Supplementary materials for: 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

Please cite the base text:

Andersen, H. K. (2021). A closer look at random and fixed effects panel regression in structural equation modeling using `lavaan`. *Structural Equation Modeling: A Multidisciplinary Journal* (forthcoming). https://doi.org/10.1080/10705511.2021.1963255

The following discusses further topics related to (static) panel regression models in SEM and explores some extensions to the basic models. The first section discusses the logic of RE and FE regression in SEM in greater detail than the base text. The next section discusses the issues of nonnormal variables and possibilities for treating missing data. Following that, it is illustrated that under ideal conditions, the SEM-based panel regression models produce the same results as other methods, as implemented in the `plm` (Croissant and Millo 2008) and `lme4` (Bates et al. 2015) packages for panel and multilevel models in `R`. Finally, the final section illustrates ways to extend the basic models: relaxing assumptions and accounting for measurement error with latent variables.

The logic of RE and FE regression in SEM

Let us begin with the RE model in SEM. The way it typically works is to convert the long-format data (NT observations stacked on top of each other) into wide-format (T individual columns of length N). We write T equations:

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

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).

For x_{it} and z_i, we simply have

$$\begin{aligned} x_{i1} &= \mu_{x_1} + \delta_{i1} \\ x_{i2} &= \mu_{x_2} + \delta_{i2} \\ &\vdots \\ x_{iT} &= \mu_{x_T} + \delta_{iT} \\ z_i &= \mu_z + \theta_i \end{aligned} \tag{2}$$

where $\mu_{x_t} = \mathbb{E}(x_{it})$, $\mu_z = \mathbb{E}(z_i)$, i.e., they are expressed by an unconditional mean and an individual deviation (δ_{it} and θ_i) from the mean (Bollen and Curran 2004).

The goal of SEM is to find estimates for the unknown parameters such that the differences between the observed mean vector, $\bar{\boldsymbol{o}}$, and covariance matrix, $\boldsymbol{S}$, and *model-implied* mean vector, $\boldsymbol{\mu}(\boldsymbol{\theta})$, and covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, are minimized (Bollen 1989, 1). The vector $\boldsymbol{\theta}$ holds the unknown model parameters, including coefficients, error variances and the variances of the latent variables.

Before discussing the SEM further, let us look at a concrete example of a four-wave RE model. This is shown in Figure 1 in the base text. The sample mean vector and covariance matrix for this model simply consist of the sample means and (co)variances between the observed variables. If we call $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^\intercal$ and $\boldsymbol{w}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, z_i)^\intercal$, then the observed mean vector and covariance matrix would be

$$\bar{\boldsymbol{o}} = \begin{bmatrix} \bar{\boldsymbol{y}} \\ \bar{\boldsymbol{w}} \end{bmatrix}, \boldsymbol{S} = \begin{bmatrix} \mathrm{var}(\boldsymbol{y}) & \mathrm{cov}(\boldsymbol{y}, \boldsymbol{w}) \\ \mathrm{cov}(\boldsymbol{w}, \boldsymbol{y}) & \mathrm{var}(\boldsymbol{w}) \end{bmatrix}$$

where $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{w}}$ are the observed means per timepoint, and in this case $\boldsymbol{S}$ is a 9×9 (four time-varying dependent variables, $\boldsymbol{y}_i$, four time-varying independent variables, $\boldsymbol{x}_i$, one time-invariant independent variable, z_i) and $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot)$ stand for the observed (co)variances.

For the model-implied mean vector and covariance matrix, we collect the observed variables into a single vector, $\boldsymbol{y}_i^* = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, x_{i1}, x_{i2}, x_{i3}, x_{i4}, z_i)^\intercal$, the observed together with the latent variable into another, $\boldsymbol{\eta}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, x_{i1}, x_{i2}, x_{i3}, x_{i4}, z_i, \alpha_i)^\intercal$, and write the model in matrix notation as

$$\begin{aligned} \boldsymbol{\eta}_i &= \boldsymbol{\kappa} + \boldsymbol{B}\boldsymbol{\eta}_i + \boldsymbol{\nu}_i \\ \boldsymbol{y}_i^* &= \boldsymbol{\Lambda}\boldsymbol{\eta}_i \end{aligned}$$

which we can write in reduced form as

$$\begin{aligned} \boldsymbol{\eta}_i &= (\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\kappa} + \boldsymbol{\nu}_i) \\ \boldsymbol{y}_i^* &= \boldsymbol{\Lambda}[(\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\kappa} + \boldsymbol{\nu}_i)] \end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{\kappa} = (\kappa_{y_1}, \kappa_{y_2}, \kappa_{y_3}, \kappa_{y_4}, \mu_{x_1}, \mu_{x_2}, \mu_{x_3}, \mu_{x_4}, \mu_z, 0)^\intercal$ is a vector of intercepts and unconditional means (with μ_α, the mean of the latent individual effects, fixed to zero for identification purposes) and $\boldsymbol{\nu}_i = (\nu_{i1}, \nu_{i2}, \nu_{i3}, \nu_{i4}, \delta_{i1}, \delta_{i2}, \delta_{i3}, \delta_{i4}, \theta_i, \alpha_i)^\intercal$ is a vector of errors or rather individual deviations from the unconditional means.[1] $\boldsymbol{\Lambda}$ is a matrix of zeros and ones that just selects the observed variables so that the resulting model-implied mean vector and covariance matrix have the same dimensions as the observed counterparts and $\boldsymbol{B}$ is a matrix of coefficients:

[1] Note that $\mu_\alpha = 0$, so we do not decompose α_i into an unconditional mean and individual deviation.

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & \beta & 0 & 0 & 0 & \gamma & 1 \\ 0 & 0 & 0 & 0 & 0 & \beta & 0 & 0 & \gamma & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & \gamma & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & \gamma & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

This notation (suggested by Graff (1979) and discussed in Bollen (1989), p. 396) may not be the most intuitive (other notations are suggested, for example, in Bollen and Brand (2010)) but it is the way the model is represented in `lavaan`.[2] If it seems confusing, keep in mind that this matrix notation is just another way of expressing the equations in (1) and (2).

From here, the model-implied mean vector is

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\theta}) &= \boldsymbol{\Lambda}\,\mathbb{E}(\boldsymbol{\eta}_i) \\ &= \boldsymbol{\Lambda}(\boldsymbol{\kappa} + \boldsymbol{B}\boldsymbol{\mu}_\eta) \end{aligned}$$

where $\boldsymbol{\mu}_\eta$ holds the expected values of the variables in $\boldsymbol{\eta}_i$. This results since $\mathbb{E}(\boldsymbol{\nu}_i) = \mathbf{0}$, by assumption, which is unproblematic with the inclusion of $\boldsymbol{\kappa}$. The model-implied covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, is

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \mathbb{E}(\boldsymbol{y}_i^* \boldsymbol{y}_i^{*\intercal}) - \mathbb{E}(\boldsymbol{y}_i^*)\,\mathbb{E}(\boldsymbol{y}_i^*)^\intercal \\ &= \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{B})^{-1\intercal}\boldsymbol{\Lambda}^\intercal \end{aligned}$$

where $\boldsymbol{\Psi} = \mathbb{E}(\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\intercal)$ and which results if we retain the assumption $\mathbb{E}(\boldsymbol{\nu}_i) = \mathbf{0}$.

