Rehabilitating the Lagged Dependent Variable with Structural Equation Modeling

Henrik Kenneth Andersen Jochen Mayerl

2022-06-22

It has been argued that including the lagged dependent variable in panel models will open up unintended back-door paths and bias the estimates of the causal variable. We show that panel analysis in the structural equation modeling framework can get around this issue. Including the lagged dependent variable has the benefit of closing back-door paths due to unobserved time-varying confounders. We demonstrate this by looking at simulated data in the `lavaan` package for `R`.

Table of contents

1 Introduction

There is a long tradition in sociology and psychology of using cross-lagged panel models to investigate dynamic processes (Rogosa 1980) in structural equation modeling (SEM). These usually look at the lagged bidirectional effects of two variables (cross-lagged paths, CL) while holding the previous values of these variables constant (autoregressive paths, AR).

At the same time, such models have been shown to be biased in the presence of time-invariant unobserved heterogeneity. Effective ways to incorporate time-invariant unobserved heterogeneity into cross-lagged panel models have been around for several decades now (e.g., Kenneth A. Bollen and Curran 2004; Curran and Bollen 2001) and this topic has experienced renewed interest in the last several years (Curran et al. 2014; Hamaker, Kuiper, and Grasman 2015; Zyphur, Allison, et al. 2020; Zyphur, Voelkle, et al. 2020).

The basic idea of cross-lagged panel models that account for unobserved time-invariant confounders can also be generalized to 'unidirectional' models that focus on the effect of one variable on another while de-emphasizing the question of reverse causality (Allison, Williams, and Moral-Benito 2017; Moral-Benito, Allison, and Williams 2018; Williams, Allison, and Moral-Benito 2018). We could call these 'dynamic fixed effects' models because they all account for both unobserved time-invariant heterogeneity as well as state dependence by including the lagged dependent variable (lagged DV or LDV).

Indeed (Kühnel and Mays 2018) argue that the inclusion of the LDV make such models more desirable than 'static fixed effects' models (which do not include the lagged dependent variable) because they potentially account for confounding by certain time-varying variables, as well.

Still, many are skeptical of the use of such models. Often, the skepticism centers directly on the use of the LDV. Indeed, many articles warn of including the LDV in a panel regression model (e.g., Brüderl and Ludwig 2014; Dafoe 2014, 2015; Foster 2010; Collischon and Eberl 2020; Keele and Kelly 2006; Leszczensky and Wolbring 2019; Mouw 2006; Walters 2019). One of the most convincing arguments is given by Morgan and Winship (2014, 111, Figure 4.3). There, they show that the inclusion of the LDV may bias the causal effect of interest in the presence of time-invariant unobserved heterogeneity.

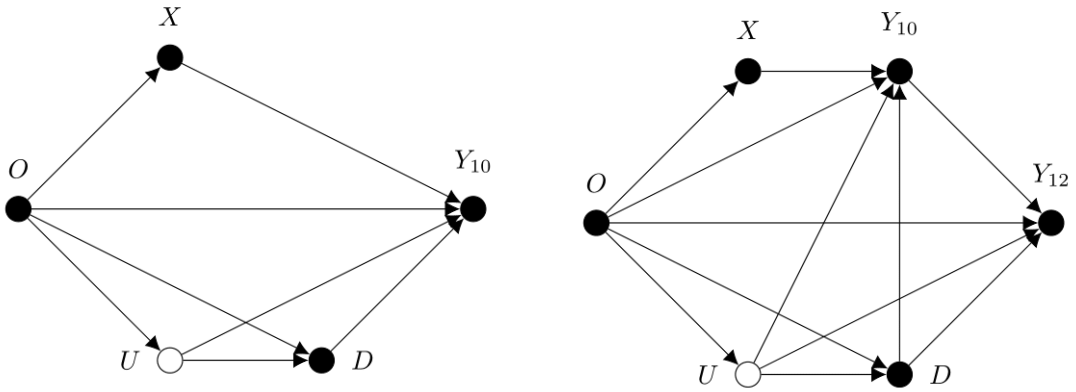
In this brief article, we will outline the arguments for and against the inclusion of the LDV and then show that the usual SEM approach to panel modelling is not generally affected by the criticism. We hope to convince readers of the usefulness of the broad class of (cross-lagged) dynamic panel models with fixed effects in SEM.

2 Benefits of the LDV

Consider an empirical example by Coleman, Hoffer, and Kilgore (1982) and outlined in Morgan and Winship (2014). There, Coleman and colleagues were looking to assess the causal effect of attending a Catholic school as opposed to public school on achievement, as measured by pupils' test scores. The direct acyclic graph (DAG) shown in Figure 1a summarizes their hypothesized data generating process (DGP), where black circles represent observed variables and white ones represent unobserved variables.[1]

In this DAG, Y_{10} represents the pupil's test score in grade 10, X are observed determinants of test scores and O are observed background factors that influence test scores, the determinants of test scores, the selection of school system, as well as the unobserved factors in U.

[1] Empirical examples chosen for this article were taken from the helpful video tutorial on LDVs by Mikko Rönkkö, https://www.youtube.com/watch?v=DhV5otUB3Jc.



(a) Without LDV

(b) With LDV

Figure 1: Catholic school example from Coleman, Hoffer, and Kilgore (1982)

For its part, U could contain any number of potentially umeasured things like motivation or intelligence, that would impact the choice of school system (Catholic schools might prefer to admit intelligent pupils) and the pupil's test scores themselves.

Conditioning on X and O would close all back-door paths except $D \leftarrow U \rightarrow Y_{10}$. Because U contains all the unobserved determinants of the causal variable and the outcome, it cannot be conditioned on and the causal effect of interest is unidentified (Morgan and Winship 2014, 270).

Coleman, Hoffer, and Kilgore (1982) came to a solution to this problem by collecting data on the pupils' test scores two years later, in grade 12. By looking at the DAG in Figure 1b, we can see that Y_{10} "screen[s] off the effects of the variables in U on Y_{12}" (Morgan and Winship 2014, 270). Indeed, by focusing on test scores in grade 12 and including the lagged measure from grade 10 in the model, the back-door paths over the unobservables in U are blocked and the causal effect of D on Y_{12} is identified.

