Centering Predictor and Mediator Variables in Multilevel and Time-Series Models

Tihomir Asparouhov and Bengt Muthén

June 3, 2018

Overview

- Introduce the different centering options: 5 types
- Two-level regression
- Contextual effect: Lüdtke's bias
  - Fixed slope
  - Random slope
  - Probit regression
  - Binary mediator
- Time series:
  - Bias due to autocorrelation
  - Nickell's bias for random autocorrelation
  - Nickell's bias for random tetrachoric autocorrelation
- Two-level mediation
- Centering with missing data

How did we get here?

- Multilevel SEM based on latent centering: separate within and between
- Random slopes for within level covariate as in standard HLM
- Random slopes for mediator (uncentered) ML estimation: hybrid model
- Define commands: center Y(grandmean/groupmean) and YB=cluster_mean(Y)
- In V8: random slope for within level covariate with missing data
- In V8: random slope for latent centered lagged regressions
- We considered the issue too complex to straighten out. Two-level mediation models pushed this over the edge: too difficult with current methods
- In V8.1: full latent centering with Bayesian estimation

The 5 types of centering

- Observed group mean centering
- Grand mean centering
- Uncentered
- The hybrid
- The latent group mean centering

The standard two-level model by Raudenbush and Bryk(2002): the observed centering

Y_{ij} is the dependent variable and X_{ij} is the predictor for individual i in cluster j

$$Y_{ij} = \alpha_j + \beta_{1j}(X_{ij} - \overline{X}_{.j}) + \varepsilon_{w,ij}$$

$$\alpha_j = \alpha + \beta_2 \overline{X}_{.j} + \varepsilon_{b,j}$$

$$\beta_{1j} = \beta_1 + \xi_j$$

The contextual effect is $\beta_2 - \beta_1$: the group level effect of the covariate, i.e., the effect of the covariate beyond the individual level effect.

- Why do we need to separate X_{ij} into within $X_{ij} - \overline{X}_{.j}$ and between and $\overline{X}_{.j}$?
- Why do we need to center X_{ij}?
- Why do we treat differently X_{ij} and Y_{ij}? Here Y_{ij} uses a random effect for its intercept and X_{ij} does not - it uses the sample average.

The uncentered model

If we don't center, we estimate the model

$$Y_{ij} = \alpha_j + \beta_{0j} X_{ij} + \varepsilon_{w,ij}$$

$$\alpha_j = \alpha + \varepsilon_{b,j}$$

$$\beta_{0j} = \beta_0 + \xi_j$$

We can not estimate the two coefficients β_1 and β_2 and therefore we can not estimate the contextual effect.
We get β_0 to be the "uninterpretable blend" of β_1 and β_2

$$\beta_0 \approx \frac{w_1\beta_1 + w_2\beta_2}{w_1 + w_2}$$

$$w_1 = 1/Var(\hat{\beta}_1), w_2 = 1/Var(\hat{\beta}_2)$$

- Many multilevel studies focus on the difference between β_1 and β_2 (contextual effect). This is also sometimes referred to as the BFSP (big fish small pond) effect, particularly education studies, see Herbert Marsh's work

The uninterpretable blend: from Raudenbush and Bryk(2002)