This is where the assumptions discussed regarding the conventional models in the base text come into play. The matrix $\boldsymbol{\Psi}$ is the covariance matrix of the errors, or rather the individual deviations from the unconditional means in the case of the exogenous variables. It looks like:

$$\boldsymbol{\Psi} = \mathbb{E}\begin{bmatrix} \nu_{i1}^2 & & & & & & & & & \\ \nu_{i2}\nu_{i1} & \nu_{i2}^2 & & & & & & & & \\ \nu_{i3}\nu_{i1} & \nu_{i3}\nu_{i2} & \nu_{i3}^2 & & & & & & & \\ \nu_{i4}\nu_{i1} & \nu_{i4}\nu_{i2} & \nu_{i4}\nu_{i3} & \nu_{i4}^2 & & & & & & \\ \delta_{i1}\nu_{i1} & \delta_{i1}\nu_{i2} & \delta_{i1}\nu_{i3} & \delta_{i1}\nu_{i4} & \delta_{i1}^2 & & & & & \\ \delta_{i2}\nu_{i1} & \delta_{i2}\nu_{i2} & \delta_{i2}\nu_{i3} & \delta_{i2}\nu_{i4} & \delta_{i2}\delta_{i1} & \delta_{i2}^2 & & & & \\ \delta_{i3}\nu_{i1} & \delta_{i3}\nu_{i2} & \delta_{i3}\nu_{i3} & \delta_{i3}\nu_{i4} & \delta_{i3}\delta_{i1} & \delta_{i3}\delta_{i2} & \delta_{i3}^2 & & & \\ \delta_{i4}\nu_{i1} & \delta_{i4}\nu_{i2} & \delta_{i4}\nu_{i3} & \delta_{i4}\nu_{i4} & \delta_{i4}\delta_{i1} & \delta_{i4}\delta_{i2} & \delta_{i4}\delta_{i3} & \delta_{i4}^2 & & \\ \theta_i\nu_{i1} & \theta_i\nu_{i2} & \theta_i\nu_{i3} & \theta_i\nu_{i4} & \theta_i\delta_{i1} & \theta_i\delta_{i2} & \theta_i\delta_{i3} & \theta_i\delta_{i4} & \theta_i^2 & \\ \alpha_i\nu_{i1} & \alpha_i\nu_{i2} & \alpha_i\nu_{i3} & \alpha_i\nu_{i4} & \alpha_i\delta_{i1} & \alpha_i\delta_{i2} & \alpha_i\delta_{i3} & \alpha_i\delta_{i4} & \alpha_i\theta_i & \alpha_i^2. \end{bmatrix}$$

Some of the assumptions, like linearity and a lack of excessive multicollinearity, are straightforward and will be taken as a given. We will also assume homoskedasticity of the composite errors conditional on the model covariates. The errors will be assumed normally distributed for the sake of simplicity. The online section discusses some ways to address nonormal variables in SEM.

[2] Those interested can examine the matrix representation of the model using `lavInspect(mx.fit, what = "coef")`, where `mx.fit` is the fitted `lavaan` model.

Let us turn to the remaining assumptions and discuss them in some detail. First, let us look at the assumption of strict exogeneity (i.e., zero conditional mean of the composite error), which can be stated in this case as $E(\nu_{it}|\boldsymbol{x}_i, z_i, \alpha_i) = 0$ and $\mathbb{E}(\alpha_i|\boldsymbol{x}_i, z_i) = 0$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\intercal$. In practical terms, strict exogeneity means that the idiosyncratic error, ν_{it}, at each point in time is uncorrelated with the model covariates (including the individual effects) at all points in time. So too are the individual effects assumed uncorrelated with the model covariates at all points in time in the RE model. Since $x_{it} = \mu_{x_t} + \delta_{it}$ and $z_i = \mu_z + \theta_i$, we set $\mathbb{E}(\delta_{it}\nu_{is})$, $\mathbb{E}(\theta_i\nu_{it})$, $\mathbb{E}(\alpha_i\nu_{it})$ as well as $\mathbb{E}(\delta_{it}\alpha_i)$ and $\mathbb{E}(\theta_i\alpha_i)$ to zero for all t and s.[3] Second, we saw in the covariance matrix of the composite errors that the source of dependency between observations on the same unit over time was the individual effect α_i. After accounting for this, we do not expect any covariance between the idiosyncratic errors of the same unit over time. So, we can set $\mathbb{E}(\nu_{it}\nu_{is})$ to zero whenever $t \neq s$. Third, we mentioned that homoskedasticity in the panel regression setup can apply to the time dimension as well. If we believe the idiosyncratic error variance is constant over time, we can constrain the variance of ν_{it} so that a single variance is estimated for all four timepoints. We can reflect this assumption by removing the t subscript from $\mathbb{E}(\nu_{it}^2)$ and just write $\mathbb{E}(\nu_i^2)$. Other than that, the models discussed so far place no assumptions on the relationships between the exogenous observed covariates, so we can allow these to covary freely. This results in the following specification of the $\boldsymbol{\Psi}$ matrix:

$$\boldsymbol{\Psi} = \mathbb{E}\begin{bmatrix}
\nu_i^2 & & & & & & & & & \\
0 & \nu_i^2 & & & & & & & & \\
0 & 0 & \nu_i^2 & & & & & & & \\
0 & 0 & 0 & \nu_i^2 & & & & & & \\
0 & 0 & 0 & 0 & \delta_{i1}^2 & & & & & \\
0 & 0 & 0 & 0 & \delta_{i2}\delta_{i1} & \delta_{i2}^2 & & & & \\
0 & 0 & 0 & 0 & \delta_{i3}\delta_{i1} & \delta_{i3}\delta_{i2} & \delta_{i3}^2 & & & \\
0 & 0 & 0 & 0 & \delta_{i4}\delta_{i1} & \delta_{i4}\delta_{i2} & \delta_{i4}\delta_{i3} & \delta_{i4}^2 & & \\
0 & 0 & 0 & 0 & \theta_i\delta_{i1} & \theta_i\delta_{i2} & \theta_i\delta_{i3} & \theta_i\delta_{i4} & \theta_i^2 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_i^2
\end{bmatrix}.$$

Once we have specified the model and applied the assumptions by placing constraints on the $\boldsymbol{B}$ and $\boldsymbol{\Psi}$ matrices, we can estimate it. The most widely used estimator is maximum likelihood, which involves minimizing the fitting function

$$F_{ML} = \ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \ln|\boldsymbol{S}| + \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{S}\right) - p + (\bar{\boldsymbol{o}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})(\bar{\boldsymbol{o}} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

where p is the number of observed variables, ln is the natural log, $|\cdot|$ is the determinant and tr is the trace (Bollen and Brand 2010, 32). The estimates for $\boldsymbol{\theta}$ are chosen as those that minimize the fitting function, thereby also minimizing the differences between the observed and model-implied mean vectors and covariance matrices.

Recall that in SEM we model the individual effects explicitly as a latent variable. Equation (1) from the RE setup pertains for the FE model as well. We relax the assumption of unrelated effects by allowing the individual effects, α_i, to correlate with the time-varying model covariates. We do so by relaxing the assumptions in $\boldsymbol{\Psi}$ and allowing $\mathbb{E}(\alpha_i x_{it})$ to be freely estimated at each point in time, rather than fixing them to zero. The assumption of $\mathbb{E}(\alpha_i z_i) = 0$ must be retained for identification.

Some notes on panel regression in SEM

Especially when one is switching from other statistical methods to SEM, some points of contention are bound to come up. In this section, we can briefly discuss and address two popular ones: nonnormality and missing data.

