Longitudinal Data Analysis

White Rose Social Sciences Doctoral Training Partnership

Thiago R. Oliveira

Lecturer in Criminology, University of Manchester

8 April 2025, University of Leeds

Housekeeping and introduction to longitudinal data

About me

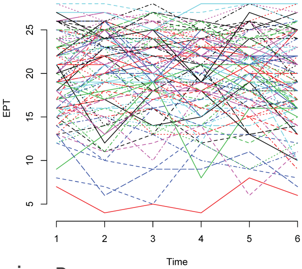
- Hi, I'm Thiago!
- Lecturer, University of Manchester
  PhD in Social Research Methods
- I am a quantitative criminologist
- Into longitudinal data analysis, causal inference, and other hard drugs...

About the course

This is a course about **longitudinal data analysis**

Specifically, we are going to cover two topics:

- How to handle **reverse causality** and **reciprocal relationships**
  Keyword: **cross-lagged panel models**
- How to leverage longitudinal data to **make causal conclusions**
  Keyword: **Difference-in-differences**



We will adopt a hands-on approach using R:

Morning session:

- 10:00—11:15: lecture on cross-lagged panel models
- 11:15—12:05: lab session: CLPMs in practice
- 12:05—12:30: Recap and wrap-up

Afternoon session:

- 1:30—2:45: lecture on causal inference with panel data
- 2:45—3:35: lab session: DiD in practice
- 3:35—4:00: Recap and wrap-up

Idea of longitudinal data

What is **longitudinal data**?

- ⇝ How is it different from **cross-sectional data**?
- ⇝ How is it different from **time series data**?

⇒ **Repeated observations of the same units over time**

- ⇝ Intuitively: Large N, small T

What can we use longitudinal data for?

General intuition: modelling **change over time**

- ⇝ Modelling *change over time*, controlling for variables measured at different points in time (LDV, CLPM, RI-CLPM, DPM, . . .)
- ⇝ Modelling *effects on change*, focusing on within-unit change over time only (FE, TWFE, DiD, M-DiD, . . .)
- ⇝ Modelling *individual trajectories*, treating TIME as the main independent variable (GCM, LGCM, LCSM, GMM, GBTM, . . .)
- ⇝ Modelling *time to event*, focusing on distal outcomes or probability of event happening (survival models, event-history, MSM, . . .)

Our focus:

⇒ **Cross-lagged panel models** for *reverse causality* and *reciprocal relationships*

⇒ **Difference-in-differences** for *causal inference* with panel data

Part I: handling reverse causality and reciprocal effects with cross-lagged panel models

Reverse causality?

When conducting empirical research, we sometimes want to examine the effect of X on Y but are afraid that Y might also affect X

- $\rightsquigarrow$ Effect of policing (X) on crime (Y)
  - More police might be deployed in high-crime areas
- $\rightsquigarrow$ Effect of education (X) on income (Y)
  - People from higher-income families may be more likely to pursue education
- $\rightsquigarrow$ Effect of social media (X) on mental health (Y)
  - People with poor mental health may use social media more

We may be interested in:

$\rightarrow$ *Controlling* for reverse causality

$\rightarrow$ *Discovering* the direction of the association

$\rightarrow$ *Discovering* a reciprocal relationship*

Reverse causality: a motivating example

Untangling the Relationship Between Fear of Crime and Perceptions of Disorder: Evidence from a Longitudinal Study of Young People in England and Wales

Ian Brunton-Smith ✉

The British Journal of Criminology, Volume 51, Issue 6, November 2011, Pages 885–899,
https://doi.org/10.1093/bjc/azr064
Published: 19 August 2011

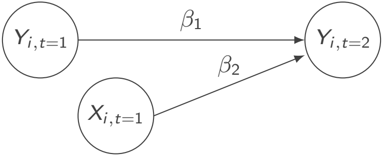
- Brunton-Smith (2011) wanted to study the relationship between fear of crime and perceptions of disorder
  - H_1 : fear of crime $\longrightarrow$ perceptions of disorder
  - H_2 : perceptions of disorder $\longrightarrow$ fear of crime
- Interesting question: which one is causing which?
  - Can we use empirical data to answer it?