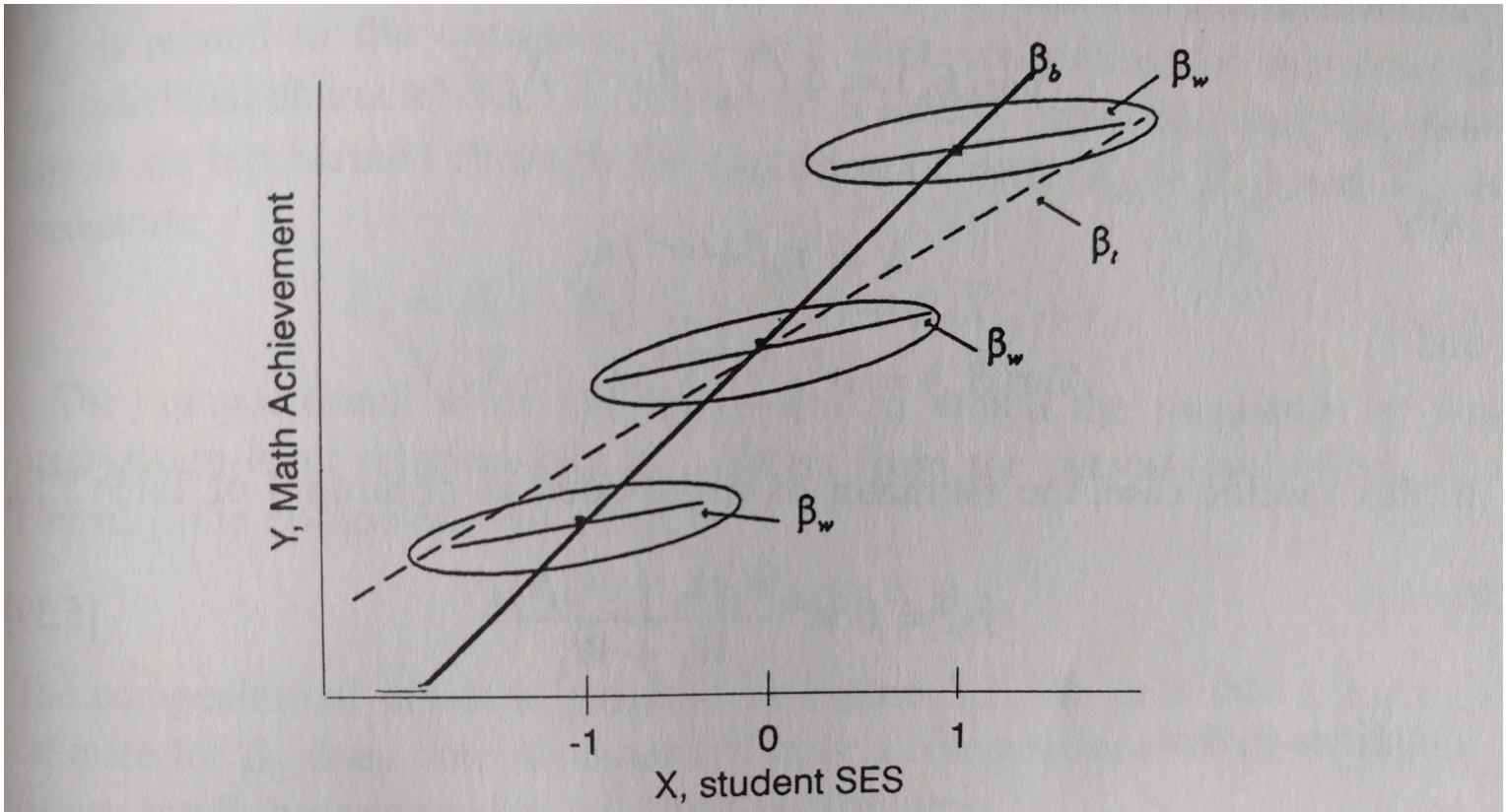


Figure 5.1. Within-Group Regression Lines (Inside the Elipses) have Slope β_w, Between Group Regression Lines (Bold Lines) has Slope β_b, and Total Regression (Dashed Line) Has Slope β_t.

The uncentered model with contextual effect

In the uncentered model we can estimate the contextual effect by adding the covariate $\overline{X}_{.j}$

$$Y_{ij} = \alpha_j + \beta_{1j} X_{ij} + \varepsilon_{w,ij}$$
$$\alpha_j = \alpha + \beta_2 \overline{X}_{.j} + \varepsilon_{b,j}$$
$$\beta_{1j} = \beta_1 + \xi_j$$

- If β_{1j} is a fixed slope this model is equivalent to the observed centered model. The within level effect is β_1 and the between level effect if $\beta_2 + \beta_1$
- If β_{1j} is random the model is not equivalent. The between level effect becomes $\beta_2 + \beta_1 + \xi_j$. The addition ξ_j is difficult to interpret. It represents the strange interaction contribution of ξ_j and $\overline{X}_{.j}$ added with a fixed slope of 1.
- The model fails to clearly separate within and between level effects. That interaction is often assumed to be zero so that the effects can be interpreted.

The grand mean centered model with contextual effect

$$Y_{ij} = \alpha_j + \beta_{1j}(X_{ij} - \overline{X}_{..}) + \varepsilon_{w,ij}$$
$$\alpha_j = \alpha + \beta_2(\overline{X}_{.j} - \overline{X}_{..}) + \varepsilon_{b,j}$$
$$\beta_{1j} = \beta_1 + \xi_j$$

- The model is equivalent to the uncentered model.
- We simply subtract a constant from the covariate.
- The reparameterization could be complex for larger models because the random effects change as well, not just the fixed effects

Lüdtke's bias

Lüdtke et al. (2008) shows that the observed centering does not estimate the contextual effect correctly. In the context of non-random slope the bias for β_2 is

$$(\beta_1 - \beta_2)\frac{(1-ICC)/n}{ICC+(1-ICC)/n} \tag{1}$$

where n is the size of the clusters, and ICC is the intraclass correlation for X_{ij}

- The bias becomes negligible if there is no contextual effect ($\beta_1 = \beta_2$) or the cluster size is large, but it increases as ICC $\to 0$
- The bias occurs because $\overline{X}_{.j}$ is a measurement of the mean and it has a measurement error that is no accounted for
- Lüdtke et al. (2008) shows that the latent centering / latent covariate approach based on multilevel SEM methodology eliminates the bias

Lüdtke's latent covariate model

$$X_{ij} = X_{w,ij} + X_{b,j}$$

$$Y_{ij} = \alpha_j + \beta_1 X_{w,ij} + \varepsilon_{w,ij}$$

$$\alpha_j = \alpha + \beta_2 X_{b,j} + \varepsilon_{b,j}$$

- $X_{b,j}$ is the true mean of X_{ij} in cluster j which is a latent variable.
- $X_{w,ij} = X_{ij} - X_{b,j}$ is the latent centered covariate on the within level.