[3] In fact, when it comes to the time-invariant covariate, we *must* fix the covariance between the individual effects and z_i to zero, or else a perfect collinearity between the two will result, and the model will not be identified (Bollen and Brand 2010, 27).

Nonnormal variables

Standard estimation of SEMs using the most common techniques such as maximum likelihood (ML) and generalized least squares (GLS) typically requires the assumption of multivariate normality along with the other common regression assumptions discussed in the base text. As the name suggests, ML estimation entails maximizing the likelihood of the observed variables given the model parameters. Getting the likelihood requires a knowing (or assuming) a well-defined probability density function. GLS, on the other hand, works by choosing a weight matrix (typically the inverse of the observed covariance matrix) to minimize the (weighted) squared differences between observed and model-implied matrices. In both cases, ML and GLS, the assumption of multivariate normality means the probability distributions of the observed variables can be completely summarized by those sample (co)variances (along with a vector of means, if the mean structure is to be examined as well). If the observed variables diverge from the normality assumption, then higher-order moments (skewness, kurtosis) would be needed to summarize the distributions (West, Finch, and Curran 1995).

It should be noted, however, that the normality assumption concerns the *error distribution* and thus usually only the endogenous, i.e., dependent variables (R. B. Kline 2015; CenterStat 2019) along with any latent exogenous variables (then the normality assumption concerns the observed indicators).

There are two main sources of nonnormality: nonnormal continuous variables and so-called coarsely categorized or simply ordered categorical variables (West, Finch, and Curran 1995). Nonnormal continuous variables are continuous variables characterized by skewness and/or kurtosis that deviate from that of the standard normal distribution. Coarsely categorized variables are those with underlying distributions that could be considered normal, but where observed measures are available as categorized data. Still other dependent variables may be dichotomous in nature, e.g., employed/unemployed or voted for party A/did not vote for party A. Even dichotomous variables, however, can sometimes be thought of as stemming from an unobserved continuous normal distribution (see Best and Wolf 2015, 155).

Nonnormal continuous variables are arguably the less problematic of the two. They can lead to a χ^2 model fit test statistic that is too often significant when it should not be (incorrect rejection of the true model) and standard errors that tend to be too low, yielding too many significant parameter estimates (West, Finch, and Curran 1995). In `lavaan`, estimators like `MLM` (for complete data) and `MLR` (for complete and incomplete data) offer standard errors and χ^2 test statistics that are robust to nonnormality, see the `lavaan` tutorial website (https://lavaan.ugent.be/tutorial/est.html) and the 'brief user's guide' (https://users.ugent.be/~yrosseel/lavaan/lavaan2.pdf). These estimators can be chosen by including the argument `estimator =` in the `sem()` function, e.g., for `MLR`:

```r
mx.fit <- sem(mx, data = df, estimator = "MLR")
```

Ordered categorical variables display lower correlations between variables than continuous counterparts (West, Finch, and Curran 1995). This leads to a type of attenuation bias where the estimated effects in ML-based models with categorical outcomes are lower than they should be. The amount of bias depends on how coarse the variable was categorized, so to speak. Generally, the fewer the categories, the weaker the observed correlations and thus the more downwards biased the parameter estimates will be (downwards in the sense of towards zero). The same that goes for nonnormal continuous variables in terms of χ^2 tests of model fit and standard errors apply.

The solution to the problem of categorical (or even dichotomous) outcomes in SEM is to simply declare them as such. This depends on the software, but in `lavaan` it can be done using the shortcut `ordered = c("y1", "y2")` in the `sem()` function call, assuming there are two dependent variables are called `y1` and `y2` and they are ordered categorical. In the model syntax, the user can includes $M_k - 1$ thresholds (where M_k is the number of categories for the kth variable) per categorical endogenous variable (Rosseel 2021). For example, if each dependent variable were measured on a Likert style scale with five categories, four thresholds could be estimated (rather than means) using

```
y1 | t1 + t2 + t3 + t4
y2 | t1 + t2 + t3 + t4
```

where `t1`, `t2` etc. are used by convention to specify the thresholds (Rosseel et al. 2020). The `ordered` option turns on the diagonal weighted least squares (DWLS) estimator along with corrections for standard errors and test statistics. Many know this as the *weighted least squares means and variance adjusted* (WLSMV) 'estimator' in `Mplus`.[4] Those who are familiar with the latent variable derivation of the logistic and probit regression model (see Best and Wolf 2015) will be familiar with the logic of these estimators. Essentially, it is assumed that a normally distributed underlying variable is responsible for the crude measures on the categorical scale. Thresholds are calculated based on the proportion of responses per category, which are used to estimate the continuous latent variable distribution. In a **lavaan** model using the DWLS estimator, the estimated coefficients are those of a probit regression model, but can be interpreted as the linear effect of the covariate on the latent continuous underlying response variable, see also this post on the **lavaan** Google Groups forum: https://groups.google.com/g/lavaan/c/mG5Mjrf2jgo.

Missing data

In most practical applications, missing data is an issue (Allison 2003). Seldom do we have valid data on all model variables for all cases, especially when it comes to survey data where confusingly worded questions or social desirability concerns, for example, can cause respondents to skip questions.

In the context of panel studies, the issue of panel attrition also plays a role. Even if complete data is observed in one wave of data collection, it is not always the case that the same observational unit is available to take part in subsequent waves.

In SEM, like other methods, missing data treatments like listwise deletion, where units with missing values on any of the model variables are eliminated, are always available. However, notice that when we transform the panel dataset from long- to wide-format (as outlined in the base text), the impact of listwise deletion is magnified. That is, for each unit in a long-format dataset, the timepoints are separate rows of data. If complete data for unit i is available at time t but data is incomplete for the same unit i at time s, then at least one row of data is retained with listwise deletion. Once the data is in wide-format, there is only a single row per unit, with each timepoint in a separate column. In this setup, if data for unit i is complete at time t but incomplete at time s, then the observations for the unit *at all points in time* are lost through listwise deletion. Therefore listwise deletion in panel SEM is even less attractive than it is for conventional methods using stacked, long-format data.

Methods like listwise and pairwise (where means and co(variances) are estimated using the available pairwise data) are appropriate in some situations, ideally when the missing data is missing completely at random (MCAR, i.e., the probability of a missing value does not depend on other variables, observed or unobserved). However, there are a number of drawbacks to these methods, especially when the data is not MCAR but rather missing at random (MAR), where missingness is dependent on the unobserved missing data itself. For example, if the probability of a missing value on income is dependent on the respondent's income then the missing data is said to be MAR. For an overview of the situations in which listwise and pairwise deletion may be appropriate, and their drawbacks, see (Allison 2003; R. Kline 2016).

Other methods like mean, single and multiple imputation have been introduced as an alternative to listwise and pairwise deletion, in which a value for the missing datum is estimated or 'imputed.' For example, for single or 'conditional' imputation, if data was missing on income, we could estimate a unit's income given their sex, age, job, etc., assuming those characteristics were available (Allison 2003; Graham and Coffman 2015).

[4] Estimator is placed in quotations because it is somewhat of a misnomer. Turning on WLSMV in `Mplus` employs the DWLS estimator plus robust standard error and test statistics, see this post on the **lavaan** Google Groups forum: https://groups.google.com/g/lavaan/c/Nymu7jmVUk8.