Note that U could potentially include the unmeasured outcome even further in the past. In this way, the inclusion of the lagged dependent variable also accounts for endogeneity, where the causal variable is impacted by previous outcomes. An example given by Wooldridge (2012, 313) concerns the explanation of crime as a function of police expenditure. It is plausible that more is spent on policing in areas where crime rates have been high in the past. Simply regressing the crime rate on expenditure will likely be misleading, because where there is a high crime rate, there will be increased spending, so the effect of spending on crime may even be positive.[2] Including the lagged crime rate allows for an intuitive interpretation of the effect of expenditure: it is the difference in crime rate between two hypothetical cities with the same crime rate in the previous period given a unit change in expenditure (Wooldridge 2012).

[2]This is not to justify police expenditure, which is a difficult topic in some parts of the world. It may still be that police spending has a positive causal effect on crime, or that there is no tangible effect. But the estimated effect in a simple regression of current crime rate on current expenditure will likely be biased in one way or another.

3 Arguments Against the LDV

To continue with outlining the issue, let us turn to a simpler DAG as shown in Figure 2, which is adapted also from Morgan and Winship (2014, 111, Fig. 4,3).[3] In this DAG, we have dropped the observed variables captured in X and O from above. Since they are observed, we can drop them from the following discussion without loss of generality. We still have D, the causal variable of interest, and two measures of the outcome, represented by Y_1 and Y_2. Consistent with the Catholic school example, the inclusion of Y_1 as an observed predictor of Y_2 closes the back-door path across $D \leftarrow U \rightarrow Y_1 \rightarrow Y_2$.

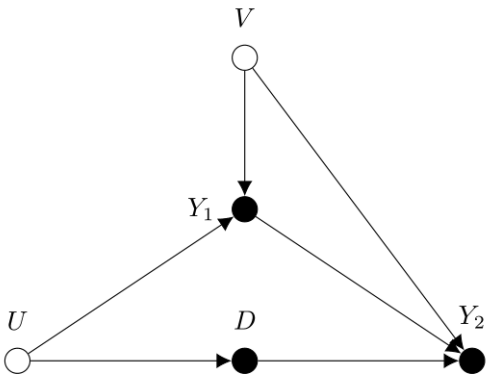


Figure 2: Lagged DV as collider from Morgan and Winship (2014)

The novelty of this DAG compared to the previous ones is the inclusion of V. It represents the time-invariant unobserved heterogeneity. It is an unobserved variable that affects the outcome at all points in time. It can also be thought of as the combined effect of all the time-invariant factors affecting the outcome. The issue, outlined in Morgan and Winship (2014), is that if V is unobserved, then Y_1 is also a collider variable and controlling for it therefore opens a new back-door path over V, rendering the causal effect biased.

This point is often the one criticisms of the LDV focus on. Keele and Kelly (2006, 187) write, for example, that "[e]ven when a lagged dependent variable is theoretically appropriate, remaining residual autocorrelation can lead to biased coefficient results." Here, the remaining residual autocorrelation is the time-invariant unobserved heterogeneity; the unobserved stable factors that cause the outcome to be correlated with itself over time (Andersen 2022). Foster (2010, 1467) echos this, stating "[s]uch analyses [that include the LDV] are problematic. As has long been known in economics and other fields, in the presence of autocorrelation (a relationship between unobservables over time), the resulting estimates have poor statistical properties." Collischon and Eberl (2020, 297) argue against the LDV in a similar fashion.

The intuition for these arguments can be shown easily. To simplify things, let us assume variables are mean centered and scaled to have a variance of one, as Cinelli, Forney, and Pearl

[3] The same thing is shown in the variation of Model 7 in Cinelli, Forney, and Pearl (2022), where Z takes the place of the LDV.

(2022) do. Then, by path tracing, labelling the structural coefficients λ, the covariances of the observed variables, Y_2, Y_1, D are given by

$$\sigma_{Y_2 D} = \lambda_{DY_2} + \lambda_{UD}\lambda_{UY_1}\lambda_{Y_1 Y_2} \tag{1}$$
$$\sigma_{DY_1} = \lambda_{UD}\lambda_{UY_1} \tag{2}$$
$$\sigma_{Y_2 Y_1} = \lambda_{Y_1 Y_2} + \lambda_{VY_1}\lambda_{VY_2} + \lambda_{UY_1}\lambda_{UD}\lambda_{DY_2}. \tag{3}$$

The partial coefficient of the causal variable, D, on the outcome, Y_2, controlling for Y_1, is given by

$$\beta_{Y_2 D \cdot Y_1} = \frac{\sigma_{Y_2 D} - \sigma_{DY_1}\sigma_{Y_2 Y_1}}{1 - \sigma^2_{DY_1}} \tag{4}$$

(Cinelli, Forney, and Pearl 2022) which, after substitution, works out to

$$\beta_{Y_2 D \cdot Y_1} = \lambda_{DY_2} - \frac{\lambda_{UD}\lambda_{UY_1}\lambda_{VY_1}\lambda_{VY_2}}{1 - (\lambda_{UD}\lambda_{VY_1})^2} \tag{5}$$

which does not equal the structural coefficient λ_{DY_2}, the average causal effect.

This shows the bias resulting from the LDV and Dafoe (2015, 139) suggests that it is therefore only safe to include the LDV when it is not a collider. That is the case when either "there are no unobserved common causes of treatment and the lagged outcome" (if U were missing from the DAG in Figure 2) or "no unobserved persistent causes of the outcome" (if V were missing). Partly because of this (along with other theoretical reasons), Brüderl and Ludwig (2014, 342) propose flatly that "LDV models are not useful at all."