In multilevel SEM context the model is written as

$$Y_{ij} = Y_{w,ij} + Y_{b,j}$$

$$Y_{w,ij} = \beta_1 X_{w,ij} + \varepsilon_{w,ij}$$

$$Y_{b,j} = \alpha_j = \alpha + \beta_2 X_{b,j} + \varepsilon_{b,j}$$

The hybrid centering

- If the slope is not random the latent centering model can be estimated with ML
- If the slope is random the likelihood does not have a closed form expression because it includes the product of two random effects on the within level

$$\beta_{1j} X_{w,ij} = \beta_{1j}(X_{ij} - X_{b,j}) = \beta_{1j} X_{ij} - \beta_{1j} X_{b,j}.$$

  Both β_{1j} and $X_{b,j}$ are random effects.
- It can be done with numerical integration, however, it is impractical for general modeling with multiple covariates.
- This hybrid model is however possible to estimate with ML.

$$X_{ij} = X_{w,ij} + X_{b,j}$$
$$Y_{ij} = \alpha_j + \beta_{1j} X_{ij} + \varepsilon_{w,ij}$$
$$\alpha_j = \alpha + \beta_2 X_{b,j} + \varepsilon_{b,j}$$
$$\beta_{1j} = \beta_1 + \xi_j$$

The hybrid centering

- This model is the same as the uncentered model except that on the between level we use the true mean $X_{b,j}$ instead of $\overline{X}_{.j}$ so potentially we could resolve Lüdtke's bias for models with random slopes.
- The model suffers from the same deficiencies as the uncentered model: it doesn't separate the within and the between effects well, and it results in the strange interaction ξ_j and $X_{b,j}$ on the between level.
- The separation of the effects could be dealt with using model analysis and reparameterizations, see Preacher et al. (2010) in mediation models

The hybrid centering

- For large cluster sizes ξ_j and $X_{b,j}$ are determined with little error and that extra interaction term will cause problems even with large samples and cluster sizes
- The hybrid method is Mplus default with the ML estimator. You can also use with ML or Bayes observed centering, uncentered or grand mean centering through the DEFINE commands center X(grandmean/groupmean) and XB=cluster_mean(X)
- The hybrid method does not accommodate missing data for X
- It requires a lot of work to make proper inference, for example Preacher et al. (2010) made incorrect computation for the indirect effect in 2-1-1 mediation model.
- Inference can become prohibitive for larger models

The hybrid centering

- Unclear if Mplus users understand what model they are running. This was big reason to turn to the latent centering.
- Major source of confusion is the fact that with non-random slopes the variables are latent centered, while with random slopes and the hybrid method the within part is uncentered while on the between level we have the latent centering variable.
- In the ML framework there are no great options
- In V8 the hybrid method was used also with Bayes.
- In V8.1 we have now switched to the latent centering with Bayes

The latent centering

$$X_{ij} = X_{w,ij} + X_{b,j}$$
$$Y_{ij} = \alpha_j + \beta_{1j} X_{w,ij} + \varepsilon_{w,ij}$$
$$\alpha_j = \alpha + \beta_2 X_{b,j} + \varepsilon_{b,j}$$
$$\beta_{1j} = \beta_1 + \xi_j$$

- The model can be estimated with Bayes.
- The change in the algorithm amounts to splitting the random effects into two blocks that are updated separately: β_{1j} is updated in one step, then α_j and $X_{b,j}$ are updated in a separate step.
- In each of the two steps the updating is based on standard two-level model updating because there is no product of random effects.
- $X_{b,j}$ participates in two equations. In the general case this is much more complex but we use multivariate modeling so it is not a problem.

The latent centering

- The model accommodates missing data on the covariate
- It separates clearly the within and the between effects
- The model estimation is more complex. In certain examples the estimation could be slower than other centering methods or it may fail to converge. In such situations you can still use all other centering methods.
- We have not seen a major drawback for that method and made this our default with Bayes.
- It can handle any number of covariates and random effects.
- Many mediation models that were previously untractable are now substantially simpler due to the clean separation in the within and the between levels

Simulation study: two-level regression model

$$X_{ij} = X_{w,ij} + X_{b,j} \tag{2}$$

$$Y_{ij} = \alpha_j + \beta_j X_{w,ij} + \varepsilon_{ij} \tag{3}$$

$$\varepsilon_{ij} \sim N(0, \sigma), X_{w,ij} \sim N(0, \psi) \tag{4}$$

$$\begin{pmatrix} X_{b,j} \\ \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \right) \tag{5}$$

- Within level regression with unstructured random effect model.
- 1000 clusters of size 15
- Technically the model is not a contextual effect model but if you convert it to such the within effect is 1 and the between effect is 0.5 so it is present

Simulation study: two-level regression model

Table: The two-level regression model: absolute bias(coverage)

