Longitudinal Data Analysis

methods@manchester summer school

Thiago R. Oliveira

Lecturer in Quantitative Criminology, University of Manchester

30/06—04/07

Housekeeping

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...

How about you?

About the course

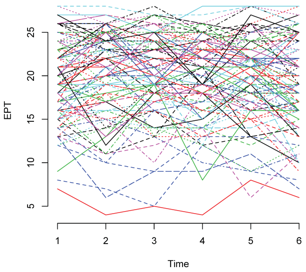
This is a course about **longitudinal data analysis**

Specifically, we are going to cover three topics:

- How to model change over time and estimate **individual trajectories**
  Keyword: **Growth curve models**

- 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 estimators**



Course structure

We will adopt a hands-on approach:

Lectures introducing theoretical concepts, followed by **lab sessions** using R applying concepts with real datasets

Morning session (9am—12:30pm)

- 9:00—10:30: Lecture

  Coffee break in The Hive
- 11:00—12:30: Lab session using R

Afternoon session (1:30—5pm)

- 1:30—3:00: Lecture

  Coffee break in The Hive
- 3:30—5:00: Lab session using R

12:30—1:30pm: Lunch in The Hive

Course outline

Course outline (short version)

- > **Monday**: Introduction to longitudinal data and R
- > **Tuesday**: Latent trajectories and growth curve models
- > **Wednesday**: Reverse causality and cross-lagged panel models
- > **Thursday**: Difference-in-differences and causal inference with panel data
- > **Friday**: Recent advancements in causal inference with panel data

Course outline (long version)

> **Monday**: Introduction to longitudinal data and R

- ⇝ Housekeeping
- ⇝ What is longitudinal data?
- ⇝ Wide vs. long datasets
- ⇝ Introduction to R
- ⇝ Dealing with longitudinal data in R

Course outline (long version)

> **Monday**: Introduction to longitudinal data and R

- ⇝ Housekeeping
- ⇝ What is longitudinal data?
- ⇝ Wide vs. long datasets
- ⇝ Introduction to R
- ⇝ Dealing with longitudinal data in R

Course outline (long version)

> **Tuesday**: Latent trajectories and growth curve models

- **Morning session**: a multilevel approach
  - Introduction to multilevel models
  - Growth curve models: a multilevel approach
  - Estimating growth curve models using `lme4`
- **Afternoon session**: a structural equation modelling approach
  - Introduction to SEM
  - Latent Growth Curve Analysis: a SEM approach
  - Estimating LGCA models using `lavaan`

Course outline (long version)

> **Tuesday**: Latent trajectories and growth curve models

⇝ **Morning session**: a multilevel approach

- Introduction to multilevel models
- Growth curve models: a multilevel approach
- Estimating growth curve models using `lme4`

⇝ **Afternoon session**: a structural equation modelling approach

- Introduction to SEM
- Latent Growth Curve Analysis: a SEM approach
- Estimating LGCA models using `lavaan`

Course outline (long version)

> **Tuesday**: Latent trajectories and growth curve models

- **Morning session**: a multilevel approach
  - Introduction to multilevel models
  - Growth curve models: a multilevel approach
  - Estimating growth curve models using `lme4`
- **Afternoon session**: a structural equation modelling approach
  - Introduction to SEM
  - Latent Growth Curve Analysis: a SEM approach
  - Estimating LGCA models using `lavaan`

Course outline (long version)

> **Wednesday**: Reverse causality and cross-lagged panel models

- **Morning session**: traditional cross-lagged panel models
  - What is reverse causality *really*?
  - Reverse causality and/vs. reciprocality
  - The traditional cross-lagged panel model
  - Estimating CLPMs using `lavaan`
- **Afternoon session**: modern cross-lagged panel models
  - Recent advancements in cross-lagged panel models
  - The Random Intercepts Cross-Lagged Panel Model
  - The Dynamic Panel Model with fixed effects
  - Estimating RI-CLPM models using `lavaan`
  - Estimating DPMs using `dpm`

Course outline (long version)