All of these methods above are available in current SEM software (see for example Allison 2003; Muthén, Muthén, and Asparouhov 2016). However, arguably the preferred method for dealing with missing data in SEM is the Full Information Maximum Likelihood (FIML) method, sometimes referred to as 'direct ML' (Allison 2003; Lei and Wu 2015). For each individual in the sample, an individual-specific vector of observed (i.e., minus the variables with missing values) is used: $\boldsymbol{x}_i$. FIML entails maximizing the likelihood of these individual-specific vectors $\boldsymbol{x}_i$ of values given an individual-specific mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} f(\boldsymbol{x}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

where $f(\cdot)$ is the multivariate normal density function (Allison 2003). To be more concrete, if $\boldsymbol{x} = (x_1, x_2, x_3, x_4)^\intercal$ was the vector of four model variables, and x_2 and x_3 were missing for unit i, then $\boldsymbol{x}_i = (x_{1i}, x_{4i})^\intercal$, $\boldsymbol{\mu}_i = (\mu_{x1}, \mu_{x4})^\intercal$ and $\boldsymbol{\Sigma}_i = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_4) \\ \mathrm{Cov}(x_4, x_1) & \mathrm{Var}(x_4) \end{bmatrix}$ where the (co)variances in $\boldsymbol{\Sigma}_i$ are the model-implied (co)variances based on the model parameters.

Loosely speaking, FIML requires that missingness is at least ignorable, i.e., MAR and the mechanism that governs the missing data should be distinct from the variables included in the model (Allison 2003). FIML is straightforward in situations of continuous normally distributed variables but can be implemented for categorical data as well, see for example this discussion on the `Mplus` forum: http://www.statmodel.com/discussion/messages/23/11336.html?1602351808.

A comparison with non-SEM methods

To summarize the previous section, there are numerous solutions available to a variety of issues and limitations to SEM panel regression. For nonnormal data, robust estimators and models akin to logistic and probit regressions are available and straightforward to implement in modern SEM software. Missing values can be dealt with in a number of ways and the FIML approach (though not without limitations, see Graham and Coffman (2015)) is currently seen by many as a state-of-the-art way to deal with missingness (Lei and Wu 2015).

Under ideal conditions, panel SEM results line up with the more traditional (i.e., non-SEM) methods. As a demonstration, take the typical fixed effects model. We can use the long-format data (file: `longData.Rda`) to run the typical FE model using the `plm` package (Croissant and Millo 2008) for panel regression. By default, the `plm` function assumes the dataframe is structured so that the first two columns correspond to the individual and time indices, see the documentation at https://cran.r-project.org/web/packages/plm/plm.pdf or Croissant and Millo (2008).

```r
library(plm)

# Run the FE model in plm
fe1 <- plm(y ~ x,
           effect = "individual", model = "within",
           data = df)
summary(fe1)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = y ~ x, data = df, effect = "individual", model = "within")
##
## Balanced Panel: n = 1000, T = 5, N = 5000
```

```
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -3.62051 -0.61151  0.02113  0.62445  3.10621
##
## Coefficients:
##    Estimate Std. Error t-value  Pr(>|t|)
## x 0.3001194  0.0081337  36.898 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5480.6
## Residual Sum of Squares: 4088.6
## R-Squared:      0.25398
## Adj. R-Squared: 0.067434
## F-statistic: 1361.48 on 1 and 3999 DF, p-value: < 2.22e-16
```

The FE-SEM in the base text gave an estimated coefficient of $\hat{\beta}_{FE-SEM} = 0.298$ (0.01). The results in the output above show that they are essentially identical with $\hat{\beta}_{FE} = 0.3$ (0.008). Note it is not possible to include time-invariant covariates in the `plm` model because they get wiped out along with the individual effects in the demeaning.

Other methods of estimating FE models work in either the random or mixed effects model framework. For example, with dichotomous covariates, it may seem strange to demean them along with the rest of the equation, so instead we could include the cluster means per individual of the time-varying independent variables, here x, in the equation to achieve within estimates (Mundlak 1978; Chamberlain 1980; Wooldridge 2002).

```r
# Generate the cluster means for x per id
clusterMeanx <- aggregate(df$x, by = list(df$id), FUN = mean)
# Rename the columns
names(clusterMeanx) <- c("id", "xbar")

# Add the cluster means back into df
df <- merge(df, clusterMeanx, by = "id")
```

Here using the `plm` function in the `random` setup:

```r
fe2 <- plm(y ~ x + xbar,
           effect = "individual", model = "random",
           data = df)
summary(fe2)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = y ~ x + xbar, data = df, effect = "individual",
##     model = "random")
##
## Balanced Panel: n = 1000, T = 5, N = 5000
##
## Effects:
```

```
##                   var std.dev share
## idiosyncratic 1.0224  1.0111 0.638
## individual    0.5810  0.7622 0.362
## theta: 0.4898
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -3.709784 -0.691096  0.021666  0.679000  3.673462
##
## Coefficients:
##              Estimate Std. Error z-value  Pr(>|z|)
## (Intercept) 1.3122101  0.0788245  16.647 < 2.2e-16 ***
## x           0.3001194  0.0081337  36.898 < 2.2e-16 ***
## xbar        0.7096142  0.0158843  44.674 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12101
## Residual Sum of Squares: 5109
## R-Squared:      0.57779
## Adj. R-Squared: 0.57762
## Chisq: 6838.42 on 2 DF, p-value: < 2.22e-16
```

And here using the `lmer` function of the `lme4` package (Bates et al. 2015) to estimate a mixed model:

```r
library(lme4)

# Run the mixed model in lmer with the cluster means for x
mixed1 <- lmer(y ~ x + xbar + (1 | id), data = df)
summary(mixed1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ x + xbar + (1 | id)
##    Data: df
##
## REML criterion at convergence: 15662.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3271 -0.6204  0.0163  0.6218  3.1667
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 0.581    0.7622
##  Residual             1.022    1.0111
## Number of obs: 5000, groups:  id, 1000
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 1.312210   0.078825   16.65
## x           0.300119   0.008134   36.90
## xbar        0.709614   0.015884   44.67
##
```

```
## Correlation of Fixed Effects:
##      (Intr) x
## x     0.000
## xbar -0.803 -0.512
```

In both cases, the models return the same estimates as the FE and FE-SEM models. Also, in both the `random` setup using the `plm` function, and the mixed model using the `lmer` function, we get estimates of the variance components, $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\nu$:

```
# Print the variance components for the plm model
print(ercomp(fe2),
      digits = 3)
```

```
##                 var std.dev share
## idiosyncratic 1.022   1.011  0.64
## individual    0.581   0.762  0.36
## theta: 0.49
```

```
# Print the variance components for the lmer model
print(VarCorr(mixed1),
      comp = c("Variance", "Std.Dev"),
      digits = 3)
```

```
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 0.581    0.762
##  Residual             1.022    1.011
```

From this we see both models report the same estimated variance components, $\hat{\sigma}^2_\alpha = 0.581$ and $\hat{\sigma}^2_\nu = 1.022$, telling us that about 36.2% of the residual variance is due to the differences between individuals (shown in the `share` column of the `ercomp()` output). This is what is referred to as the intraclass correlation coefficient, or ICC (Hox 2010).

Extensions

Relaxing assumptions meant to mimic traditional FE models

There are a number of implicit assumptions attached to the typical FE model that can be relaxed in SEM. Some of these assumptions have been discussed already, and a fairly comprehenisve list of assumptions can be found in Bollen and Brand (2010). Here, I will go over just a few, concentrating on the implementation in `lavaan` and the opportunity to empirically test whether the adjustments are justified or not.

