Quantitative Social Research II
Workshop 9: Longitudinal Data

Jose Pina-Sánchez

Workshop Aims

- Consider the range of new research questions that can be explored using longitudinal data
- Present the main forms of longitudinal data
- Introduce two approaches to model longitudinal data
  - growth curve models
  - dynamic models

What is Longitudinal Data?

- Data that can be used to trace the individual trajectories of multiple subjects across time
- Important difference with cross-sectional data
  - which observations refer to a single time point

What is Longitudinal Data?

- Data that can be used to trace the individual trajectories of multiple subjects across time
- Important difference with cross-sectional data
  - which observations refer to a single time point
- Somehow different to time-series too
  - time-series is a common form of 'longitudinal' data
  - normally focused on one or a limited number of subjects (e.g. countries) for a longer period of time
  - a stronger focus on prediction
  - modelling normally informed by the time-series own past behaviour

New Research Questions

New Research Questions

- Distinction between *cohort* and *age* effects
  - is the crime drop due to an ageing population or to new generations being less prone to offending behaviour?
  - will pro-EU support grow in the UK in the next 20 years after most people aged 60+ have died? (Frese et al., 2024)

New Research Questions

- Distinction between *cohort* and *age* effects
  - is the crime drop due to an ageing population or to new generations being less prone to offending behaviour?
  - will pro-EU support grow in the UK in the next 20 years after most people aged 60+ have died? (Frese et al., 2024)
- Exploring *variability* of trajectories/pathways
  - do grammar schools reduce education/academic inequalities?
  - do between judge disparities change with judicial experience?

New Research Questions

- Distinction between *cohort* and *age* effects
  - is the crime drop due to an ageing population or to new generations being less prone to offending behaviour?
  - will pro-EU support grow in the UK in the next 20 years after most people aged 60+ have died? (Frese et al., 2024)
- Exploring *variability* of trajectories/pathways
  - do grammar schools reduce education/academic inequalities?
  - do between judge disparities change with judicial experience?
- Estimating the speed of *transitions* from one status to another and the factors influencing them
  - how long does it take women to go back to work after they give birth?
  - does the gender of a partner affect the duration of paternity/maternity leave period?

New Research Questions

- Distinction between *cohort* and *age* effects
  - is the crime drop due to an ageing population or to new generations being less prone to offending behaviour?
  - will pro-EU support grow in the UK in the next 20 years after most people aged 60+ have died? (Frese et al., 2024)
- Exploring *variability* of trajectories/pathways
  - do grammar schools reduce education/academic inequalities?
  - do between judge disparities change with judicial experience?
- Estimating the speed of *transitions* from one status to another and the factors influencing them
  - how long does it take women to go back to work after they give birth?
  - does the gender of a partner affect the duration of paternity/maternity leave period?
- Obtain better causal estimates and explore *reverse causality*
  - is the effect of procedural justice on legitimacy unidirectional?

Longitudinal Designs

- Two main longitudinal designs (data collection strategies)
  - panel (aka prospective, longitudinal) designs
  - retrospective questions

Longitudinal Designs

- Two main longitudinal designs (data collection strategies)
  - panel (aka prospective, longitudinal) designs
  - retrospective questions
- Panel designs
  - multiple contacts with the same participants
  - detect changes across contact points

Longitudinal Designs

- Two main longitudinal designs (data collection strategies)
  - panel (aka prospective, longitudinal) designs
  - retrospective questions
- Panel designs
  - multiple contacts with the same participants
  - detect changes across contact points
  - ex.1: *Do you trust your neighbours?*
  - ex.2: *Which of the following reflect your work status more accurately: employed, unemployed, or out of the labour force?*

Longitudinal Designs

- Two main longitudinal designs (data collection strategies)
  - panel (aka prospective, longitudinal) designs
  - retrospective questions
- Panel designs
  - multiple contacts with the same participants
  - detect changes across contact points
  - ex.1: *Do you trust your neighbours?*
  - ex.2: *Which of the following reflect your work status more accurately: employed, unemployed, or out of the labour force?*
  - can miss changes taking place between interviews that go back to their previous value
  - affected by attrition, i.e. compound non-response

Longitudinal Designs

- Retrospective questions
  - single contact with the participants
  - unaffected by attrition
  - can potentially identify any changes within the recall period

Longitudinal Designs

- Retrospective questions
  - single contact with the participants
  - unaffected by attrition
  - can potentially identify any changes within the recall period
  - ex.1: *How old were you when you first started smoking Marijuana?*
  - ex.2: *Please identify whether you were employed, unemployed, or out of the labour force from January 2018 to the present day, indicating the date(s) when your status changed*