Parameter	True Value	Latent	Observed	Uncentered	Hybrid
μ_1	1	.00(.93)	.00(.95)	.00(.95)	.00(.95)
μ_2	2	.00(.93)	.00(.96)	1.46(.00)	1.46(.00)
μ_3	1	.00(.95)	.00(.95)	.00(.96)	.00(.96)
σ_{11}	1	.01(.93)	.07(.73)	.07(.73)	.00(.94)
σ_{22}	1	.01(.91)	.13(.32)	1.75(.00)	1.75(.00)
σ_{33}	1	.01(.93)	.00(.93)	.10(.39)	.10(.39)
σ_{12}	0.5	.00(.93)	.06(.63)	1.45(.00)	1.45(.00)
σ_{13}	0.5	.01(.94)	.00(.96)	.04(.78)	.04(.78)
σ_{23}	0.5	.01(.91)	.00(.93)	1.38(.00)	1.38(.00)
σ	1	.00(.96)	.00(.97)	1.00(.00)	.01(.94)
ψ	1	.00(.96)	.07(.00)	.01(.94)	.00(.97)

Simulation study: two-level regression model

- Latent centering works well
- Observed centering has biased estimates for the variance covariance parameters of the random effects and the within level residual variance. These biases will disappear if we increase the cluster sizes to 100 or more
- Uncentered and Hybrid are similar and require major model reparametrization to make inference for the generating model

Simulation study: The contextual effect model with random slopes

$$X_{ij} = X_{w,ij} + X_{b,j} \tag{6}$$

$$Y_{ij} = \alpha_j + \beta_{1,j} X_{w,ij} + \varepsilon_{w,ij} \tag{7}$$

$$\alpha_j = \alpha + \beta_2 X_{b,j} + \varepsilon_{b,j} \tag{8}$$

$$\beta_{1,j} = \beta_1 + \beta_3 X_{b,j} + \xi_{b,j} \tag{9}$$

$$\begin{pmatrix} \varepsilon_{b,j} \\ \xi_{b,j} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}) \tag{10}$$

$$\varepsilon_{w,ij} \sim N(0, \sigma_w), X_{w,ij} \sim N(0, \psi_w), X_{b,j} \sim N(\mu, \psi_b). \tag{11}$$

500 clusters of size 15. Added a contextual effect for the random slope as well ($X_{b,j}$ predicts $\beta_{1,j}$).

Simulation study: The contextual effect model with random slopes

Table: Lüdtke's bias with random slope: absolute bias(coverage)

Parameter	True Value	Latent Centering	Observed Centering
α	2	.00(.95)	.06(.85)
β_1	-1	.01(.89)	.06(.84)
β_2	1	.00(.96)	.06(.71)
β_3	1	.00(.97)	.06(.73)
σ_w	1	.00(.97)	.00(.98)
σ_{11}	.9	.02(.94)	.20(.21)
σ_{12}	.5	.01(.96)	.06(.87)
σ_{22}	1	.01(.96)	.06(.95)
μ	1	.01(.97)	.01(.98)
ψ_w	1	.00(.98)	.06(.01)
ψ_b	.9	.01(.94)	.07(.84)

Simulation study: The contextual effect model with random slopes

- Latent centering works well
- Observed centering shows biased results for almost every parameter
- Lüdtke's bias is a misnomer term - the bias does not occur only for the estimation of the contextual effect - it is everywhere

Simulation study: The contextual effect model with random slopes

Table: Lüdtke's bias with random slope: results for β_2

Centering	Latent	Observed Group	Observed Grand	Uncentered	Hybrid
Reparameterization	β_2	β_2	$\beta_1+\beta_2$	$\beta_1+\beta_2+2\mu\beta_3$	$\beta_1+\beta_2+2\mu\beta_3$
Estimate	1.00	0.94	0.95	0.95	1.02
Standard error	.054	.049	.093	0.122	0.128
Coverage	.96	.71	.87	.94	.96
SMSE	.054	.082	0.123	0.123	0.121

- It is not just bias. The observed centering underestimates the SE by a factor of almost 2, much bigger SMSE
- The precision (SMSE) is much worse for Grand, Uncentered, Hybrid - 3 times bigger error
- The Hybrid reduces bias but at a substantial cost of precision - hard to recommend

Simulation study: Multilevel probit model with contextual effect

- Y is categorical, X is continuous

$$X_{ij} = X_{w,ij} + X_{b,j} \tag{12}$$

$$P(Y_{ij} = 0) = \Phi(\alpha_j + \beta_{1,j} X_{w,ij}) \tag{13}$$

$$\alpha_j = \alpha + \beta_2 X_{b,j} + \varepsilon_{b,j} \tag{14}$$

$$\beta_{1,j} = \beta_1 + \xi_{b,j} \tag{15}$$

$$\varepsilon_{b,j} \sim N(0, \sigma_b), X_{w,ij} \sim N(0, \psi_w), X_{b,j} \sim N(\mu, \psi_b), \xi_{b,j} \sim N(0, v) \tag{16}$$

- New feature in Mplus: cluster specific polyserial correlation
- Without random slope you can run the model in V8 (Bayes, WLSMV)
- With random slope only in V8.1 Bayes

Simulation study: Multilevel probit model with contextual effect

Table: Lüdtke's bias in multilevel probit regression with random slope: absolute bias(coverage) for $r = Cor(Y_{b,j}, X_{b,j})$

Cluster Size	Contextual Effect	Latent Centering	Observed Centering
15	Yes	.01(.93)	.09(.09)
15	No	.00(.96)	.00(.96)
50	Yes	.00(.94)	.03(.84)

