained.

Conclusion

Panel regression in SEM has been outlined in well-known articles by P. Allison (2011); Bollen and Brand (2010); Teachman et al. (2001). This article provides a focused look at the implementation of the basic

model using the `lavaan` package in `R`. The online portion of the article further discuss common extensions and some tools for evaluating and loosening model assumptions.

The benefits of the SEM-based approach as opposed to traditional least squares-based FE-models are largely the same ones that apply to the SEM framework in general: for one, SEM allows for a great deal of flexibility. For example, it is easy to loosen model constraints as necessary. Measurement error in both the dependent and independent variables can be dealt with using latent variables to address attenuation bias in variables measured with error. Researchers interested in time-invariant predictors can integrate them into a hybrid FE/RE model with ease. Further extensions, like measurement invariance testing (Schoot, Lugtig, and Hox 2012; Millsap 2011; Steenkamp and Baumgartner 1998) as well as lagged dependent variables (Bollen and Brand 2010; P. Allison, Williams, and Moral-Benito 2017) for example, can also be implemented in a straightforward fashion.

There are drawbacks or at least important caveats associated with the SEM-based approach to panel regression. For one, the models can become very cumbersome to specify. Even fairly simple models can turn in into dozens of lines of code, depending on the number of observed waves of data. By introducing multiple indicator measurement models, the code can quickly expand to over 100 lines. The normality assumption is often a point of contention, and while there are numerous ways to estimate such models with nonnormal and categorical outcomes in SEM (discussed briefly in the online portion), there remains some uncertainty regarding which alternative to use in a given situation, and under what circumstances one can expect desirable results. Simulation studies can help guide one's choice, but there is always a chance that some neglected boundary conditions could thwart one's analysis.

The most basic static panel SEM regression model is the basis for a variety of currently popular extended models, such as Latent Curve Models in general (Curran and Bollen 2001; Bollen and Curran 2004), as well as special implementations like the Dynamic Panel Model (P. Allison, Williams, and Moral-Benito 2017), the Random-Intercept Cross-Lagged Panel Model (Hamaker, Kuiper, and Grasman 2015) and the Latent Curve Model with Structured Residuals (Curran et al. 2014). For this reason, it is all the more important for researchers to have a good grasp on the method of applying panel regression in SEM, and understanding the intuition of controlling for time-invariant confounders. This article is meant to serve as a consolidated resource for researchers looking for concrete advise on specifying RE, FE and more general panel models in SEM.

References

Allison, Paul. 2011. *Fixed Effects Regression Models*. Thousand Oaks: Sage Publications.

Allison, Paul D. 2009. *Fixed Effects Regression Models*. Thousand Oaks: Sage Publications. https://doi.org/10.4135/9781412993869.

Allison, Paul, Richard Williams, and Enrique Moral-Benito. 2017. "Maximum Likelihood for Cross-Lagged Panel Models with Fixed Effects." *Socius* 3: 1–17. https://doi.org/10.1177/2378023117710578.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, Oxford: Princeton University Press.

Bianconcini, Silvia, and Kenneth A. Bollen. 2018. "The Latent Variable-Autoregressive Latent Trajectory Model: A General Framework for Longitudinal Data Analysis." *Structural Equation Modeling: A Multidisciplinary Journal* 25 (5): 791–808.

Bollen, Kenneth. 1989. *Structural Equations with Latent Variables*. New York, Chichester: Wiley.

Bollen, Kenneth, and Jennie Brand. 2010. "A General Panel Model with Random and Fixed Effects: A Structural Equations Approach." *Social Forces* 89(1): 1–34.

Bollen, Kenneth, and Patrick Curran. 2004. "Autoregressive Latent Trajectory (ALT) Models: A Synthesis of Two Traditions." *Sociological Methods and Research* 32(3): 336–83. https://doi.org/10.1177/0049124103260222.

Brüderl, Josef, and Volker Ludwig. 2015. "Fixed-Effects Panel Regression." In *The Sage Handbook of Regression Analysis and Causal Inference*, edited by Henning Best and Christof Wolf, 327–57. London, Thousand Oaks: Sage Publications.

Curran, Patrick, and Daniel Bauer. 2011. "The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change." *Annual Review of Psychology* 62: 583–619.

Curran, Patrick, and Kenneth Bollen. 2001. "The Best of Both Worlds: Combining Autoregressive and Latent Curve Models." In *New Methods for the Analysis of Change*, edited by L. Collins and A. Sayer, 107–35. Washington, DC: American Psychological Press. https://doi.org/%20doi:10.1037/10409-004.

Curran, Patrick, Andrea Howard, Sierra Bainter, Stephanie Lane, and James McGinley. 2014. "The Separation of Between-Person and Within-Person Components of Individual Change over Time: A Latent Growth Curve Model with Structured Residuals." *Journal of Consulting and Clinical Psychology* 82(5): 879–94. https://doi.org/10.1037/a0035297.

Ferron, John M., and Melinda R. Hess. 2007. "Estimation in SEM: A Concrete Example." *Journal of Educational and Behavioral Statistics* 32 (1): 110–20.

Hamaker, Ellen, Rebecca Kuiper, and Raoul Grasman. 2015. "A Critique of the Cross-Lagged Panel Model." *Psychological Methods* 20(1): 102–16. https://doi.org/10.1037/a0038889.