Longitudinal Designs

- Retrospective questions
  - single contact with the participants
  - unaffected by attrition
  - can potentially identify any changes within the recall period
  - ex.1: *How old were you when you first started smoking Marijuana?*
  - ex.2: *Please identify whether you were employed, unemployed, or out of the labour force from January 2018 to the present day, indicating the date(s) when your status changed*
  - rely on memory → prone to recall errors and telescoping effects

Growth Curve Models

- Useful to explore variability in individual trajectories
  - require repeated measurements across time for the same subjects

Growth Curve Models

- Useful to explore variability in individual trajectories
  - require repeated measurements across time for the same subjects
- In essence based on multilevel modelling
  - longitudinal data taken as 2-level hierarchical data
  - individual responses (level 1) clustered within subjects (level 2)
  - we can change the notation to make it more intuitive

    level 2 cases (normally the individual level): $i = 1, 2, 3, ..., I$

    level 1 cases, time level: $t = 1, 2, 3, ..., T_i$
  - ex.1: reading scores of pupils in England throughout the 'Key Stages' national Curriculum
  - ex.2: the probability of judges imposing custodial sentence throughout their careers
  - allow initial scores (intercept) and progress (slope) to vary across individuals

Modelling Time using Standard Regression

- We can assess the average trajectory across individuals using standard regression models
  - $Y_{it} = \beta_0 + \beta_1 \, TIME_{it} + \beta_k X_{kit} + e_{it}$
  - with $TIME_{it}$ as a variable capturing the sequence of contact times across the window of observation

    $TIME_i = 1, 2, 3, ..., T$
  - and β_1 indicating the average change of the outcome variable across the window of observation
  - in ex.1, reading scores probably go up across time, $\beta_1 > 0$; while for ex.2 severity might remain stable, $\beta_1 \approx 0$

Modelling Time using Standard Regression

- We can assess the average trajectory across individuals using standard regression models
  - $Y_{it} = \beta_0 + \beta_1 TIME_{it} + \beta_k X_{kit} + e_{it}$
  - with $TIME_{it}$ as a variable capturing the sequence of contact times across the window of observation

    $TIME_i = 1, 2, 3, ..., T$
  - and β_1 indicating the average change of the outcome variable across the window of observation
  - in ex.1, reading scores probably go up across time, $\beta_1 > 0$; while for ex.2 severity might remain stable, $\beta_1 \approx 0$
- This model can be extended in multiple ways
  - can modify the linear assumption using polynomial regression
  - AGE can be included (instead of $TIME$) together with $COHORT$ to disentangle them
  - can use the sandwich estimator to obtain robust standard errors

Modelling Time using Random Effects

- We can use random slopes to model the variability in individual trajectories
  - $Y_{it} = \overbrace{\beta_{0i}}^{\beta_0+u_{0i}} + \overbrace{\beta_{1i}}^{\beta_1+u_{1i}} TIME_{it} + \beta_k X_{kit} + \epsilon_{it}$
  - σ_{u0} captures subject variability in initial scores
    σ_{u1} captures subject variability in progress across time

Modelling Time using Random Effects

- We can use random slopes to model the variability in individual trajectories
  - $Y_{it} = \overbrace{\beta_{0i}}^{\beta_0+u_{0i}} + \overbrace{\beta_{1i}}^{\beta_1+u_{1i}} \; TIME_{it} + \beta_k X_{kit} + \epsilon_{it}$
  - σ_{u0} captures subject variability in initial scores

    σ_{u1} captures subject variability in progress across time
  - the key role of $cov(u_0, u_1)$

    if positive the slopes will diverge, i.e. higher intercepts are associated with higher slopes and vice versa

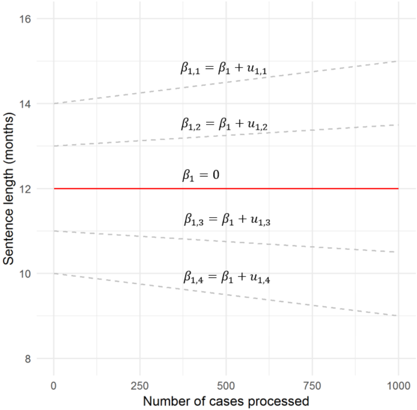
    if negative the slopes will converge, i.e. higher intercepts are associated with lower slopes and vice versa
  - invokes additional assumptions