4 The Structural Equation Modeling Approach

The key to the SEM approach, and the reason why cross-lagged and other panel models that include the LDV are so widespread in SEM, has to do with the fact that by using latent variables, the LDV does not open an unblocked back-door path. Panel models in SEM that account for time-invariant unobserved heterogeneity normally work by specifying a latent variable that causes the observed outcome at all points in time. This explicitly accounts for the "remaining residual autocorrelation" (Keele and Kelly 2006, 187), or "persistent causes of the outcome" (Dafoe 2015, 139).

In SEM, the time-invariant unobserved heterogeneity is not unobserved in the classical sense. It represents the conditional (on the other observed variables) covariance between the outcome and itself over time (Andersen 2022). Say we had the linear model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + \epsilon_{it},\ i = 1, \dots, N,\ t = 1, \dots, T \tag{6}$$

where $\mathbf{x}_{it} = (d_{it}, y_{it-1})$ and $\boldsymbol{\beta} = (\beta, \rho)^\top$, α_i is a latent variable representing the stable factors that change between individuals but not within them, and ϵ_{it} is the time-varying error component. Then the covariance of any two columns of the wide-format outcome, conditional on the observed covariates, is just

$$\text{Cov}(y_{it}, y_{is}|\mathbf{x}_{it}) = \text{Var}(\alpha_i),\ t \neq s \tag{7}$$

since we normally assume $\text{Cov}(\epsilon_{it}, \epsilon_{is}) = 0,\ t \neq s$ and $\text{Cov}(\epsilon_{it}, \alpha_i) = 0,\ t = 1, \dots, T$.

The SEM approach to modelling time-invariant unobserved heterogeneity as a latent variable in dynamic models relies on our ability to decompose the correlations observed between columns of the wide-format outcome into a part that is due to unobserved time-invariant factors, and the autoregressive component. Notice that this means we need at least three observed timepoints for the outcome to do so properly. Consider a simplified model with just two observed timepoints where we are trying to estimate separately the autoregressive effect and the stable unobserved factors:

$$y_{i1} = \alpha_i + \epsilon_{i1} \tag{8}$$
$$y_{i2} = \rho y_{i1} + \alpha_i + \epsilon_{i2} \tag{9}$$

For this, we would need to estimate four parameters, $\theta = (\phi_\alpha, \psi_1, \psi_2, \rho)$, where $\phi_\alpha = \text{Var}(\alpha)$ and $\psi_t = \text{Var}(\epsilon_t)$. But we have just three pieces of observed information, $\text{Var}(y_1), \text{Var}(y_2), \text{Cov}(y_1, y_2)$. This would mean our model is underidentified with -1 degrees of freedom.

Furthermore, (implictly) fixing the factor loading of $\alpha \to y_1$ to 1.0 assumes the process under investigation is occurring within a vacuum, and that the cumulative effects of the previous realizations of the outcome have no bearing on the initial observation (Ou et al. 2017 call this the 'history independent endogenous' model); an unrealistic assumption in most cases. Another option is to assume the process under investigation is 'ongoing endogenous' (Ou et al. 2017) and fix the initial factor loading to a specific value (see, for example Andersen 2021), but this does not fix the problem of an underidentified model, either. In fact, in order to avoid making assumptions about the process before the observation period began, it is generally advised to treat the initial observations as 'predetermined', i.e., enter it into the model as an exogenous variable, allowing it to covary freely with the individual effects (Allison, Williams, and Moral-Benito 2017; Andersen 2021; Ou et al. 2017). This is the most robust choice but uses up another degree of freedom in the process.

If we have three measures of the outcome, however, we can treat the initial observation as predetermined and generally have enough degrees of freedom to properly estimate the

autoregressive effect as well as the variance of the time-invariant factors affecting the outcome. Doing so requires we assume the effect of the unobserved stable factors is constant over time (this assumption is reflected in the usual practice of fixing the factor loadings from the latent variable to the outcome to 1.0 at each point in time (Andersen 2022; K. A. Bollen and Brand 2010)), and that the autoregressive effect is constant over time, as well (in other words, there is no moderation by period effects; a plausible assumption in many sociological applications based on observational data). For example, a simple model with three measures of the outcome and the above-mentioned assumptions is just-identified with zero degrees of freedom. The longer the observed timeframe, and the more observed covariates we include in the model, the more degrees of freedom we have to be able to relax these and other assumptions, if necessary.

More on the method for accounting for stable unobserved characteristics in panel models in SEM has been outlined elsewhere (e.g., Allison 2011; Andersen 2022; K. A. Bollen and Brand 2010; Teachman et al. 2001), so we will not describe it in detail here. Instead, we show that by explicitly modeling the time-invariant unobserved heterogeneity, the causal effect of interest can be estimated without bias.

5 A Simulated Example

We construct a simulated dataset to demonstrate the use of SEM in accounting for time-varying and invariant confounders. We generate three measures of the outcome to be able to treat the initial observation as predetermined.[4] Figure 3 shows the simulated DGP, where we focus on three measures of the outcome over time, $Y_0 - Y_2$, along with the causal variable, D, and account for the unobserved variables U and V indirectly.

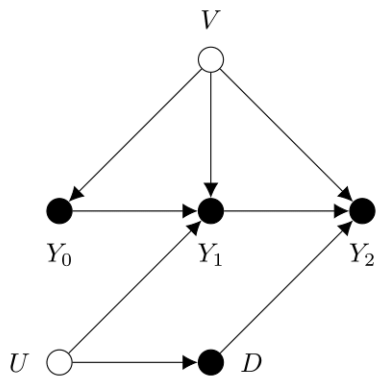


Figure 3: Simulated DGP

[4] We need to treat the initial observation as predetermined in this simulated example because the process will not yet have reached equilibrium (unless the autoregressive effect were to be diminishly small). This means that the ongoing endogenous and history independent endogenous specifications will be inappropriate, see Ou et al. (2017).