The assumptions we will discuss here pertain to the time-invariance of the effects of both the latent individual effects and the observed covariates, as well as a time-invariant error variance.

For example, we can rewrite the original FE equation as

$$y_{it} = \beta_t x_{it} + \gamma_t z_i + \lambda_t \alpha_i + \nu_{it}$$

where β becomes β_t, γ becomes γ_t and the implicit regression weight of one turns to λ_t to highlight the fact that the effect of x, z as well as α on y may vary over time. We can furthermore easily relax the assumption of time-constant error variance, i.e., $\sigma^2_{\nu_t}$. As noted in the base text, the assumption regarding the relatedness of the model covariate(s) and the individual effects determines whether we have an FE or RE model. We can set the correlations to zero and test whether the RE model would be preferable to the FE model. In general, if the individual effects are truly uncorrelated with the model covariates, it is advisable to switch to an RE model since because it uses up less degrees of freedom, it will have smaller standard errors (Bollen and Brand 2010).

In the following `lavaan` code, we simply remove the factor loadings of one for the latent individual effect variable which allows them to be estimated freely at each timepoint. The first factor loading, however, is set to one by default for identification purposes. For the effect of the covariates, we can either delete the constraints `beta` in `yt ~ beta*xt` or give each regression a different label, e.g., `beta1`, `beta2`, `beta3`, etc. Similarly, to allow the error variance to vary over time, we turn the constraints `nu` into simple labels, i.e., `nu1`, `nu2`, `nu3`, etc., or again just delete them. In fact, regarding the error variances, they will be estimated necessarily, and do not need to be explicitly mentioned in the model syntax at all. Finally, to move from an FE to an RE model, we could simply constrain the correlations between the individual effects and the covariates to zero, i.e., `alpha ~~ 0*x1 + 0*x2 + 0*x3 + 0*x4 + 0*z`. The following model `fe_sem_fullyrelaxed` demonstrates a model in which all of the spoken of assumptions have been relaxed.

```
fe_sem_fullyrelaxed <- '
# Define individual effects variable
alpha =~ y1 + y2 + y3 + y4
# Regressions
y1 ~ beta1*x1 + gamma1*z
y2 ~ beta2*x2 + gamma2*z
y3 ~ beta3*x3 + gamma3*z
y4 ~ beta4*x4 + gamma4*z
# Allow unrestricted correlation between eta and covariates
alpha ~~ x1 + x2 + x3 + x4 + 0*z
# Alternatively: constrain all to 0 for RE model, or
# just individual correlations
# 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 ~~ nu1*y1
y2 ~~ nu2*y2
y3 ~~ nu3*y3
y4 ~~ nu4*y4
'
fe_sem_fullyrelaxed.fit <- sem(model = fe_sem_fullyrelaxed,
                               meanstructure = TRUE,
                               data = dfw,
                               estimator = "ML")
```

As outlined in Bollen and Brand (2010), the researcher has the opportunity to test each of the assumptions empirically and decide whether a more parsimonious, i.e., restrictive model is justifiable. In this case, for each assumption, a likelihood ratio test can be carried out to determine whether the improvement to model fit resulting from the relaxation of various assumptions is significant or whether the more parsimonious model is preferable after all.

If we use the original model `fe_sem.fit` (from the base text) as a starting point, the best strategy for testing these assumptions is to work in a stepwise fashion, relaxing one assumption at a time. We can begin by first constraining the correlation between α and x_t to zero (`re_sem`) for an RE model. If turning from an FE to an RE model does not significantly worsen model fit, we can go forward with the rest of the steps with the RE model. If, however, the fit does worsen significantly, it is likely better to stick with the FE model; moving forward then with it to see if a less restrictive FE model is preferable. We can perform a likelihood ratio test in R using the anova() function:

```
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_t 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 for the probability that the difference is solely due to chance. Here, the change in χ^2 is highly significant, so the FE model should be retained.

After now having established, based on the likelihood ratio test, that FE is our preferred model, we can begin relaxing the rest of the assumptions. I show the following merely as a demonstration of the procedure, we know already from the DGP that the parsimonious model as specified in fe_sem.fit is appropriate. We can next allow the error variances (fe_semb.fit), the effect of x on y (fe_semc.fit) and finally the factor loadings of the individual effects (fe_semd.fit) all to vary over time.

```
anova(fe_sem.fit, fe_semb.fit, fe_semc.fit, fe_semd.fit)
```

```
## Chi-Squared Difference Test
##
##             Df   AIC   BIC  Chisq Chisq diff Df diff Pr(>Chisq)
## fe_semd.fit 10 34911 35127 10.176
## fe_semc.fit 13 34908 35109 12.505     2.3294       3     0.5069
## fe_semb.fit 19 34900 35071 16.435     3.9303       6     0.6861
## fe_sem.fit  22 34895 35052 18.209     1.7742       3     0.6206
```

Keep in mind that a less parsimonious nested model (fewer degrees of freedom) will tend to fit the data better. I.e., even if, say, a population correlation was truly zero, chance variations due to sampling error mean that the observed correlation will tend not to equal zero exactly, so adding a constraint to a model will tend to worsen fit, at least minimally. The question here is whether the improvement to fit by loosening constraints is meaningful or not. In the table above, we should not expect any meaningful improvements moving from fe_sem.fit to fe_semd.fit. Here, using simulated data, we have the luxury of knowing that any significant differences in χ^2 are due to chance. With real data, it is up to the researcher to apply their best judgment and decide whether the results are plausible or not.

Measurement error

What if the observed variables are not measured perfectly? Then what we observe, call them $\tilde{x}_t$ and $\tilde{y}_t$ are composites of the true score we are after, i.e., x_t and y_t, plus an additive measurement error portion:

$$\begin{aligned}\tilde{x}_t &= x_t + \upsilon_t,\\ \tilde{y}_t &= y_t + \nu_t.\end{aligned}$$

How does this affect our model? Well, first notice that measurement error in the dependent variable is typically less of a serious problem than measurement error in the independent variables. Let us assume mean-centered variables so that we can ignore the intercept, and consider the following simple bivariate equation:

$$y = \beta x + \varepsilon$$

if y is measured imperfectly and what we observe is $\tilde{y} = y + \nu$, then we can rewrite the equation as:

$$\begin{aligned}(\tilde{y} - \nu) &= \beta x + \varepsilon\\ \tilde{y} &= \beta x + \varepsilon + \nu.\end{aligned}$$

The measurement error in y just gets added to the regression error. As long as ν is uncorrelated with x, then the regression coefficient will be unbiased (Pischke 2007; Wooldridge 2009). However, this will increase the error variance and thus make the estimates less precise.

We will look at the effect of measurement error in the dependent variable using an example shortly. For now though, let us be safe in the knowledge that the coefficient of interest is likely unbiased, and concentrate on the more serious problem of error in the independent variable.

The intuition behind the problem of measurement error in the independent variable(s) can be explained as follows. Take $\tilde{x} = x + \upsilon$ and substitute this into the equation for y:

$$\begin{aligned}y &= \beta x + \varepsilon\\ &= \beta(\tilde{x} - \upsilon) + \varepsilon\\ &= \beta\tilde{x} + (\varepsilon - \beta\upsilon).\end{aligned}$$

Since $\tilde{x}$ is obviously correlated with υ (unless the variance of υ is so small so that the correlation is essentially negligible), then the composite error in this regression is also correlated with the independent variable and thus the estimated coefficient of β will be biased.

The consequences of measurement error

To demonstrate the effect of measurement error on the FE-SEM model, and then provide a strategy for dealing with measurement error in SEM, the simulated dataset (file: `wideData.Rda`) generates multiple *indicators* of the time-varying independent and dependent variables that all measure the intended variable imprecisely. Let us drop the time-invariant covariate, z_i, for now for the sake of simplicity. Returning to our panel data, we have three indicators of each the independent and dependent variable, per timepoint:

$$\begin{aligned}\tilde{x}_{kt} &= x_t + \upsilon_{kt},\\ \tilde{y}_{kt} &= y_t + \nu_{kt}\end{aligned}$$

where $k = 1, 2, 3$ and $t = 1, ..., T$. This is like repeatedly presenting a respondent with a multi-item scale designed to measure things like stress, depression, xenophobia, etc. over the course of a panel study. To create the observed indicators, a random amount of measurement error (ranging from $\{\sigma^2_{\upsilon_k}, \sigma^2_{\nu_k}\} \in \{1.0, 1.1, 1.2, 1.3, 1.4, 1.5\}$) was added to the true variables, see the simulation code (file: `simulation-code.R`).

Let us first focus on the issue of imprecise measurements of the independent variable of interest and run the same FE-SEM model above, but this time we will use one of the measurement error sullied indicators, here $\tilde{x}_{1t}$, instead of the true independent variable, x_t. As for the naming conventions in the `R` code, `x11` stands for the first indicator ($k = 1$) at the first point in time ($t = 1$), whereas for example `x35` stands for the third indicator ($k = 3$) at the fifth point in time ($t = 5$).