- Lüdtke's bias exist with and without random slope

Simulation study: Multilevel linear model with contextual effect of a binary predictor

- Y is continuous, X is categorical
- X^* is the underlying continuous variable that is cut to obtain X

$$X^*_{ij} = X^*_{w,ij} + X^*_{b,j} \tag{17}$$

$$Y_{ij} = \alpha_j + \beta_{1,j} X^*_{w,ij} + \varepsilon_{w,ij} \tag{18}$$

$$\alpha_j = \alpha + \beta_2 X^*_{b,j} + \varepsilon_{b,j} \tag{19}$$

$$X_{ij} = 0 \iff X^*_{ij} < \tau \tag{20}$$

$$P(X_{ij} = 0|j) = \Phi(\tau - X^*_{b,j}) \tag{21}$$

$$\varepsilon_{b,j} \sim N(0, \sigma_b), \varepsilon_{w,ij} \sim N(0, \sigma_w), X^*_{w,ij} \sim N(0,1), X^*_{b,j} \sim N(0, \psi_b). \tag{22}$$

Why/when should we consider binary X on the latent scale instead of the observed? Enders and Tofighi (2007)

- If the distribution of X is not invariant across cluster: the binary variable is already dependent. It is affected by the cluster so it is best treat it as such, otherwise you get model misspecification. The best way to treat the binary variable in multilevel model is through the multilevel probit regression and the latent scale
- Missing data on the binary predictor - that requires full modeling for proper missing data imputation
- The predictor is a mediator. The proper regression modeling requires the multilevel probit model and the latent scale.
- Observed centering for binary items fails to separate the within and the between level into two-independent parts, i.e., not possible to evaluate properly the contextual effect and some of the between effect is still channelled through the within level. Variance on the within is determined by the between variable (the mean determines the variance for binary).

Simulation study: Multilevel linear model with contextual effect of a binary predictor: non-random slope

Table: Lüdtke's bias in multilevel regression with binary predictor: absolute bias(coverage) for r

Cluster Size	Contextual Effect	Latent Centering	Observed Centering
15	Yes	.00(.94)	.12(.01)
50	Yes	.00(.95)	.06(.35)
100	Yes	.00(.93)	.04(.63)
15	No	.00(.91)	.03(.79)
50	No	.00(.95)	.03(.87)
100	No	.00(.93)	.03(.80)

Lüdtke's bias exists. Additional bias due to the non-linearity of the link function (increases as ICC for X increases, opposite direction of Lüdtke's bias). Latent centering resolves the bias. The bias on the within-level is much higher: 0.25 and it doesn't disappear for large clusters or without contextual effect.

Simulation study: Multilevel linear model with contextual effect of a binary predictor: random slope

Table: Lüdtke's bias in multilevel regression with binary predictor and random slope: absolute bias(coverage) for r

Cluster Size	Contextual Effect	Latent Centering	Observed Centering
15	Yes	.01(.99)	.07(.65)
50	Yes	.00(.96)	.05(.94)
15	No	.01(1.00)	.01(.98)
50	No	.00(.96)	.00(.93)

Lüdtke's bias exists. Latent centering resolves the bias. Non-linearity bias is small due to smaller ICC.

Time-series models

- There are two separate issues
  - Contemporaneous centering for a predictor from the same time period: Lüdtke's bias, bias due to autocorrelation
  - Lag centering of the lag-variable predictor (same variable from a different period): Nickell's bias
- Latent centering available in DSEM V8.1: Lag and Contemporaneous, for both continuous and categorical, for both random and non-random slopes
- Latent centering available in RDSEM V8.1: For continuous, Lag and Contemporaneous, for both random and non-random slopes
- Latent centering available in RDSEM V8.1: For categorical, Some models (not all), Contemporaneous, for both random and non-random slopes
- New applications studies featuring these:
  - Contemporaneous continuous latent centering by Hamaker in DSEM
  - Contemporaneous continuous latent centering by Muthén in RDSEM

Time-series models: bias due to autocorrelation

Y_{it} and X_{it} are the dependent variable and the covariate for individual i at time t

$$X_{it} = X_{b,i} + X_{w,it} \tag{23}$$

$$Y_{it} = \alpha_i + \beta_1 X_{w,it} + \varepsilon_{it} \tag{24}$$

$$\alpha_i = \alpha + \beta_2 X_{b,i} + \varepsilon_i \tag{25}$$

We now add the two new autocorrelation equations

$$\varepsilon_{it} = r_y \varepsilon_{i,t-1} + \delta_{it} \tag{26}$$

$$X_{w,it} = r_x X_{w,i,t-1} + \xi_{it} \tag{27}$$

$$\delta_{it} \sim N(0, \sigma_1), \xi_{it} \sim N(0, \psi_1), \varepsilon_i \sim N(0, \sigma_2), X_{b,i} \sim N(\mu, \psi_2) \tag{28}$$

This is a new model in Mplus 8.1 based on RDSEM(residual dynamic structural equation models). Simulation with N=200 individuals.