    $u_{0i} \sim N(0, \sigma_{u0})$; $cov(u_{0i}, u_{0i'}) = 0$

    $u_{1i} \sim N(0, \sigma_{u1})$; $cov(u_{1i}, u_{1i'}) = 0$

    $\epsilon_i \sim N(0, \sigma_\epsilon)$; $cov(\epsilon_i, \epsilon_{i'}) = 0$
  - can include additional levels, e.g. reading scores clustered within pupils clustered within schools

Growth Curve Models (+cov)



Dynamic Models

- We can also use longitudinal data to delve deeper into the exploration of causal effects
  - we can explore whether an assumed *cause* precedes its corresponding *effect*
  - essential condition to be able to ascertain a causal mechanism
  - we will be able to explore the possibility of reverse causal paths

Dynamic Models

- We can also use longitudinal data to delve deeper into the exploration of causal effects
  - we can explore whether an assumed *cause* precedes its corresponding *effect*
  - essential condition to be able to ascertain a causal mechanism
  - we will be able to explore the possibility of reverse causal paths
- The simplest case does not require repeated measures
  - only that the effect $Y_{i,2}$ is measured at a later period than the cause $X_{i,1}$
  - $Y_{i,2} = \beta_0 + \beta_1 X_{i,1} + e_i$

Dynamic Models

- We can also use longitudinal data to delve deeper into the exploration of causal effects
  - we can explore whether an assumed *cause* precedes its corresponding *effect*
  - essential condition to be able to ascertain a causal mechanism
  - we will be able to explore the possibility of reverse causal paths
- The simplest case does not require repeated measures
  - only that the effect $Y_{i,2}$ is measured at a later period than the cause $X_{i,1}$
  - $Y_{i,2} = \beta_0 + \beta_1 X_{i,1} + e_i$
  - commonly used in the literature, yet far from ideal
  - β_1 represents the effect of X_1 on Y_2, not the change of Y from $t = 1$ to $t = 2$

Autorregressive Models

- Models where *lagged* outcome variables are included as explanatory variables
  - requires repeated measures across time

Autorregressive Models

- Models where *lagged* outcome variables are included as explanatory variables
  - requires repeated measures across time
- The simplest form is the first order autoregressive model
  - where only the previous time point is considered
  - for the simple case of $t = 1, 2$, and one explanatory variable
    $Y_{i,2} = \beta_0 + \beta_1 Y_{i,1} + \beta_2 X_{i,1} + e_i$

Autorregressive Models

- Models where *lagged* outcome variables are included as explanatory variables
  - requires repeated measures across time
- The simplest form is the first order autoregressive model
  - where only the previous time point is considered
  - for the simple case of $t = 1, 2$, and one explanatory variable

    $Y_{i,2} = \beta_0 + \beta_1 Y_{i,1} + \beta_2 X_{i,1} + e_i$
  - if we have $t = 1, 2, ..., T$, and $k = 1, 2, ..., K$, then the first order autoregressive model will take the more general form

    $Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_k X_{kit-1} + e_{it}$

Autorregressive Models

- Models where *lagged* outcome variables are included as explanatory variables
  - requires repeated measures across time
- The simplest form is the first order autoregressive model
  - where only the previous time point is considered
  - for the simple case of $t = 1, 2$, and one explanatory variable
    $Y_{i,2} = \beta_0 + \beta_1 Y_{i,1} + \beta_2 X_{i,1} + e_i$
  - if we have $t = 1, 2, ..., T$, and $k = 1, 2, ..., K$, then the first order autoregressive model will take the more general form
    $Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_k X_{kit-1} + e_{it}$
- β_1 is known as the stability coefficient
  - indicates how much of the variability in the outcome is due to its previous value
  - after controlling for β_1, then β_2 can be taken as the effect of $X_{1,t-1}$ on the change of Y from one period to the next
  - this offers a neater approximation to causal effects and controls for lots of time-invariant confounders

Cross-Lagged Models

- To explore a potential reverse causal effect we can use cross-lagged models
  - estimating autoregressive models for X and Y simultaneously using structural equation modelling (SEM)

Cross-Lagged Models

- To explore a potential reverse causal effect we can use cross-lagged models
  - estimating autoregressive models for X and Y simultaneously using structural equation modelling (SEM)
  - for the simple case of $t = 1, 2$, and one explanatory variable:

    $Y_{i,2} = \beta_0 + \beta_1 Y_{i,1} + \beta_2 X_{i,1} + e_i$

    $X_{i,2} = \beta_0' + \beta_1' X_{i,1} + \beta_2' Y_{i,1} + e_i'$

Cross-Lagged Models

- To explore a potential reverse causal effect we can use cross-lagged models
  - estimating autoregressive models for X and Y simultaneously using structural equation modelling (SEM)
  - for the simple case of $t = 1, 2$, and one explanatory variable:

    $Y_{i,2} = \beta_0 + \beta_1 Y_{i,1} + \beta_2 X_{i,1} + e_i$

    $X_{i,2} = \beta_0^{'} + \beta_1^{'} X_{i,1} + \beta_2^{'} Y_{i,1} + e_i^{'}$
  - two stability effects (β_1 and $\beta_1^{'}$) and two cross-lagged effects (β_2 and $\beta_2^{'}$)
  - time 2 residuals are correlated, $cov(e_i, e_i^{'} \neq 0)$, that's why we rely on SEM rather than estimating each model separately

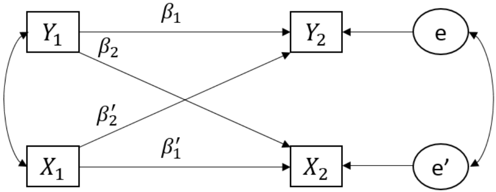
Cross-Lagged Models

- To explore a potential reverse causal effect we can use cross-lagged models
  - estimating autoregressive models for X and Y simultaneously using structural equation modelling (SEM)
  - for the simple case of $t = 1, 2$, and one explanatory variable:

    $$Y_{i,2} = \beta_0 + \beta_1 Y_{i,1} + \beta_2 X_{i,1} + e_i$$

    $$X_{i,2} = \beta_0^{'} + \beta_1^{'} X_{i,1} + \beta_2^{'} Y_{i,1} + e_i^{'}$$
  - two stability effects (β_1 and $\beta_1^{'}$) and two cross-lagged effects (β_2 and $\beta_2^{'}$)
  - time 2 residuals are correlated, $cov(e_i, e_i^{'} \neq 0)$, that's why we rely on SEM rather than estimating each model separately
- Interpreting the cross-lagged coefficients we can...
  - detect the presence of reverse causality; if $\beta_2^{'}$ is statistically significant
  - if the measures of X and Y are standardised we can also determine which causal path is stronger by comparing β_2 and $\beta_2^{'}$

Cross-Lagged Models



Assumptions

- The cross-lagged model represents a great tool to investigate causal mechanisms
- But it is still affected by multiple problems
  - uses *observational* (not *experimental*) data, i.e. confounder bias
  - the causal effect takes place at the exact length of the interval between measurement periods
  - the variables used might be affected by measurement error
  - the causal effect might be non-linear

Assumptions

- The cross-lagged model represents a great tool to investigate causal mechanisms
- But it is still affected by multiple problems
  - uses *observational* (not *experimental*) data, i.e. confounder bias
  - the causal effect takes place at the exact length of the interval between measurement periods
  - the variables used might be affected by measurement error
  - the causal effect might be non-linear
- Many of those problems can be tackled by extending the model
  - we have explored a 'two-wave cross-lagged model' but could add many more waves
  - could include 'contemporaneous effects' accounting for the relationship between X and Y at the same wave
  - can also specify fixed or random effects to control for all the time-invariant heterogeneity
  - can use latent variable estimation to obtain better measures of X and Y
  - can use polynomial regression to explore non-linearity

Recap

- Longitudinal data allows exploring a whole new range of research questions
- Growth curve models
  - to explore variability in individual trajectories

Recap

- Longitudinal data allows exploring a whole new range of research questions
- Growth curve models
  - to explore variability in individual trajectories
- Cross-lagged panel models
  - to approximate causal effects
  - and explore reverse causality

Recap

- Longitudinal data allows exploring a whole new range of research questions
- Growth curve models
  - to explore variability in individual trajectories
- Cross-lagged panel models
  - to approximate causal effects
  - and explore reverse causality
- Many other modelling strategies we do not have time to cover
  - event history analysis models

Recap

- Longitudinal data allows exploring a whole new range of research questions
- Growth curve models
  - to explore variability in individual trajectories
- Cross-lagged panel models
  - to approximate causal effects
  - and explore reverse causality
- Many other modelling strategies we do not have time to cover
  - event history analysis models
- To learn more about growth curve models
  - Steele (2008) 'Multilevel models for longitudinal data'
- To learn more about cross-lagged panel models
  - Selig and Little (2012) 'Autoregressive and cross-lagged panel analysis for longitudinal data'
  - Brunton-Smith (2011) 'Untangling the relationship between fear of crime and perceptions of disorder'