Reverse causality: a motivating example

⇝ Panel data allows us to model changes in various ways

- e.g. including **autoregressive** parameters

$$Y_{it} = \alpha + \beta_1 \cdot Y_{i,t-1} + \beta_2 \cdot X_{i,t} + \varepsilon$$



⇝ Because of the inclusion of the autoregressive parameter β_1:

- β_2 represents the 'effect' of X on changes in Y

Reverse causality: a motivating example

- Good! So we can assess the association between perceptions of disorder and changes in fear of crime:

$$\begin{tikzcd}
\textit{Fear of crime at } T = 1 \arrow[rr] & & \textit{Fear of crime at } T = 2 \\
& \textit{Perceptions of disorder} \arrow[ur] &
\end{tikzcd}$$

- But perceptions of disorder also vary in time, so we can **also** assess the association between fear of crime and changes in disorder perceptions:

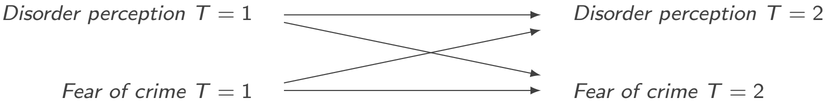
$$\begin{tikzcd}
\textit{Disorder perception } T = 1 \arrow[rr] & & \textit{Disorder perception } T = 2 \\
& \textit{Fear of crime} \arrow[ur] &
\end{tikzcd}$$

The cross-lagged panel model

The cross-lagged panel model

- What if we estimate both simultaneously?

$$Y_{i,t} = \alpha + \beta_1 \cdot Y_{i,t-1} + \beta_2 \cdot X_{i,t-1} + \varepsilon$$
$$X_{i,t} = \mu + \beta_3 \cdot X_{i,t-1} + \beta_4 \cdot Y_{i,t-1} + \upsilon$$



- CLPM allows us to model reciprocal relationships
- Estimated using Structural Equation Modelling (SEM) framework
- Temporal order: very useful in social research

The cross-lagged panel model

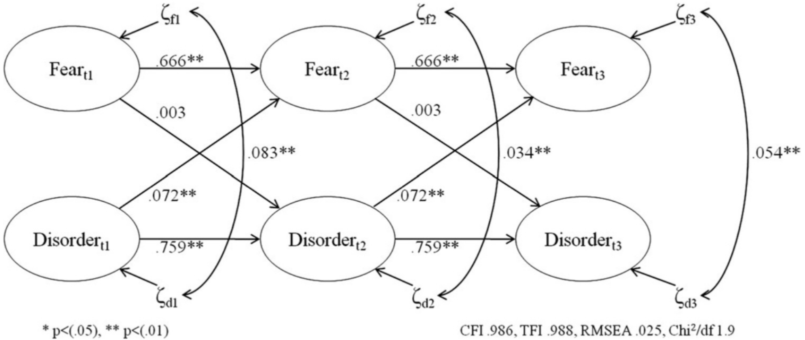


FIG. 2 Cross-lagged panel model (age 16–25).

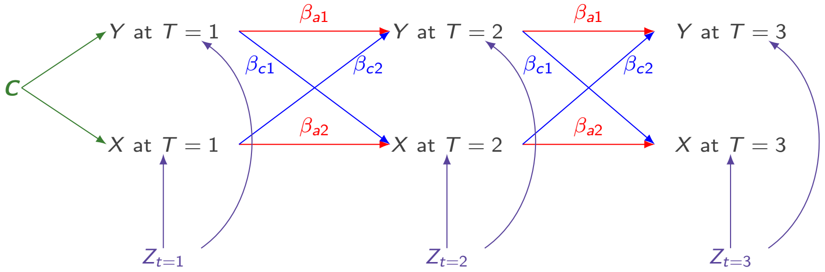
⇝ Brunton-Smith (2011) estimated a CLPM

- Changes in perceptions of disorder lead to changes in fear of crime
- Changes in fear of crime **do not** lead to changes in perceptions of disorder

The cross-lagged panel model

Some technical details. . .