> **Wednesday**: Reverse causality and cross-lagged panel models

- **Morning session**: traditional cross-lagged panel models
  - What is reverse causality *really*?
  - Reverse causality and/vs. reciprocality
  - The traditional cross-lagged panel model
  - Estimating CLPMs using `lavaan`
- **Afternoon session**: modern cross-lagged panel models
  - Recent advancements in cross-lagged panel models
  - The Random Intercepts Cross-Lagged Panel Model
  - The Dynamic Panel Model with fixed effects
  - Estimating RI-CLPM models using `lavaan`
  - Estimating DPMs using `dpm`

Course outline (long version)

> **Wednesday**: Reverse causality and cross-lagged panel models

- **Morning session**: traditional cross-lagged panel models
  - What is reverse causality *really*?
  - Reverse causality and/vs. reciprocality
  - The traditional cross-lagged panel model
  - Estimating CLPMs using `lavaan`
- **Afternoon session**: modern cross-lagged panel models
  - Recent advancements in cross-lagged panel models
  - The Random Intercepts Cross-Lagged Panel Model
  - The Dynamic Panel Model with fixed effects
  - Estimating RI-CLPM models using `lavaan`
  - Estimating DPMs using `dpm`

Course outline (long version)

> **Thursday**: Difference-in-differences and causal inference with panel data

- **Morning session**: Basic intuition for causal inference with panel data
  - Causality and the potential outcomes framework
  - Leveraging longitudinal data to identify average causal effects
  - Difference-in-differences in a two-period scenario
  - Simple difference-in-differences analysis in R
- **Afternoon session**: Difference-in-differences regression estimators
  - Using linear regression in the two-period scenario
  - Multi-period difference-in-differences
  - Two-way fixed effects regression estimator
  - Using regression in difference-in-differences analysis in R

Course outline (long version)

> **Thursday**: Difference-in-differences and causal inference with panel data

- **Morning session**: Basic intuition for causal inference with panel data
  - Causality and the potential outcomes framework
  - Leveraging longitudinal data to identify average causal effects
  - Difference-in-differences in a two-period scenario
  - Simple difference-in-differences analysis in R
- **Afternoon session**: Difference-in-differences regression estimators
  - Using linear regression in the two-period scenario
  - Multi-period difference-in-differences
  - Two-way fixed effects regression estimator
  - Using regression in difference-in-differences analysis in R

Course outline (long version)

> **Thursday**: Difference-in-differences and causal inference with panel data

⇝ **Morning session**: Basic intuition for causal inference with panel data

- Causality and the potential outcomes framework
- Leveraging longitudinal data to identify average causal effects
- Difference-in-differences in a two-period scenario
- Simple difference-in-differences analysis in R

⇝ **Afternoon session**: Difference-in-differences regression estimators

- Using linear regression in the two-period scenario
- Multi-period difference-in-differences
- Two-way fixed effects regression estimator
- Using regression in difference-in-differences analysis in R

Course outline (long version)

> **Friday**: Recent advancements in causal inference with panel data

- ⇝ Issues with TWFE estimator
- ⇝ Staggered difference-in-differences
- ⇝ Difference-in-differences when treatment status switches on and off
- ⇝ Discussion of projects

Introduction to longitudinal data

Idea of longitudinal data

What is **longitudinal data**?

- $\rightsquigarrow$ How is it different from **cross-sectional data**?
- $\rightsquigarrow$ How is it different from **time series data**?

$\Rightarrow$ **Repeated observations of the same units over time**

- $\rightsquigarrow$ Intuitively: Large N, small T

Idea of longitudinal data

What is **longitudinal data**?

- $\rightsquigarrow$ How is it different from **cross-sectional data**?
- $\rightsquigarrow$ How is it different from **time series data**?

$\Rightarrow$ **Repeated observations of the same units over time**

- $\rightsquigarrow$ Intuitively: Large N, small T

Idea of longitudinal data

What is **longitudinal data**?

- $\rightsquigarrow$ How is it different from **cross-sectional data**?
- $\rightsquigarrow$ How is it different from **time series data**?

$\Rightarrow$ **Repeated observations of the same units over time**

- $\rightsquigarrow$ Intuitively: Large N, small T

What can we use longitudinal data for?