We look at linear additive effects. The exogenous variables, as well as the errors, are standard normal. The effect sizes were chosen arbitrarily and can be seen in the code below. The main causal effect of interest, $D \to Y2$ is set to the value 0.4.

```r
# Set seed
set.seed(45678)

# Load packages
library(lavaan)
library(dplyr)

# Set large sample size
n <- 1000L

rho   = 0.3 # Autoregressive effect
gamma = 0.6 # Effect U -> D
delta = 0.5 # Effect U -> Y6
beta  = 0.4 # Causal effect of interest

# Time-invariant unobserved heterogeneity
V = rnorm(n, 0, 1)

# Simulate initial realization of outcome
Y0 = 1 * V + rnorm(n, 0, 1)

# Time-varying confounder
U = rnorm(n, 0, 1)

# Causal variable
D = gamma * U + rnorm(n, 0, 1)

# Remaining realizations of outcome
Y1 = delta * U + rho * Y0 + 1 * V + rnorm(n, 0, 1)
Y2 = beta  * D + rho * Y1 + 1 * V + rnorm(n, 0, 1)

# Put into dataframe
df = data.frame(Y0, Y1, Y2,
                D, V, U)
```

Obviously, if both U and V were observed, then we could estimate the model without bias. We can estimate the entire model simultaneously in SEM using the `lavaan` package in `R`.

```r
m1 = "
  Y2 ~ beta*D + rho*Y1 + V
  Y1 ~ delta*U + V
  D  ~ gamma*U
```

```
"
m1.fit = sem(model = m1, data = df, estimator = "ML") %>%
  summary()
```

```
lavaan 0.6-11 ended normally after 1 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         9

  Number of observations                          1000

Model Test User Model:

  Test statistic                                 0.742
  Degrees of freedom                                 3
  P-value (Chi-square)                           0.863

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y2 ~
    D       (beta)    0.396    0.027   14.608    0.000
    Y1       (rho)    0.304    0.028   10.929    0.000
    V                 1.007    0.048   20.934    0.000
  Y1 ~
    U       (delt)    0.431    0.034   12.747    0.000
    V                 1.321    0.033   39.458    0.000
  D ~
    U       (gamm)    0.572    0.033   17.453    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Y2                0.952    0.043   22.361    0.000
   .Y1                1.096    0.049   22.361    0.000
   .D                 1.026    0.046   22.361    0.000
```

Here, the causal effect of interest (labeled `beta`) is unbiased at about 0.396 (the slight discrepancy is due to sampling error).

Now, to see the point Morgan and Winship (2014) were making, let us assume V is unobserved. Including the lagged dependent variable will close the back-door path over U but the bias will still be present because V is unobserved.

```
m2 = "
  Y2 ~ beta*D + rho*Y1
  Y1 ~ delta*U
  D  ~ gamma*U
"
m2.fit = sem(model = m2, data = df, estimator = "ML") %>%
  summary()
```

```
lavaan 0.6-11 ended normally after 2 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         7

  Number of observations                          1000

Model Test User Model:

  Test statistic                                22.891
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.000

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y2 ~
    D       (beta)    0.317    0.032    9.850    0.000
    Y1       (rho)    0.745    0.022   34.444    0.000
  Y1 ~
    U       (delt)    0.412    0.054    7.611    0.000
  D ~
    U       (gamm)    0.572    0.033   17.453    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Y2                1.370    0.061   22.361    0.000
   .Y1                2.802    0.125   22.361    0.000
```

```
.D                1.026    0.046   22.361    0.000
```

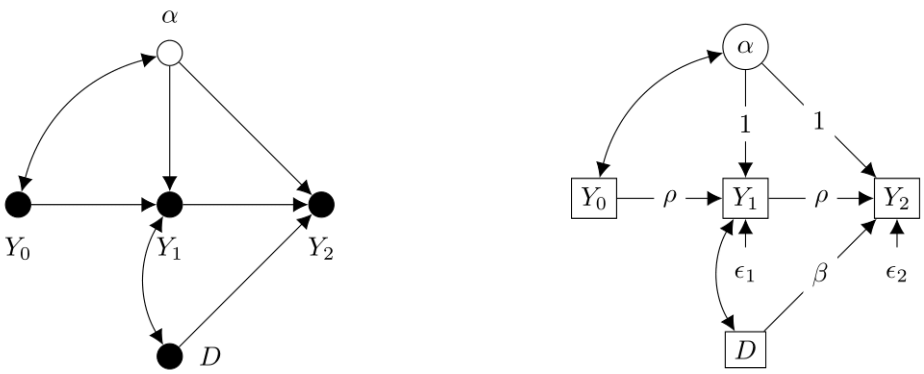
The effect of D on Y_2 is biased in this model with an estimated coefficient of 0.317.

In SEM though, V is explicitly accounted for as a latent variable. It represents the correlation of the outcome with itself over time, over and above the lagged causal variable and the autoregressive effect. In the following model, both U and V are still treated as unobserved.

We specify a latent variable to account for time-invariant unobserved heterogeneity with `alpha =~ 1*Y1 + 1*Y2` and regress `Y2` on both `D` and `Y1`. We include `Y0` in the model as an exogenous variable, allowing it to covary freely with the stable factors represented by `alpha`. Allowing the initial observation to covary freely with the stable factors accounts for situations in which the outcome is not yet at equilibrium, see Andersen (2021) and it is the common approach in dynamic models (e.g., Allison, Williams, and Moral-Benito 2017).

Note that we use the name `alpha` here instead of `V` because `lavaan` will usually return an error if one of the names of the latent variables overlaps with the name of one of the observed variables.

Finally, since `U` is now also unobserved, we allow `Y1` and `D` to covary to account for this common cause. Figure 4 shows the SEM approach graphically, both as a DAG (Figure 4a) and in the style of a traditional SEM path model (Figure 4b).