Time-series models: bias for β_1 due to autocorrelation

Table: Absolute bias(coverage) for β_1

Time Points	Contextual Effect	r_y/r_x	Latent Centering with autocorrelation	Latent Centering without autocorrelation	Observed Centering	REML Observed Centering with r_y
30	Yes	.7/.7	.00(.93)	.00(.69)	.00(.69)	.00(.95)
60	Yes	.7/.7	.00(.94)	.00(.82)	.00(.82)	.00(.96)
100	Yes	.7/.7	.00(.90)	.00(.70)	.00(.70)	.00(.90)
30	No	.7/.7	.00(.95)	.00(.69)	.00(.69)	.00(.96)
100	No	.7/.7	.00(.91)	.00(.70)	.00(.70)	.00(.90)
30	Yes	.7/.0	.00(.93)	.00(.97)	.00(.97)	.00(.92)
30	Yes	.0/.7	.00(.89)	.00(.90)	.00(.90)	.00(.89)
30	Yes	.0/.0	.00(.94)	.00(.95)	.00(.95)	.00(.95)

Time-series models: bias for β_1 due to autocorrelation

- No bias in the parameter estimates
- Bias in the standard error when $r_y > 0$ and $r_x > 0$ for the observed and latent centering without the autocorrelation
- The bias exist even with large samples, large number of time points, and without contextual effect.
- The bias disappear if either $r_y = 0$ or $r_x = 0$
- The bias occurs due to overestimation of the number of independent observations when ignoring the autocorrelation
- Latent centering with autocorrelation and REML resolves the problem

Time-series models: bias for β_2 due to autocorrelation

Table: Absolute bias(coverage) for β_2

Time Points	Contextual Effect	r_y/r_x	Latent Centering with autocorrelation	Latent Centering without autocorrelation	Observed Centering	REML Observed Centering with r_y	Analytically Derived Observed Centering Bias
30	Yes	.7/.7	.14(.83)	.44(.01)	.50(.00)	.51(.00)	.51
60	Yes	.7/.7	.02(.92)	.26(.18)	.31(.03)	.30(.05)	.31
100	Yes	.7/.7	.02(.93)	.16(.54)	.19(.36)	.18(.37)	.20
30	No	.7/.7	.01(.96)	.00(.98)	.00(.96)	.00(.96)	.00
100	No	.7/.7	.00(.96)	.00(.95)	.00(.95)	.00(.95)	.00
30	Yes	.7/.0	.00(.93)	.00(.97)	.07(.89)	.07(.88)	.13
30	Yes	.0/.7	.16(.85)	.44(.00)	.50(.00)	.50(.00)	.51
30	Yes	.0/.0	.01(.98)	.01(.97)	.07(.84)	.07(.87)	.13

Time-series models: bias for β_2 due to autocorrelation

- Large parameter estimate bias when $r_x > 0$ and contextual effect, for REML the observed and latent centering without the autocorrelation
- The bias disappears when $r_x = 0$ or with zero contextual effect
- The bias adds on to Lüdtke's bias so observed centering and REML are worse than latent centering without the autocorrelation
- The bias depends on the size of r_x but it can be much worse than Lüdtke's bias: 6 times bigger in this simulation
- The bias tends to disappear as the number of time points increases but much slower than Lüdtke's bias and can be found also when number of time points is > 100. The number of time points needed to eliminate the bias depends on r_x and ICC
- The bias occurs due to not properly accounting for the measurement error in the mean of the covariate

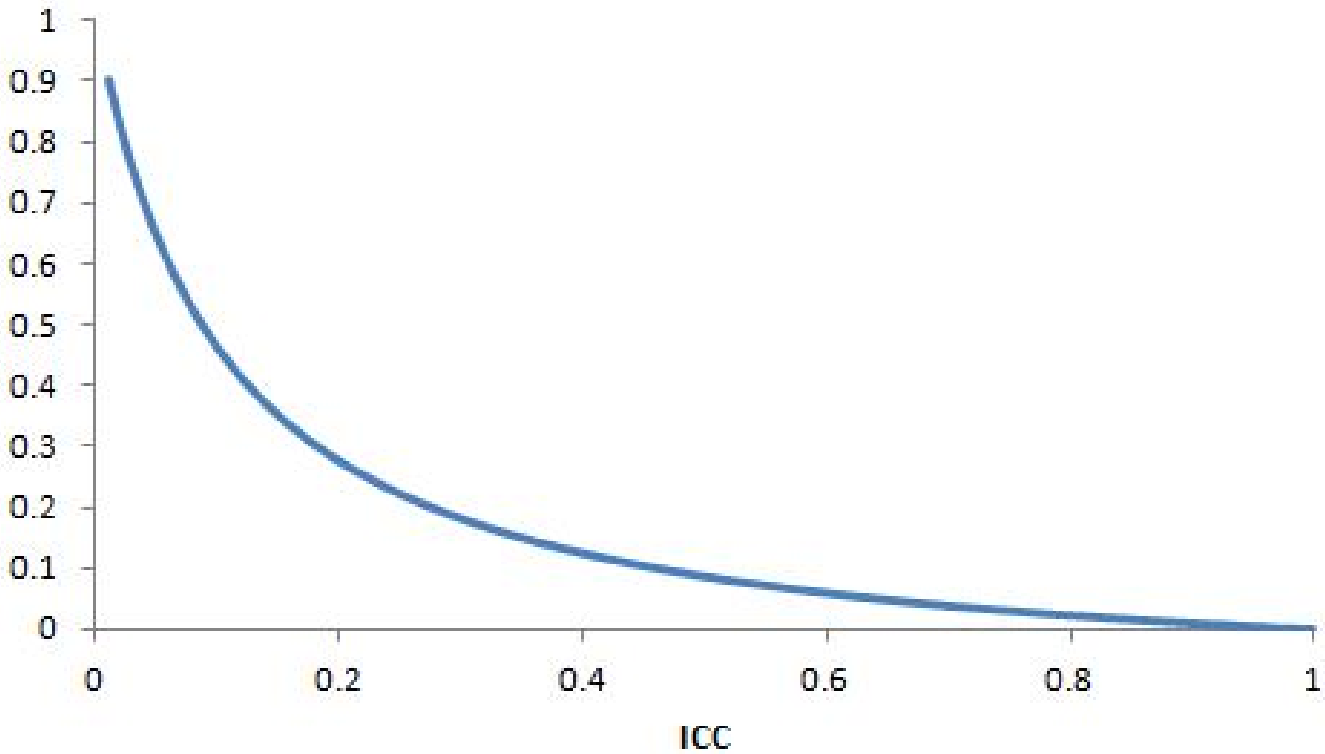
Time-series models: bias for β_2 due to autocorrelation