- **Autoregressive** and **cross-lagged** parameters are *conventionally* constrained to equality
- Time-constant covariates are included as predictors of both initial states (i.e., $X_{i,t=1}$ and $Y_{i,t=1}$)
- Time-varying covariates are included as predictors of each X_{it} and each Y_{it}



A critique of the cross-lagged panel model

A critique of the cross-lagged panel model

Psychological Methods
2015, Vol. 20, No. 1, 102–116

1082-989X/15/$12.00 http://dx.d

A Critique of the Cross-Lagged Panel Model

Ellen L. Hamaker and Rebecca M. Kuiper
Utrecht University

Raoul P. P. P. Grasman
University of Amsterdam

The cross-lagged panel model (CLPM) is believed by many to overcome the problems associated with the use of cross-lagged correlations as a way to study causal influences in longitudinal panel data. The current article, however, shows that if stability of constructs is to some extent of a trait-like, time-invariant nature, the autoregressive relationships of the CLPM fail to adequately account for this. As a result, the lagged parameters that are obtained with the CLPM do not represent the actual within-person relationships over time, and this may lead to erroneous conclusions regarding the *presence*, *predominance*, and *sign* of causal influences. In this article we present an alternative model that separates the within-person process from stable between-person differences through the inclusion of random intercepts, and we discuss how this model is related to existing structural equation models that include cross-lagged relationships. We derive the analytical relationship between the cross-lagged parameters from the CLPM and the alternative model, and use simulations to demonstrate the spurious results that may arise when using the CLPM to analyze data that include stable, trait-like individual differences. We also present a modeling strategy to avoid this pitfall and illustrate this using an empirical data set. The implications for both existing and future cross-lagged panel research are discussed.

Keywords: cross-lagged panel, reciprocal effects, longitudinal model, trait–state models, within-person dynamics

(Hamaker et al., 2015)

Socius
Volume 3, 2017

https://doi.org/10.1177/2378023117710578

SAGE journals

Original Article

Maximum Likelihood for Cross-lagged Panel Models with Fixed Effects

Paul D. Allison[1], Richard Williams[2], and Enrique Moral-Benito[3]

Abstract
Panel data make it possible both to control for unobserved confounders and allow for lagged, reciprocal causation. Trying to do both at the same time, however, leads to serious estimation difficulties. In the econometric literature, these problems have been solved by using lagged instrumental variables together with the generalized method of moments (GMM). Here we show that the same problems can be solved by maximum likelihood (ML) estimation implemented with standard software packages for structural equation modeling (SEM). Monte Carlo simulations show that the ML-SEM method is less biased and more efficient than the GMM method under a wide range of conditions. ML-SEM also makes it possible to test and relax many of the constraints that are typically embodied in dynamic panel models.

Keywords
panel data, dynamic panel model, fixed effects, cross-lagged model, generalized method of moments, GMM, Arellano-Bond, FIML, SEM, structural equation model, maximum likelihood, predetermined variable, sequentially exogenous variable, xtdpdml, instrumental variable

(Allison et al., 2017)

A critique of the cross-lagged panel model

Common issues have emerged recently

- ⇝ Unobserved stable heterogeneity (Hamaker et al., 2015; Allison et al., 2017)
- ⇝ Correct specification of temporal lags (Vaisey and Miles, 2017)
- ⇝ Low inter-temporal variation (Lanfear et al., 2020)

Unobserved stable heterogeneity

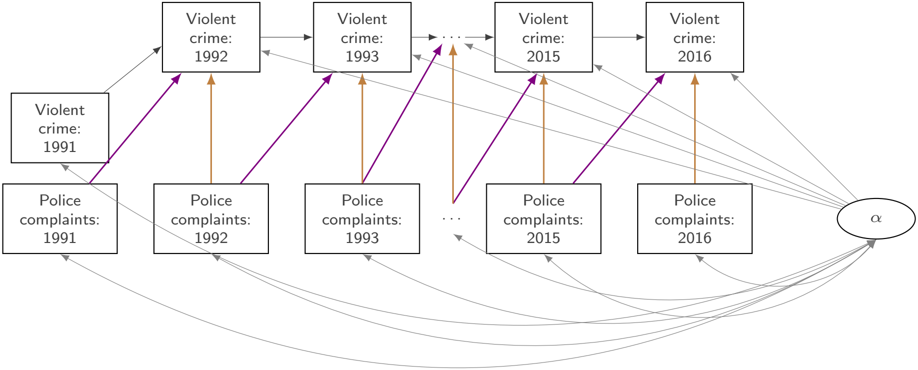
Point is relatively simple: the autoregressive parameter alone is not sufficient!