Heckman, James J. 1981. "Heterogeneity and State Dependence." In *Studies in Labor Markets*, edited by Sherwin Rosen, 91–140. Chicago: University of Chicago Press.

Hox, Joop J. 2010. *Multilevel Analysis: Techniques and Applications. Second Edition.* New York, Hove: Routledge.

Hoyle, Rick H. 2015a. "Introduction and Overview." In *Handbook of Structural Equation Modeling*, edited by Rick H. Hoyle, 3–16. New York, London: The Guilford Press.

———. 2015b. "Model Specification in Structural Equation Modeling." In *Handbook of Structural Equation Modeling*, edited by Rick H. Hoyle, 126–44. New York, London: The Guilford Press.

Hsiao, Cheng. 2014. *Analysis of Panel Data. Third Edition.* New York: Cambridge University Press.

Millsap, Roger. 2011. *Statistical Approaches to Measurement Invariance.* New York, London: Routledge.

Muthén, Bengt O., Linda K. Muthén, and Tihomir Asparouhov. 2016. *Regression and Mediation Analysis Using Mplus.* Los Angeles, CA: Muthén & Muthén.

Muthén, Linda, and Bengt Muthén. 1998--2017. *Mplus User's Guide. Eighth Edition.* Los Angeles, California: Muthén & Muthén.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosseel, Yves. 2012. "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36. http://www.jstatsoft.org/v48/i02/.

Rosseel, Yves, Terrence D. Jorgensen, Daniel Oberski, Jarrett Byrnes, Leonard Vanbrabant, Victoria Savalei, Ed Merkle, et al. 2020. *lavaan: Latent Variable Analysis.* https://CRAN.R-project.org/package=lavaan.

Schmidheiny, Kurt. 2019. "Panel Data: Fixed and Random Effects." Unversität Basel.

Schoot, Rens van de, Peter Lugtig, and Joop Hox. 2012. "A Checklist for Testing Measurement Invariance." *European Journal of Developmental Psychology* 9(4): 486–92.

Seifert, Nico, and Henrik Andersen. 2020. "The Impact of Social Isolation on Subjective Well-Being." December 16, 2020. https://www.researchgate.net/publication/348548953_The_impact_of_social_isolation_on_subjective_well-being.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling.* Boca Raton, London, New York, Washington, D.C.: Chapman & Hall/CRC.

Steenkamp, Jan-Benedict, and Hans Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25(1): 78–90.

Teachman, Jay, Greg Duncan, Jean Yeung, and Dan Levy. 2001. "Covariance Structure Models for Fixed and Random Effects." *Sociological Methods and Research* 30(2): 271–88.

Wooldridge, Jeffery. 2002. *Econometric Analysis of Cross Sectional and Panel Data.* Cambridge, Massachusetts: The MIT Press.

———. 2009. *Introductory Econometrics: A Modern Approach, 4^{th} Edition.* Mason, Ohio: South-Western Cengage Learning.

———. 2012. *Introductory Econometrics: A Modern Approach, 5^{th} Edition.* Mason, Ohio: Thomson South-Western.

Zyphur, Michael J., Paul D. Allison, Louis Tay, Manuel C. Voelkle, Kristopher J. Preacher, Zhen Zhang, Ellen L. Hamaker, et al. 2020a. "From Data to Causes i: Building a General Cross-Lagged Panel Model (GCLM)." *Organizational Research Methods* 23 (4): 651–87. https://doi.org/10.1177/1094428119847278.

———, et al. 2020b. "From Data to Causes II: Comparing Approaches to Panel Data Analysis." *Organizational Research Methods* 23 (4): 688–716. https://doi.org/10.1177/1094428119847280.

Figure 1: Four-wave random effects model

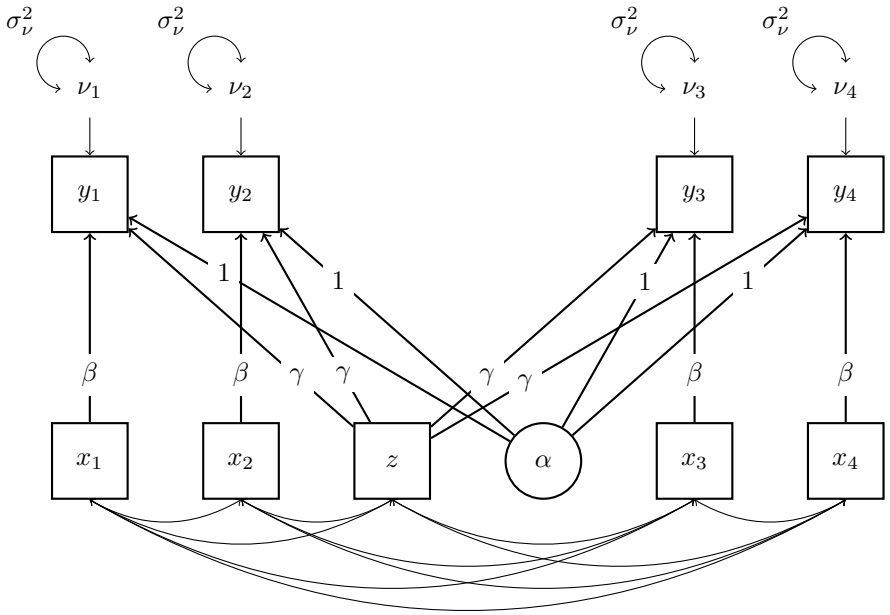

Figure 2: Four-wave fixed effects model