(a) DAG

(b) SEM path model

Figure 4: Modelling strategy in SEM (from m3.fit)

```
m3 = "
  # Individual effects to account for V
  alpha =~ 1*Y1 + 1*Y2
  # Regressions
  Y2 ~ beta*D + rho*Y1
  Y1 ~ rho*Y0
  # Allow initial outcome to correlate with unit effects
  alpha ~~ Y0
  # Account for U, common cause of Y6 and D
```

```
  D ~~ Y1
"
m3.fit = sem(model = m3, data = df, estimator = "ML") %>%
  summary()
```

```
lavaan 0.6-11 ended normally after 34 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        10
  Number of equality constraints                     1

  Number of observations                          1000

Model Test User Model:

  Test statistic                                 0.151
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.697

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  alpha =~
    Y1                1.000
    Y2                1.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y2 ~
    D       (beta)    0.378    0.042    9.006    0.000
    Y1       (rho)    0.343    0.183    1.879    0.060
  Y1 ~
    Y0       (rho)    0.343    0.183    1.879    0.060

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  alpha ~~
    Y0                0.931    0.328    2.839    0.005
 .Y1 ~~
```

```
    D                 0.196    0.048    4.126    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Y1                1.236    0.306    4.040    0.000
   .Y2                0.982    0.157    6.260    0.000
    D                 1.338    0.060   22.361    0.000
    Y0                1.965    0.088   22.361    0.000
    alpha             0.863    0.607    1.422    0.155
```

Now, the estimate of the causal effect is relatively close to the true parameter at 0.378. And this is not a fluke: we can draw many samples to show that the estimated coefficient is approximately equal to the causal effect.

```r
sim_func = function(rho = 0.3, beta = 0.4, gamma = 0.6, delta = 0.5) {

  # Set large sample size
  n = 1000L

  # Time-invariant unobserved heterogeneity
  V = rnorm(n, 0, 1)

  # Simulate initial realization of outcome
  Y0 = 1 * V + rnorm(n, 0, 1)

  # Time-varying confounder
  U = rnorm(n, 0, 1)

  # Causal variable
  D = gamma * U + rnorm(n, 0, 1)

  # Remaining realizations of outcome
  Y1 = delta * U + rho * Y0 + 1 * V + rnorm(n, 0, 1)
  Y2 = beta  * D + rho * Y1 + 1 * V + rnorm(n, 0, 1)

  # Put into dataframe
  df = data.frame(Y0, Y1, Y2,
                  D, V, U)

  # Fit the model
  mx = "
    # Individual effects to account for V
    alpha =~ 1*Y1 + 1*Y2
    # Regressions
    Y2 ~ beta*D + rho*Y1
    Y1 ~ rho*Y0
```

```
    # Allow initial outcome to correlate with unit effects
    alpha ~~ Y0
    # Account for U, common cause of Y6 and D
    D ~~ Y1
  "
  mx.fit = sem(model = mx, data = df, estimator = "ML")

  # Get estimate of beta
  est = lavInspect(mx.fit, "list") %>%
    filter(op == "~") %>%
    filter(label == "beta") %>%
    select(est) %>%
    as.numeric()

  # Return estimate of beta
  return(est)
}

res = replicate(n = 10000L, expr = sim_func())
```

The estimated coefficient is unbiased

```
mean(res); median(res)
```

```
[1] 0.3956767
```

```
[1] 0.3963745
```

```
sd(res)
```

```
[1] 0.05102995
```

and approximately normally distributed around the true structural coefficient of 0.4, see Figure 5.

```
hist(res, main = NULL, xlab = NULL, breaks = 30)
```

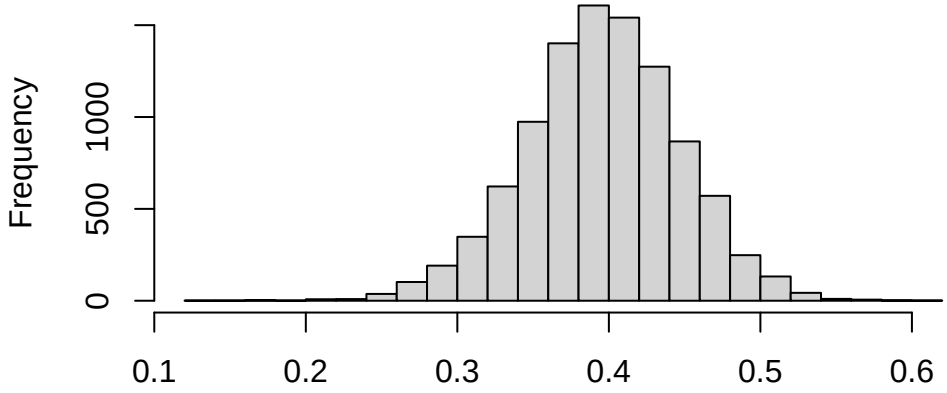


Figure 5: Histogram, sampling distribution of res

Readers will notice the distribution is somewhat skewed and has a large variance. Even if the estimated coefficient is unbiased, the statistical properties are improved if the underlying data has had more time to reach equilibrium. In other words, the way we simulated the data above means the variances of the outcome over time, as well as the covariances of adjacent columns of the outcome are changing rapidly near the beginning of the process (Andersen 2021). If we first allow the process to reach equilibrium (variances, covariances, as well as means are fairly constant over time), then the sampling distribution becomes more normal, and the variance is reduced. We can show this by first simulating a 'spin-up' phase for the outcome, and then focusing in on the causal model at a later point in time.

```
sim_func2 = function(rho = 0.3, beta = 0.4, gamma = 0.6, delta = 0.5) {

  # Set large sample size
  n <- 1000L

  # Time-invariant unobserved heterogeneity
  V = rnorm(n, 0, 1)

  # Simulate spin-up phase to allow Yt to reach equilibrium
  Y0 = 1 * V + rnorm(n, 0, 1)
  Y1 = rho * Y0 + 1 * V + rnorm(n, 0, 1)
  Y2 = rho * Y1 + 1 * V + rnorm(n, 0, 1)
  Y3 = rho * Y2 + 1 * V + rnorm(n, 0, 1)
  Y4 = rho * Y3 + 1 * V + rnorm(n, 0, 1)
```

```r
  Y5 = rho * Y4 + 1 * V + rnorm(n, 0, 1)

  # Time-varying confounder
  U = rnorm(n, 0, 1)

  # Causal variable
  D = gamma * U + rnorm(n, 0, 1)

  # Focus on effect D -> Y7, holding Y6 constant
  Y6 = delta * U + rho * Y5 + 1 * V + rnorm(n, 0, 1)
  Y7 = beta * D + rho * Y6 + 1 * V + rnorm(n, 0, 1)

  # Put into dataframe
  df = data.frame(Y0, Y1, Y2, Y3, Y4, Y5, Y6, Y7, D, V, U)

  # Fit the model
  mx = "
    # Individual effects to account for V
    alpha =~ 1*Y6 + 1*Y7
    # Regressions
    Y7 ~ beta*D + rho*Y6
    Y6 ~ rho*Y5
    # Allow initial outcome to correlate with unit effects
    alpha ~~ Y5
    # Account for U, common cause of Y6 and D
    D ~~ Y6
    "
  mx.fit = sem(model = mx, data = df, estimator = "ML")

  # Get estimate of beta
  est = lavInspect(mx.fit, "list") %>%
    filter(op == "~") %>%
    filter(label == "beta") %>%
    select(est) %>%
    as.numeric()

  # Return estimate of beta
  return(est)
}

res2 = replicate(n = 10000L, expr = sim_func2())
```