General intuition: modelling **change over time**

- Modelling **individual trajectories**, treating *TIME* as the main *independent* variable
- Modelling **change over time**, controlling for variables measured at different points in time
- Modelling **effects on change**, focusing on within-unit change over time only
- Modelling **time to event**, focusing on distal outcomes or probability of event happening (survival models, event-history, marginal structural models, ...)

What can we use longitudinal data for?

General intuition: modelling **change over time**

- ⇝ Modelling **individual trajectories**, treating *TIME* as the main *independent* variable
- ⇝ Modelling **change over time**, controlling for variables measured at different points in time
- ⇝ Modelling **effects on change**, focusing on within-unit change over time only
- ⇝ Modelling **time to event**, focusing on distal outcomes or probability of event happening (survival models, event-history, marginal structural models, . . .)

What can we use longitudinal data for?

General intuition: modelling **change over time**

- ⇝ Modelling **individual trajectories**, treating *TIME* as the main *independent* variable
- ⇝ Modelling **change over time**, controlling for variables measured at different points in time
- ⇝ Modelling **effects on change**, focusing on within-unit change over time only
- ⇝ Modelling **time to event**, focusing on distal outcomes or probability of event happening (survival models, event-history, marginal structural models, . . .)

What can we use longitudinal data for?

General intuition: modelling **change over time**

- ⇝ Modelling **individual trajectories**, treating *TIME* as the main *independent* variable
- ⇝ Modelling **change over time**, controlling for variables measured at different points in time
- ⇝ Modelling **effects on change**, focusing on within-unit change over time only
- ⇝ Modelling **time to event**, focusing on distal outcomes or probability of event happening (survival models, event-history, marginal structural models, ...)

What can we use longitudinal data for?

General intuition: modelling **change over time**

- ⇝ Modelling **individual trajectories**, treating *TIME* as the main *independent* variable
- ⇝ Modelling **change over time**, controlling for variables measured at different points in time
- ⇝ Modelling **effects on change**, focusing on within-unit change over time only
- ⇝ Modelling **time to event**, focusing on distal outcomes or probability of event happening (survival models, event-history, marginal structural models, ...)

From cross-sectional to longitudinal data

Imagine we conduct a survey with a sample of 100 individuals and collect information on 5 variables:

`age`, `gender`, `income`, `education`, `employment status`

This gives us a dataset with:

100 rows (individuals) and **5 columns** (variables)

This is a typical cross-sectional dataset.

In standard regression analysis, we often assume observations are ***i.i.d.*** *— independent and identically distributed.*

From cross-sectional to longitudinal data

Imagine we conduct a survey with a sample of 100 individuals and collect information on 5 variables:

`age`, `gender`, `income`, `education`, `employment status`

This gives us a dataset with:

100 rows (individuals) and **5 columns** (variables)

This is a typical cross-sectional dataset.

*In standard regression analysis, we often assume observations are **i.i.d.** — independent and identically distributed.*

From cross-sectional to longitudinal data

Imagine we conduct a survey with a sample of 100 individuals and collect information on 5 variables:

`age`, `gender`, `income`, `education`, `employment status`

This gives us a dataset with:

100 rows (individuals) and **5 columns** (variables)

This is a typical cross-sectional dataset.

*In standard regression analysis, we often assume observations are **i.i.d.** — independent and identically distributed.*

From cross-sectional to longitudinal data

Imagine we conduct a survey with a sample of 100 individuals and collect information on 5 variables:

`age`, `gender`, `income`, `education`, `employment status`

This gives us a dataset with:

100 rows (individuals) and **5 columns** (variables)

This is a typical cross-sectional dataset.

In standard regression analysis, we often assume observations are ***i.i.d.*** *— independent and identically distributed.*

From cross-sectional to longitudinal data

Imagine we conduct a survey with a sample of 100 individuals and collect information on 5 variables:

`age`, `gender`, `income`, `education`, `employment status`

This gives us a dataset with:

100 rows (individuals) and **5 columns** (variables)

This is a typical cross-sectional dataset.

*In standard regression analysis, we often assume observations are **i.i.d.** — independent and identically distributed.*

Repeated measurements – what changes?

Now imagine we go back one year later and collect the same 5 variables from the same 100 individuals.

How should we organise the new data?