- Common misconception in statistical practice for multilevel time-series modeling: the focus is on r_y while r_x is typically ignored. As seen here r_x has much bigger impact.
- Latent centering with autocorrelation resolves the problem
- The analytically derived observed centering bias is

$$(\beta_1 - \beta_2)\frac{(1-ICC)/T^*}{ICC+(1-ICC)/T^*} \tag{29}$$

$$T^* = T\frac{1-r_x}{(1+r_x)(1-2r_x/(T(1-r_x^2)))} \tag{30}$$

Figure: Relative contextual bias as a function of ICC, $r_x = 0.5$, $T = 30$



Time-series models: Nickell's bias

Consider the following two-level AR model

$$Y_{it} - \mu_i = \phi_i(Y_{i,t-1} - \mu_i) + \xi_{it} \tag{31}$$

$$\mu_i = \mu + \varepsilon_{i1} \tag{32}$$

$$\phi_i = \phi + \varepsilon_{i2} \tag{33}$$

$$\xi_{it} \sim N(0, \sigma), \varepsilon_{i1} \sim N(0, \theta_1), \varepsilon_{i2} \sim N(0, \theta_2). \tag{34}$$

If observed centering is used for the predictor the estimate for ϕ is biased. The bias is know as Nickell's bias and is approximately

$$-\frac{1+\phi}{T-1}, \tag{35}$$

where T is the number of observations in the time-series (applies for fixed and random slopes). The bias again occurs because sample mean is used intead of true mean without accounting for the error. The bias is resolved with latent centering in V8.

Time-series models with categorical variables: Random tetrachoric autocorrelation

Y_{ij} is binary, Y^*_{it} the underlying continuous variable

$$Y^*_{it} - \mu_i = \phi_i(Y^*_{i,t-1} - \mu_i) + \xi_{it} \tag{36}$$

$$\mu_i = \mu + \varepsilon_{1i} \tag{37}$$

$$\phi_i = \phi + \varepsilon_{2i} \tag{38}$$

$$P(Y_{it} = 1) = P(Y^*_{it} > 0) \tag{39}$$

$$\xi_{it} \sim N(0,1), \varepsilon_{i1} \sim N(0,\theta_1), \varepsilon_{i2} \sim N(0,\theta_2). \tag{40}$$

Time-series models with categorical variables

- New model in V8.1.
- Time series models for categorical variables in V8 required large samples, a latent variable behind that allows the AR model, and slow to converge
- This AR model is directly for the categorical variable (no need of a latent variable), can work with as low as 20 time points, and it is much faster to converge.
- ϕ_i is a random tetrachoric autocorrelation - new concept
- The model is a competitive alternative to Markov chain models - lags$>$1, covariates, fits in two-level DSEM well, random transition probabilities
- It can be used for categorical variables with more than two categories and is much more parsimonious than Markov chain models. With 5 categories Markov chain estimates 24 parameters while this model estimates 5.

Table: Nickell's bias for the random tetrachoric autocorrelation, $\phi = 0.3$, 5000 clusters

Cluster Size	Latent Centering	Observed Centering	Uncentered	Bivariate centered	Bivariate uncentered
20	.01	-.16	-.05	-.08	.04
50	.00	-.12	-.06	-.03	.03
200	.00	-.09	-.05	.01	.03

The bias is computed on correlation scale: standardized estimate of slope in each cluster using the OUTPUT:STAND(CLUSTER); Nickell's bias does not disappear as cluster size increases. Latent centering is the only viable alternative.