The estimate of the causal effect is still unbiased

```r
mean(res2); median(res2)
```

```
[1] 0.3988873
```

```
[1] 0.399189
```

```
sd(res2)
```

```
[1] 0.03995896
```

but now the sampling distribution better approximates the normal distribution and the spread of the estimated coefficients is tighter, see Figure 6.

```
hist(res2, main = NULL, xlab = NULL, breaks = 30)
```

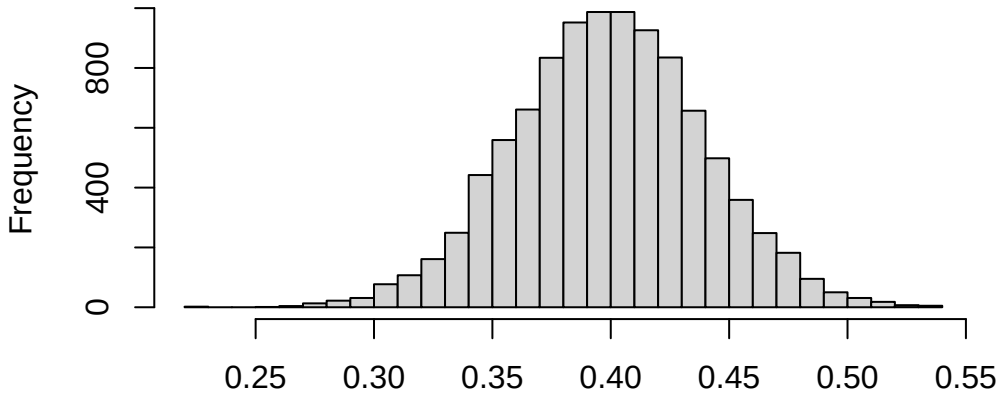


Figure 6: Histogram, sampling distribution of res2

6 Conclusion

The goal of this article was to renew the reader's confidence in the use of LDVs in panel models. As we discussed, there are good reasons to consider doing so. From a theoretical standpoint, LDVs can capture inertial effects (Wooldridge 2012, 313) where there is expected to be carry-over of the outcome at one point in time to the next (Keele and Kelly 2006). The inclusion of the LDV also gives the main coefficient of interest a desirably intuitive interpretation: a comparison of outcomes between hypothetical units that displayed the same outcome in the previous period, but whose values on the causal variable differ. But

perhaps most importantly, the LDV can be effective at closing back-door paths due to unobserved time-varying confounders.

The practice of including the LDV is often criticized and for good reason. In the presence of time-invariant unobserved heterogeneity, the LDV acts as a collider and opens up an unintended back-door path, thus biasing the estimate of the causal effect. But leaving the LDV out means the onus is on the researcher to measure all the potential time-varying confounders, so you're often "damned if you do, damned if you don't" (Cinelli, Forney, and Pearl 2022).

The SEM approach gets around this specific criticism of the LDV by explicitly accounting for time-invariant unobserved heterogeneity. This blocks both back-door paths, across (certain) time-varying confounders and time-invariant ones. Thus, the wide use of cross-lagged and other panel models in SEM that account for LDVs is arguably justified.

LDVs are not a silver bullet, however. The researcher's qualitative hypotheses must hold, as always. The LDV will stop confounding if the simplified DAG in Figure 2 (or something equivalent, such as Figure 3) is the true DGP. If the unobserved time-varying confounders affect the current outcome over and above the mediated path over the lagged version, then another unblocked path is opened up. And other assumptions, like the appropriateness of a linear model, must be scrutinized in SEM just as in any other methodology (Kenneth A. Bollen and Pearl 2013).

Finally, we did not look at other potential approaches to LDVs, such as the Arellano-Bond (AB) differenced model (which is perhaps the most promising approach outside of SEM), which tries to use lags even further back from the current outcome as instruments (Brüderl and Ludwig 2014). In a series of simulations, Leszczensky and Wolbring (2019) showed that AB and SEM both performed well under a wide variety of underlying DGPs and assumptions.

Instead, this paper chose to draw attention to an apparently overlooked aspect. At this very moment, a researcher at a sociological conference is likely being alerted to the fact that the LDV is a collider and that their SEM and its results are biased. We believe this should not be a blanket criticism, and that SEM provides a flexible framework for modeling various dynamic processes that suit researchers' qualitative hypotheses.