Option A – Wide format:
One row per individual, now with 10 columns (e.g., `income_t1`, `income_t2`)

Option B – Long format:
One row per observation (i.e., person-year): 200 rows $\times$ 5 variables + a time column

These are two ways of structuring longitudinal data: *wide* and *long* formats.

Repeated measurements – what changes?

Now imagine we go back one year later and collect the same 5 variables from the same 100 individuals.

How should we organise the new data?

Option A – Wide format:
One row per individual, now with 10 columns (e.g., `income_t1`, `income_t2`)

Option B – Long format:
One row per observation (i.e., person-year): 200 rows $\times$ 5 variables $+$ a time column

These are two ways of structuring longitudinal data: *wide* and *long* formats.

Repeated measurements – what changes?

Now imagine we go back one year later and collect the same 5 variables from the same 100 individuals.

How should we organise the new data?

Option A – Wide format:
One row per individual, now with 10 columns (e.g., `income_t1`, `income_t2`)

Option B – Long format:
One row per observation (i.e., person-year): 200 rows $\times$ 5 variables + a time column

These are two ways of structuring longitudinal data: *wide* and *long* formats.

Repeated measurements – what changes?

Now imagine we go back one year later and collect the same 5 variables from the same 100 individuals.

How should we organise the new data?

Option A – Wide format:
One row per individual, now with 10 columns (e.g., `income_t1`, `income_t2`)

Option B – Long format:
One row per observation (i.e., person-year): 200 rows $\times$ 5 variables + a time column

These are two ways of structuring longitudinal data: *wide* and *long* formats.

Repeated measurements – what changes?

Now imagine we go back one year later and collect the same 5 variables from the same 100 individuals.

How should we organise the new data?

Option A – Wide format:

One row per individual, now with 10 columns (e.g., `income_t1`, `income_t2`)

Option B – Long format:

One row per observation (i.e., person-year): 200 rows $\times$ 5 variables $+$ a time column

These are two ways of structuring longitudinal data: *wide* and *long* formats.

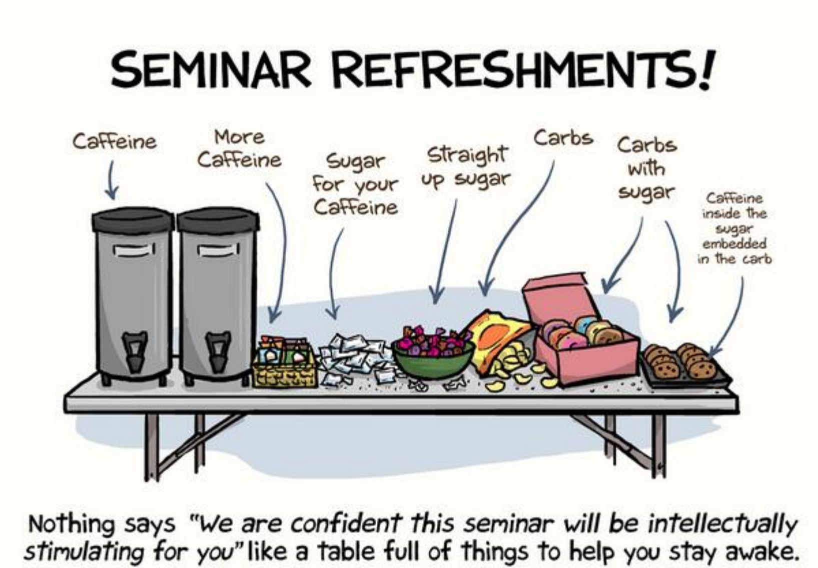
Wide vs Long format

	Wide Format	Long Format
Structure	One row per individual	One row per individual-time point
Columns	One variable per time point (e.g., `income_t1`, `income_t2`)	One outcome column, plus a time indicator (e.g., `income`, `wave`)
Pros	Maintains individuals as the unit of analysis	Much more flexible and efficient
Cons	Inefficient and static	Violates i.i.d. assumptions

⇒ Different modelling strategies require different longitudinal datasets

Your turn!

Coffee break



Thank you!

thiago.oliveira@manchester.ac.uk

ThiagoROliveira.com

@oliveiratr.bsky.social