```
fe_sem2 <- '
# Define individual effects variable
alpha =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
# Regressions, constrain coefficient to be equal over time
# Now the imprecisely measured indicator tilde{x}_kt
# instead of the true variable x_t
y1 ~ beta*x11
y2 ~ beta*x12
y3 ~ beta*x13
y4 ~ beta*x14
# Allow unrestricted correlation between eta and covariates
alpha ~~ x11 + x12 + x13 + x14
x11 ~~ x12 + x13 + x14
x12 ~~ x13 + x14
x13 ~~ x14
# Constrain residual variances to be equal over time
y1 ~~ nu*y1
y2 ~~ nu*y2
y3 ~~ nu*y3
y4 ~~ nu*y4
'
fe_sem2.fit <- sem(model = fe_sem2,
                   meanstructure = TRUE,
                   data = dfw,
                   estimator = "ML")
```

Now, for the sake of brevity, let us look just at the estimated coefficients for β.

```
summary(fe_sem2.fit)
```

```
...
##   y1 ~
##     x11     (beta)    0.192    0.008   23.800    0.000
##   y2 ~
##     x12     (beta)    0.192    0.008   23.800    0.000
##   y3 ~
##     x13     (beta)    0.192    0.008   23.800    0.000
##   y4 ~
##     x14     (beta)    0.192    0.008   23.800    0.000
...
```

Obviously, the estimated coefficient $\hat{\beta} = 0.192$ is substantially smaller than the true population coefficient of $\beta = 0.3$. And the discrepancy is not just due to sampling error. In fact, we can derive the bias we are observing here.

For a simple bivariate regression model, it is straightforward to quantify the bias due to measurement error. It will be

$$
\begin{aligned}
\operatorname{Cov}(y, \tilde{x}) &= \operatorname{Cov}[y\tilde{x}] \\
&= \operatorname{Cov}[(\beta\tilde{x} + \varepsilon)\tilde{x}] \\
&= \operatorname{Cov}[\beta\tilde{x}^2 + \varepsilon\tilde{x}] \\
&= \beta \operatorname{Var}(\tilde{x}) \\
\hat{\beta} &= \frac{\operatorname{Cov}(y, \tilde{x})}{\operatorname{Var}(\tilde{x})} \\
&= \frac{\operatorname{Cov}[(\beta x + \varepsilon)(x + \upsilon)]}{\operatorname{Var}[(x + \upsilon)^2]} \\
&= \frac{\operatorname{Cov}[\beta x^2 + \beta x\upsilon + \varepsilon x + \varepsilon\upsilon]}{\operatorname{Cov}[x^2 + 2x\upsilon + \upsilon^2]} \\
&= \beta \frac{\operatorname{Var}(x)}{\operatorname{Var}(x) + \operatorname{Var}(\upsilon)}.
\end{aligned}
$$

which results if we assume that $\operatorname{Cov}(x, \upsilon) = 0$, $\operatorname{Cov}(x, \varepsilon) = 0$, $\operatorname{Cov}(\tilde{x}, \varepsilon) = 0$ and $\operatorname{Cov}(\varepsilon, \upsilon) = 0$ (Wooldridge 2009). However, the model we are interested is not a bivariate model, so what was the point of showing the this? For one, it points out that the bias will always move the estimated coefficient closer to 0, since $\operatorname{Var}(x) \leq \operatorname{Var}(x) + \operatorname{Var}(\upsilon)$. This means positive effects will be biased downwards and negative effects biased upwards, always towards zero. This is why it is referred to as *attenuation bias*. Second, it will help to familiarize ourselves with this equation to better understand the one for the multivariate case.

Indeed, the magnitude of the bias in a multivariate model is somewhat more complex to derive, but it will be

$$
\hat{\beta} = \beta \frac{\operatorname{Var}(\theta)}{\operatorname{Var}(\theta) + \operatorname{Var}(\upsilon)}
$$

where θ is just the residual of a regression in which the underlying theoretical variable is regressed on all other covariates. In this case, we need to regress x_t on α for: $x_t = \tau + \phi\alpha + \theta_t$ where τ is the intercept, and ϕ is the regression coefficient and θ_t is the residual (Wooldridge 2009, 318–20).

Normally it is not possible to reconstruct the bias since in cases where we have to rely on indicators, we would not have observed the underlying theoretical variable. Furthermore, in the case of a fixed-effects model, the covariates are the unobserved time-invariant characteristics. However, because we are working with simulated data, we have everything we need. Going back to the results above, we can get the residuals of x_t by either running a regression and saving the residuals, or we could skip a step and get them directly using the 'residual maker' matrix (Rüttenauer and Ludwig 2020) which is $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^\intercal\boldsymbol{A})^{-1}\boldsymbol{A}^\intercal$ and $\boldsymbol{A} = (\boldsymbol{\iota}_n, \boldsymbol{\alpha})$ is the $n \times 2$ matrix of covariate(s) plus a constant.