7 `lavaan` Code

```
# Set seed
set.seed(45678)

# Load packages
library(lavaan)
library(dplyr)


# Generate data ----------------------------------------------------------
```

```r

# Set large sample size
n <- 1000L

rho   = 0.3 # Autoregressive effect
gamma = 0.6 # Effect U -> D
delta = 0.5 # Effect U -> Y6
beta  = 0.4 # Causal effect of interest

# Time-invariant unobserved heterogeneity
V = rnorm(n, 0, 1)

# Simulate initial realization of outcome
Y0 = 1 * V + rnorm(n, 0, 1)

# Time-varying confounder
U = rnorm(n, 0, 1)

# Causal variable
D = gamma * U + rnorm(n, 0, 1)

# Remaining realizations of outcome
Y1 = delta * U + rho * Y0 + 1 * V + rnorm(n, 0, 1)
Y2 = beta  * D + rho * Y1 + 1 * V + rnorm(n, 0, 1)

# Put into dataframe
df = data.frame(Y0, Y1, Y2,
                D, V, U)


# Model 1: All observed variables ----------------------------------------

m1 = "
  Y2 ~ beta*D + rho*Y1 + V
  Y1 ~ delta*U + V
  D  ~ gamma*U
"
m1.fit = sem(model = m1, data = df, estimator = "ML") %>%
  summary()


# Model 2: Ignoring time-invariant unob. hetero.  ------------------------

m2 = "
  Y2 ~ beta*D + rho*Y1
  Y1 ~ delta*U
```

```
  D  ~ gamma*U
"
m2.fit = sem(model = m2, data = df, estimator = "ML") %>%
  summary()


# Model 3: Time-invariant unob. hetero. as latent variable ----------------

m3 = "
  # Individual effects to account for V
  alpha =~ 1*Y1 + 1*Y2
  # Regressions
  Y2 ~ beta*D + rho*Y1
  Y1 ~ rho*Y0
  # Allow initial outcome to correlate with unit effects
  alpha ~~ Y0
  # Account for U, common cause of Y6 and D
  D ~~ Y1
"
m3.fit = sem(model = m3, data = df, estimator = "ML") %>%
  summary()


# Repeated sampling ------------------------------------------------------

sim_func = function(rho = 0.3, beta = 0.4, gamma = 0.6, delta = 0.5) {

  # Set large sample size
  n = 1000L

  # Time-invariant unobserved heterogeneity
  V = rnorm(n, 0, 1)

  # Simulate initial realization of outcome
  Y0 = 1 * V + rnorm(n, 0, 1)

  # Time-varying confounder
  U = rnorm(n, 0, 1)

  # Causal variable
  D = gamma * U + rnorm(n, 0, 1)

  # Remaining realizations of outcome
  Y1 = delta * U + rho * Y0 + 1 * V + rnorm(n, 0, 1)
  Y2 = beta  * D + rho * Y1 + 1 * V + rnorm(n, 0, 1)

```

```r
  # Put into dataframe
  df = data.frame(Y0, Y1, Y2,
                  D, V, U)

  # Fit the model
  mx = "
    # Individual effects to account for V
    alpha =~ 1*Y1 + 1*Y2
    # Regressions
    Y2 ~ beta*D + rho*Y1
    Y1 ~ rho*Y0
    # Allow initial outcome to correlate with unit effects
    alpha ~~ Y0
    # Account for U, common cause of Y6 and D
    D ~~ Y1
  "
  mx.fit = sem(model = mx, data = df, estimator = "ML")

  # Get estimate of beta
  est = lavInspect(mx.fit, "list") %>%
    filter(op == "~") %>%
    filter(label == "beta") %>%
    select(est) %>%
    as.numeric()

  # Return estimate of beta
  return(est)
}

res = replicate(n = 10000L, expr = sim_func())


# Repeated sampling, data at equilibrium  ---------------------------------

sim_func2 = function(rho = 0.3, beta = 0.4, gamma = 0.6, delta = 0.5) {

  # Set large sample size
  n <- 1000L

  # Time-invariant unobserved heterogeneity
  V = rnorm(n, 0, 1)

  # Simulate spin-up phase to allow Yt to reach equilibrium
  Y0 = 1 * V + rnorm(n, 0, 1)
  Y1 = rho * Y0 + 1 * V + rnorm(n, 0, 1)
  Y2 = rho * Y1 + 1 * V + rnorm(n, 0, 1)
```

```r
  Y3 = rho * Y2 + 1 * V + rnorm(n, 0, 1)
  Y4 = rho * Y3 + 1 * V + rnorm(n, 0, 1)
  Y5 = rho * Y4 + 1 * V + rnorm(n, 0, 1)

  # Time-varying confounder
  U = rnorm(n, 0, 1)

  # Causal variable
  D = gamma * U + rnorm(n, 0, 1)

  # Focus on effect D -> Y7, holding Y6 constant
  Y6 = delta * U + rho * Y5 + 1 * V + rnorm(n, 0, 1)
  Y7 = beta * D + rho * Y6 + 1 * V + rnorm(n, 0, 1)

  # Put into dataframe
  df = data.frame(Y0, Y1, Y2, Y3, Y4, Y5, Y6, Y7, D, V, U)

  # Fit the model
  mx = "
    # Individual effects to account for V
    alpha =~ 1*Y6 + 1*Y7
    # Regressions
    Y7 ~ beta*D + rho*Y6
    Y6 ~ rho*Y5
    # Allow initial outcome to correlate with unit effects
    alpha ~~ Y5
    # Account for U, common cause of Y6 and D
    D ~~ Y6
    "
  mx.fit = sem(model = mx, data = df, estimator = "ML")

  # Get estimate of beta
  est = lavInspect(mx.fit, "list") %>%
    filter(op == "~") %>%
    filter(label == "beta") %>%
    select(est) %>%
    as.numeric()

  # Return estimate of beta
  return(est)
}

res2 = replicate(n = 10000L, expr = sim_func2())
```

References

Allison, Paul D. 2011. *Fixed Effects Regression Models*. Los Angeles, London: Sage Publications.

Allison, Paul D., Richard Williams, and Enrique Moral-Benito. 2017. "Maximum Likelihood for Cross-Lagged Panel Models with Fixed Effects." *Socius: Sociological Research for a Dynamic World* 3 (January): 237802311771057. https://doi.org/10.1177/2378023117710578.