Table: Bias and coverage for the mean of the random tetrachoric autocorrelation, $\phi = 0.3$, 100 clusters

Cluster Size	Latent Centering
20	.01(.98)
50	.01(.90)
200	.00(.98)

No bias, good coverage.

Two-level mediation: 2-1-1 case

$$Y_{ij} = Y_{w,ij} + Y_{b,j} \tag{41}$$

$$M_{ij} = M_{w,ij} + M_{b,j} \tag{42}$$

$$Y_{w,ij} = \beta_{1,j} M_{w,ij} + \varepsilon_{w,ij} \tag{43}$$

$$Y_{b,j} = \alpha_1 + \beta_2 M_{b,j} + \beta_3 X_j + \varepsilon_{b,j}. \tag{44}$$

$$M_{b,j} = \alpha_2 + \beta_4 X_j + \xi_{b,j}. \tag{45}$$

$$\varepsilon_{w,ij} \sim N(0, \sigma_w), M_{w,ij} \sim N(0, \psi_w) \tag{46}$$

$$\varepsilon_{b,j} \sim N(0, \sigma_b), \beta_{1,j} \sim N(\beta_1, \theta), \xi_{b,j} \sim N(0, \psi_b) \tag{47}$$

Preacher et al. (2010) (1154 citations) used the hybrid method to compute the indirect effect using $(\beta_1 + \beta_2)\beta_4$. Using latent or observed centering the indirect effect is $\beta_2\beta_4$

Two-level mediation: 2-1-1 case

Table: Indirect effect: absolute bias / coverage / SMSE

Number of Clusters	Cluster Size	Contextual effect	Latent	Observed	Hybrid
500	50	No	.00/.94/.064	.00/.92/.063	.01/.93/.122
500	50	Yes	.00/.93/.066	.03/.90/.071	.01/.92/.122
500	20	No	.01/.93/.058	.01/.94/.057	.00/.98/.106
500	20	Yes	.01/.95/.062	.09/.77/.104	.00/.97/.106
20	20	No	.06/.89/.376	.03/.88/.357	.05/.88/.559
20	20	Yes	.07/.93/.393	.56/.65/.856	.52/.71/.637
15	15	No	.03/.98/.528	.01/.91/.423	.04/.90/.746
15	15	Yes	.08/.97/.486	.58/.68/.901	.49/.67/.692

Latent centering is much better than the Observed or Hybrid in terms of bias, coverage and MSE. Hybrid method is the worst with large clusters, see MSE. McNeish (2017) claimed that Observed centering is best - generating data only without contextual effect.

Two-level mediation: 2-1-1 case

Random effects can be correlated. Add to the previous model the correlation between the mediator random intercept and the random slope

$$\begin{pmatrix} \beta_{1,j} \\ \xi_{b,j} \end{pmatrix} \sim N(\begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}, \begin{pmatrix} \theta & \rho \\ \rho & \psi_b \end{pmatrix}). \tag{48}$$

Preacher et al. (2010) formula (for the hybrid/uncentered) is no longer accurate. We can obtain the correct formula using causal methods / potential outcomes

$$TNIE = E[Y(1,M(1))] - E[Y(1,M(0))] = (\beta_1 + \beta_2 + \alpha_2\rho/\psi_b)\beta_4$$

This computation however can be very complicated for more advanced models.

Two-level mediation: 2-1-1 case

Table: Indirect effect for correlated random slope and between mediator

Centering	Latent	Hybrid Preacher et al. (2010)	Hybrid
Formula	$\beta_2\beta_4$	$(\beta_1+\beta_2)\beta_4$	$(\beta_1+\beta_2+\alpha_2\rho/\psi_b)\beta_4$
Bias	.00	-.77	.00
Coverage	.93	.00	.90
SMSE	.057	.782	.149

The correct formula is clearly better but still latent centering gives even better results. Latent centering is also more general (doesn't depend on which correlations are included in the model) and it is much simpler to compute.

Missing Data on the mediator

In the two-level mediation example (2-1-1 case) we generate missing data for the mediator.

- MAR

$$P(M_{ij} \text{ is missing}) = \frac{1}{1+Exp(1+0.5Y_{ij})}. \quad (49)$$

- MCAR

$$P(M_{ij} \text{ is missing}) = \frac{1}{1+Exp(1)} \quad (50)$$

- Observed centering can still be performed using the sample average of the non-missing values
- We do not include contextual effect for this simulation to separate this bias from Lüdtke's bias

Missing Data on the mediator

Table: Indirect effect with missing data on the mediator: absolute bias/coverage/SMSE

Missing Data	Latent Bayes	Observed Bayes	Observed ML + montecarlo	Observed ML + listwise
MCAR	.00/.93/.064	.01/.92/.062	.02/.91/.068	.00/.91/.064
MAR	.00/.92/.063	.09/.68/.108	.10/.58/.121	.13/.39/.142
Comp Time	3 sec	5 sec	16 min	1 sec

Latent centering outperforms if the missing data is MAR.

Summary

Latent centering solves many problems

- Separation of level effects
- Nickell's bias
- Lüdtke's bias
- Bias due to autocorrelation
- Missing data bias
- Link function related bias
- More accurate
- Simpler to interpret