```r
# Make the n x n identity matrix
Id <- diag(n)

# n x 2 matrix of covariates alpha
A <- matrix(c(rep(1, n), dfw$a1),
            nrow = n, ncol = 2)

# The residual maker matrix M = I - A(A'A)^-1 A'
M <- Id - A %*% solve(t(A) %*% A) %*% t(A)

# Save the residuals, t for 'theta'
```

```
t <- M %*% dfw$x1

# Re-run the FE model from above without z with the 'true'
# independent variable for the correct estimate for beta
fe_semx <- '
# Define individual effects variable
alpha =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
# Regressions, constrain coefficient to be equal over time
# Now the imprecisely measured indicator tilde{x}_kt
# instead of the true variable x_t
y1 ~ beta*x1
y2 ~ beta*x2
y3 ~ beta*x3
y4 ~ beta*x4
# Allow unrestricted correlation between eta and covariates
alpha ~~ x1 + x2 + x3 + x4
x11 ~~ x2 + x3 + x4
x12 ~~ x3 + x4
x13 ~~ x4
# Constrain residual variances to be equal over time
y1 ~~ nu*y1
y2 ~~ nu*y2
y3 ~~ nu*y3
y4 ~~ nu*y4
'
fe_semx.fit <- sem(model = fe_semx,
                   meanstructure = TRUE,
                   data = dfw,
                   estimator = "ML")

# The equation for the biased beta
lavInspect(fe_semx.fit, "list")[6, 14]*
  ((var(t))/(var(t) + var(dfw$x11 - dfw$x1)))
```

```
##            [,1]
## [1,] 0.2293898
```

From this we can see that the biased estimate above of $\hat{\beta} = 0.192$ roughly comes from $\beta\frac{\text{Var}(\theta_t)}{\text{Var}(\theta_t)+\text{Var}(\upsilon_t)} = 0.306\frac{6.36}{8.476} = 0.229$; 'roughly' because the equation here is the population equation. Due to sampling error, the estimates will tend vary slightly.

Using latent variables to deal with measurement error

The way we deal with measurement error in SEM is surprisingly similar to the logic of fixed-effects regression. Namely, if we have multiple cross-sectional observations of the underlying construct of interest, then we can define a latent variable that represents the common variance across those multiple variables. Contrast this with the use of longitudinal repeated measures to isolate the common variance across time.

So, if we do in fact have multiple cross-sectional indicators for the underlying variables of interest, then we can partition them into an explained and unexplained portion:

$$x_{kt} = \lambda^x_{kt}\xi_t + \delta_{kt},$$
$$y_{kt} = \lambda^y_{kt}\eta_t + \epsilon_{kt},$$

where x_{kt} and y_{kt} are the k^{th} indicators, ξ_t and η_t are latent factors representing the common variance across the cross-sectional repeated measures, and δ_{kt} and ϵ_{kt} are the unexplained portions of x_t and y_t, respectively. The latent factors are linked to the observed indicators through the factor loadings λ_{kt}.

Thus, our FE regression equation changes from $y_t = \beta x_t + \alpha + \nu_t$ to:

$$\eta_t = \beta\xi_t + \alpha + \zeta_t$$

where ζ_t represents the disturbance, in other words the residual of the latent dependent variable η_t. The model is shown in Figure 1. For the sake of legibility, the measurement model portion is shown only for the first timepoint.

[Figure 1 about here.]

First, however, let us double-check that measurement error in the dependent variable only increases the error variance (thus also increasing standard errors and reducing R^2), but does not systematically bias the coefficients of interest. The next model uses the indicators of x and specifies latent variables (ξ_t, `xi` in the code) to represent the valid cross-sectional variance. The dependent variable in the model is one of the imprecisely measured indicators of y.

```
fe_sem3 <- '
# Define individual effects variable
alpha =~ 1*y11 + 1*y12 + 1*y13 + 1*y14
# ----- MEASUREMENT MODEL FOR INDEPENDENT VARIABLE, xi
xi1 =~ 1*x11 + x21 + x31
xi2 =~ 1*x12 + x22 + x32
xi3 =~ 1*x13 + x23 + x33
xi4 =~ 1*x14 + x24 + x34
# Regressions, constrain coefficient to be equal over time
y11 ~ b*xi1
y12 ~ b*xi2
y13 ~ b*xi3
y14 ~ b*xi4
# Allow unrestricted correlation between eta and covariates
alpha ~~ xi1 + xi2 + xi3 + xi4
xi1 ~~ xi2 + xi3 + xi4
xi2 ~~ xi3 + xi4
xi3 ~~ xi4
# Constrain residual variances to be equal over time
y11 ~~ nu*y11
y12 ~~ nu*y12
y13 ~~ nu*y13
y14 ~~ nu*y14
'
fe_sem3.fit <- sem(model = fe_sem3,
                   meanstructure = TRUE,
                   data = dfw,
                   estimator = "ML")
```

```
summary(fe_sem3.fit)
```

```
...
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   y11 ~
##     xi1        (b)    0.309    0.015   20.790    0.000
##   y12 ~
##     xi2        (b)    0.309    0.015   20.790    0.000
##   y13 ~
##     xi3        (b)    0.309    0.015   20.790    0.000
##   y14 ~
##     xi4        (b)    0.309    0.015   20.790    0.000
...
```

The estimated coefficient here in model `fe_sem3.fit` is $\hat{\beta}_{y_{1t},\xi_t} = 0.309$ which is very close to the estimated coefficient in the first, correctly specified model `fe_sem.fit`, where $\hat{\beta}_{y_t,x_t} = 0.298$. Notice, however, that the standard error of the estimate is substantially larger, with 0.015 in `fe_sem3.fit` vs. 0.01 in `fe_sem.fit` in which y was measured without error. The explained variance (R^2) in the dependent variable was also much higher in the first model:

```
lavInspect(fe_sem.fit, "r2")[1:4]
```

```
##        y1        y2        y3        y4
## 0.8301378 0.8335652 0.8315385 0.8270869
```

compared to the current model:

```
lavInspect(fe_sem3.fit, "r2")[1:4]
```

```
##       y11       y12       y13       y14
## 0.7109445 0.7209264 0.7190412 0.7116103
```

Finally, to examine the effect of removing measurement error from the dependent variable in terms of standard errors and R^2 statistics, we can specify a model with latent variables representing the valid cross-sectional variance in y (`eta` for η in the code). The model is displayed in Figure 2, where again, for the sake of legibility, only the measurement models for the first timepoint are shown.

[Figure 2 about here.]

```
fe_sem4 <- '
# ----- NOW MEASUREMENT MODEL FOR DEPENDENT VARIABLE, n for eta
eta1 =~ 1*y11 + y21 + y31
eta2 =~ 1*y12 + y22 + y32
eta3 =~ 1*y13 + y23 + y33
eta4 =~ 1*y14 + y24 + y34
# Define individual effects variable
alpha =~ 1*eta1 + 1*eta2 + 1*eta3 + 1*eta4
# Measurement model for independent variables, xi
xi1 =~ 1*x11 + x21 + x31
```

```r
xi2 =~ 1*x12 + x22 + x32
xi3 =~ 1*x13 + x23 + x33
xi4 =~ 1*x14 + x24 + x34
# Regressions, constrain coefficient to be equal over time
eta1 ~ beta*xi1
eta2 ~ beta*xi2
eta3 ~ beta*xi3
eta4 ~ beta*xi4
# Allow unrestricted correlation between eta and covariates
alpha ~~ xi1 + xi2 + xi3 + xi4
xi1 ~~ xi2 + xi3 + xi4
xi2 ~~ xi3 + xi4
xi3 ~~ xi4
# Constrain residual variances to be equal over time
eta1 ~~ nu*eta1
eta2 ~~ nu*eta2
eta3 ~~ nu*eta3
eta4 ~~ nu*eta4
'
fe_sem4.fit <- sem(model = fe_sem4,
                   meanstructure = TRUE,
                   data = dfw,
                   estimator = "ML")
```