Andersen, Henrik Kenneth. 2021. "Equivalent Approaches to Dealing with Unobserved Heterogeneity in Cross-Lagged Panel Models? Investigating the Benefits and Drawbacks of the Latent Curve Model with Structured Residuals and the Random Intercept Cross-Lagged Panel Model." *Psychological Methods* Advanced online publication. https://doi.org/10.1037/met0000285.

———. 2022. "A Closer Look at Random and Fixed Effects Panel Regression in Structural Equation Modeling Using Lavaan." *Structural Equation Modeling: A Multidisciplinary Journal* 29 (3): 476–86. https://doi.org/10.1080/10705511.2021.1963255.

Bollen, K. A., and J. E. Brand. 2010. "A General Panel Model with Random and Fixed Effects: A Structural Equations Approach." *Social Forces* 89 (1): 1–34. https://doi.org/10.1353/sof.2010.0072.

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

Bollen, Kenneth A., and Judea Pearl. 2013. "Eight Myths about Causality and Structural Equation Models." In *Handbook of Causal Analysis for Social Research*, edited by Stephen L. Morgan, 301–28. Dordrecht: Springer Science+Business Media.

Brüderl, Josef, and Volker Ludwig. 2014. "Fixed-Effects Panel Regression." In, 327–58. SAGE Publications Ltd. https://doi.org/10.4135/9781446288146.n15.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2022. "A Crash Course in Good and Bad Controls." June 3, 2022. https://www.researchgate.net/publication/340082755_A_Crash_Course_in_Good_and_Bad_Controls.

Coleman, James S., Thomas Hoffer, and Sally Kilgore. 1982. *High School Achievement: Public, Catholic, and Private Schools Compared*. New York: Basic Books.

Collischon, Matthias, and Andreas Eberl. 2020. "Let's Talk About Fixed Effects: Let's Talk About All the Good Things and the Bad Things." *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie* 72 (2): 289–99. https://doi.org/10.1007/s11577-020-00699-8.

Curran, Patrick J., and Kenneth A. Bollen. 2001. "The Best of Both Worlds: Combining Autoregressive and Latent Curve Models." In, 107–35. American Psychological Association. https://doi.org/10.1037/10409-004.

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

Dafoe, Allan. 2014. "Prescriptions for Temporal Dependence: First Do No Harm." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2422871.

———. 2015. "Nonparametric Identification of Causal Effects Under Temporal Dependence." *Sociological Methods & Research* 47 (2): 136–68. https://doi.org/10.1177/0049124115613784.

Foster, E. Michael. 2010. "Causal Inference and Developmental Psychology." *Developmental Psychology* 46 (6): 1454–80. https://doi.org/10.1037/a0020204.

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

Keele, Luke, and Nathan J. Kelly. 2006. "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables." *Political Analysis* 14 (2): 186–205. https://doi.org/10.1093/pan/mpj006.

Kühnel, Steffen, and Anja Mays. 2018. "Probleme von Cross-Lagged Panelmodellen Zur Analyse Gegenseitiger Beeinflussung von Einstellung Und Verhalten." In, 359–86. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-16348-8_15.

Leszczensky, Lars, and Tobias Wolbring. 2019. "How to Deal With Reverse Causality Using Panel Data? Recommendations for Researchers Based on a Simulation Study." *Sociological Methods & Research* 51 (2): 837–65. https://doi.org/10.1177/0049124119882473.

Moral-Benito, Enrique, Paul Allison, and Richard Williams. 2018. "Dynamic Panel Data Modelling Using Maximum Likelihood: An Alternative to Arellano-Bond." *Applied Economics* 51 (20): 2221–32. https://doi.org/10.1080/00036846.2018.1540854.

Morgan, Stephen L., and Christopher Winship. 2014. "Counterfactuals and Causal Inference." https://doi.org/10.1017/cbo9781107587991.

Mouw, Ted. 2006. "Estimating the Causal Effect of Social Capital: A Review of Recent Research." *Annual Review of Sociology* 32 (1): 79–102. https://doi.org/10.1146/annurev.soc.32.061604.123150.

Ou, Lu, Sy-Miin Chow, Linying Ji, and Peter C. M. Molenaar. 2017. "(Re)evaluating the Implications of the Autoregressive Latent Trajectory Model Through Likelihood Ratio Tests of Its Initial Conditions." *Multivariate Behavioral Research* 52 (2): 178–99. https://doi.org/10.1080/00273171.2016.1259980.

Rogosa, David. 1980. "A Critique of Cross-Lagged Correlation." *Psychological Bulletin* 88 (2): 245–58. https://doi.org/10.1037/0033-2909.88.2.245.

Teachman, Jay, Greg J. Duncan, Jean Yeung, and Dan Levy. 2001. "Covariance Structure Models for Fixed and Random Effects." *Sociological Methods and Research* 30 (2): 271–88. https://doi.org/10.1177/0049124101030002005.

Walters, Glenn D. 2019. "Criminal Thinking as a Moderator of the Perceived Certainty–offending Relationship: Age Variations." *Psychology, Crime & Law* 26 (3): 267–86. https://doi.org/10.1080/1068316x.2019.1652749.

Williams, Richard, Paul D. Allison, and Enrique Moral-Benito. 2018. "Linear Dynamic Panel-Data Estimation Using Maximum Likelihood and Structural Equation Modeling." *The Stata Journal: Promoting Communications on Statistics and Stata* 18 (2): 293–326. https://doi.org/10.1177/1536867x1801800201.

Wooldridge, Jeffrey M. 2012. *Introductory Econometrics: A Modern Approach, 5th Edition.* Mason: South-Western.

Zyphur, Michael J., Paul D. Allison, Louis Tay, Manuel C. Voelkle, Kristopher J. Preacher, Zhen Zhang, Ellen L. Hamaker, et al. 2020. "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.

Zyphur, Michael J., Manuel C. Voelkle, Louis Tay, Paul D. Allison, Kristopher J. Preacher, Zhen Zhang, Ellen L. Hamaker, et al. 2020. "From Data to Causes II: Comparing Approaches to Panel Data Analysis." *Organizational Research Methods* 23 (4): 688–716. https://doi.org/10.1177/1094428119847280.