- It does not fully capture all time-constant traits
- Therefore, the model does not properly model change over time

Solution: use some recently developed robust estimator:

- Hamaker's (2015) **Random Intercepts-Cross-lagged panel model** (RI-CLPM)
  - Inspired by random effects models, it explicitly partitions the variance in *between-unit* variation and *within-unit* variation
  - Have a look at Pina-Sánchez and Brunton-Smith's (2020) paper!
- Allison et al.'s (2017) **dynamic panel model with fixed effects** (DPM)
  - Inspired by econometric models, considered the most robust approach by Vaisey and Miles (2017)

Allison et al.'s dynamic panel model



Pros: properly controls for unobserved stable heterogeneity, dismisses the threat of reverse causality, and properly models change over time

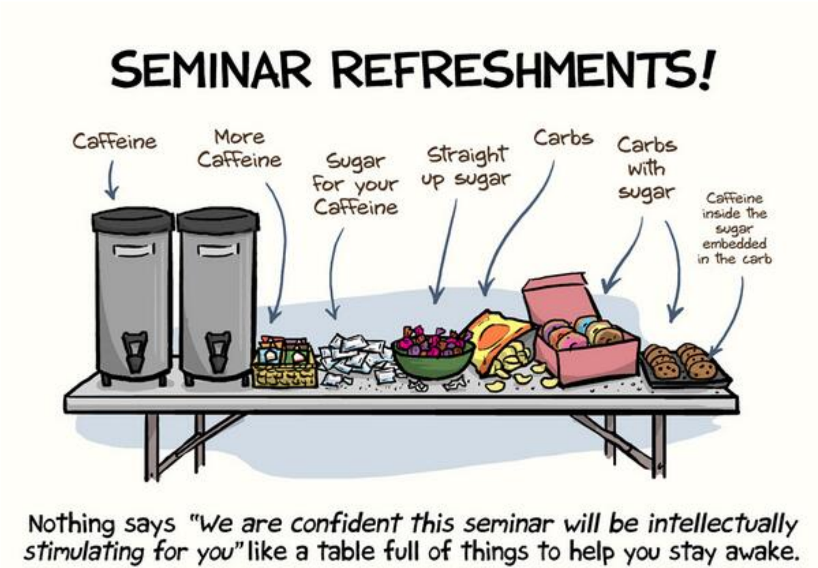
Cons: reciprocal effects cannot be simultaneously estimated, still sensitive to the correct specification of temporal lags

Summary

Summary

- Cross-lagged panel models are a powerful method that permits
  - modelling reciprocal relationship
  - establishing temporal order
  - handling reverse causality
- In general, default to more robust estimators
  - Hamaker et al.'s RI-CLPM
  - Allison et al.'s DPM
- Models are *very* sensitive to the correct specification of temporal lags
  - Not something that can be solved empirically. Think carefully about the phenomenon you are studying...
- Now let's see how to estimate those models using `R`!
  - Find the lab notes here: thiagoroliveira/1-LDA-lab.html

Coffee break



REFERENCES

Allison, P. D., R. Williams, and E. Moral-Benito (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius 3*.

Brunton-Smith, I. (2011). Untangling the relationship between fear of crime and perceptions of disorder: Evidence from a longitudinal study of young people in England and Wales. *The British Journal of Criminology 51*(6), 885–899.

Hamaker, E. L., R. M. Kuiper, and R. P. Grasman (2015). A critique of the cross-lagged panel model. *Psychological methods 20*(1), 102.

Lanfear, C. C., R. L. Matsueda, and L. R. Beach (2020). Broken windows, informal social control, and crime: Assessing causality in empirical studies. *Annual review of criminology 3*(1), 97–120.

Pina-Sánchez, J. and I. Brunton-Smith (2020). Reassessing the relationship between procedural justice and police legitimacy. *Law and Human Behavior 44*(5), 377–393.

Vaisey, S. and A. Miles (2017). What you can – and can't – do with three-wave panel data. *Sociological Methods & Research 46*(1), 44–67.