```r
summary(fe_sem4.fit)
```

```
...
## Regressions:
##                     Estimate  Std.Err  z-value  P(>|z|)
##   eta1 ~
##     xi1     (beta)    0.300    0.013   23.723    0.000
##   eta2 ~
##     xi2     (beta)    0.300    0.013   23.723    0.000
##   eta3 ~
##     xi3     (beta)    0.300    0.013   23.723    0.000
##   eta4 ~
##     xi4     (beta)    0.300    0.013   23.723    0.000
...
```

Here, the effect $\hat{\beta}_{\eta_t,\xi_t}$ is again very close to the true effect of 0.3. Again, however, if the main goal of the model is to avoid bias, it may be advisable to just leave the manifest dependent variable as it is, and worry about measurement error in the independent variables.

References

Allison, Paul D. 2003. "Missing Data Techniques for Structural Equation Modeling." *Journal of Abnormal Psychology* 112 (4): 545–57.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Best, Henning, and Christof Wolf. 2015. “Logistic Regression.” In *The Sage Handbook of Regression Analysis and Causal Inference*, edited by Henning Best and Christof Wolf, 153–71. London, Thousand Oaks: Sage Publications.

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.

CenterStat. 2019. “Can i Estimate an SEM If the Sample Data Are Not Normally Distributed?” 2019. https://centerstat.org/can-i-estimate-an-sem-if-the-sample-data-are-not-normally-distributed/.

Chamberlain, Gary. 1980. “Analysis of Covariance with Qualitative Data.” *Review of Economic Studies* 47(1): 225–38.

Croissant, Yves, and Giovanni Millo. 2008. “Panel Data Econometrics in R: The plm Package.” *Journal of Statistical Software* 27 (2): 1–43. https://doi.org/10.18637/jss.v027.i02.

Graff, J. 1979. “Verallgemeinertes LISREL-Modell.” Mannheim, Germany.

Graham, John W., and Donna L. Coffman. 2015. “Structural Equation Modeling with Missing Data.” In *Handbook of Structural Equation Modeling*, edited by Rick H. Hoyle, 277–95. New York, London: The Guilford Press.

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

Kline, Rex. 2016. *Principles and Practice of Structural Equation Modeling. Fourth Edition*. New York: The Guilford Press.

Kline, Rex B. 2015. “Assumptions in Structural Equation Modeling.” In *Handbook of Structural Equation Modeling*, edited by Rick H. Hoyle, 111–25. New York, London: The Guilford Press.

Lei, Pui-Wa, and Qiong Wu. 2015. “Estimation in Structural Equation Modeling.” In *Handbook of Structural Equation Modeling*, edited by Rick H. Hoyle, 164–79. New York, London: The Guilford Press.

Mundlak, Yair. 1978. “On the Pooling of Time Series and Cross Section Data.” *Econometrica* 46(1): 69–85.

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

Pischke, Jörn-Steffen. 2007. “Lecture Notes on Measurement Error.” http://econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf.

Rosseel, Yves. 2021. *The Lavaan Tutorial*. https://lavaan.ugent.be/tutorial/tutorial.pdf.

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.

Rüttenauer, Tobias, and Volker Ludwig. 2020. “Fixed Effects Individual Slopes: Accounting and Testing for Heterogeneous Effects in Panel Data or Other Multilevel Models.” *Sociological Methods and Research*.

West, Stephen G., John F. Finch, and Patrick J. Curran. 1995. “Structural Equation Modeling with Nonnormal Variables: Problems and Remedies.” In *Structural Equation Modeling: Concepts, Issues, and Applications*, edited by Rick H. Hoyle, 56–75. Thousand Oaks: Sage Publications.

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.

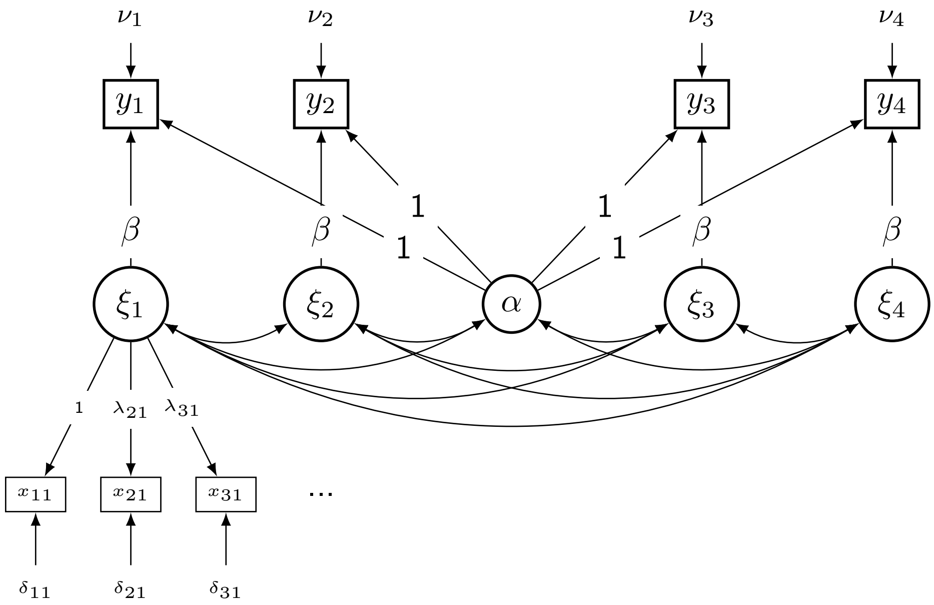

Figure 1: Four-wave FE model with measurement model for independent variable, shown only at $t = 1$

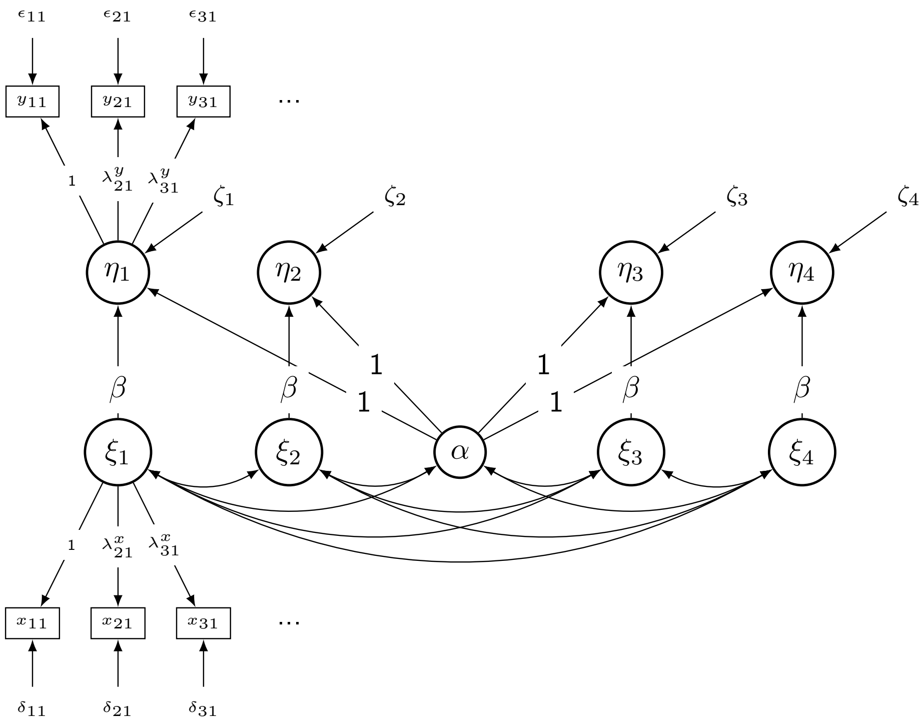

Figure 2: Four-wave FE model with measurement model for both variables, shown only at $t = 1$