Applying Structural Equation Modeling to Analyzing Longitudinal Data

Hsueh-Sheng Wu
CFDR Workshop Series
February 28, 2024

BGSU
Center for Family and Demographic Research

Outline of Workshop

- What are longitudinal data?
- Casual inference
- Panel data and causal inference
- A brief introduction to SEM
- Select SEM models in longitudinal data analyses
  - Cross-lagged panel model
  - Latent transition model
  - Growth curve model
- Conclusions

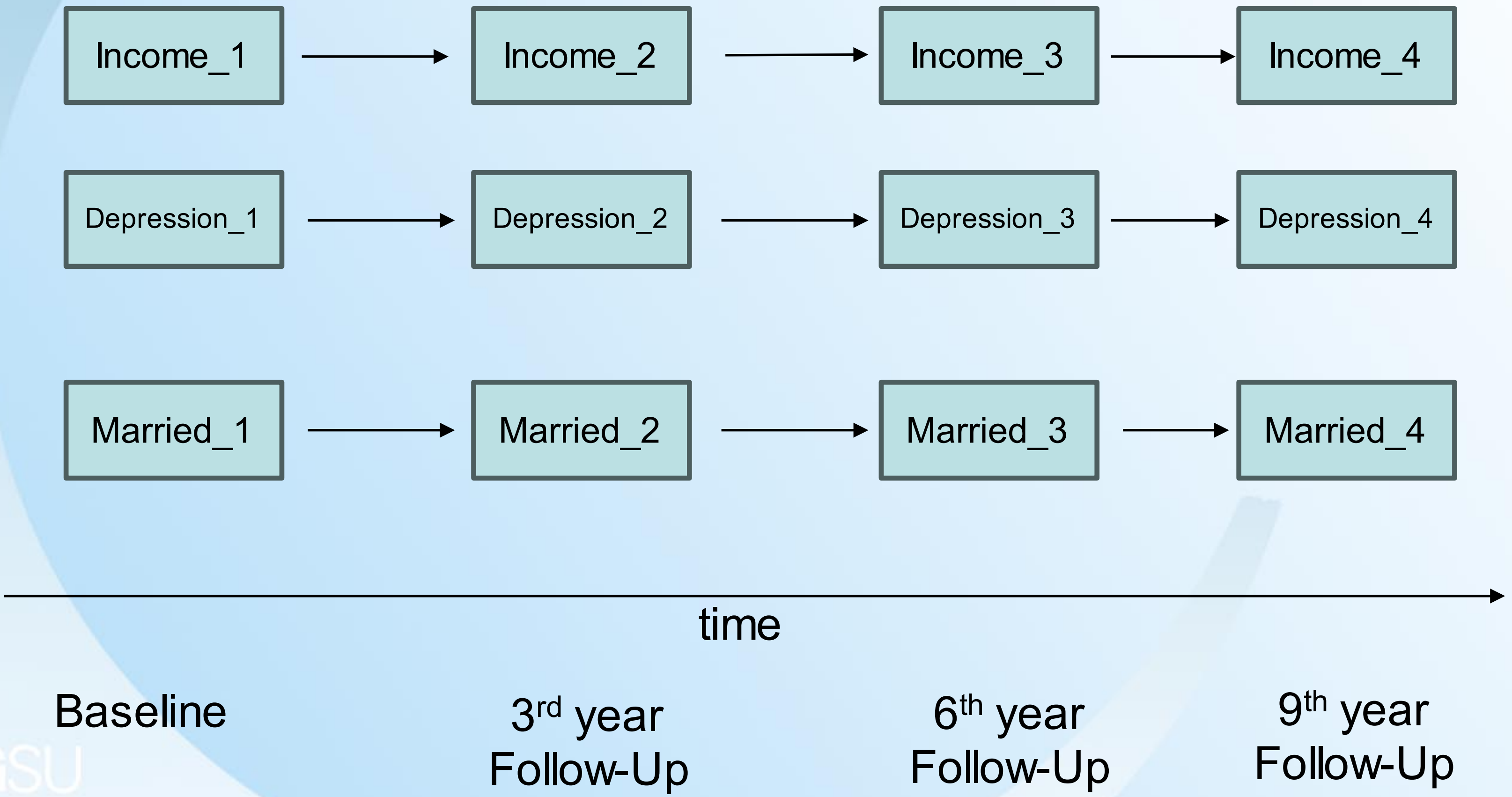
What Are Longitudinal Data?

- Longitudinal data refer to a type of data collected from the same individuals (e.g., panel data) or different individuals of the same population (e.g., repeated cross-sectional data) over a period of time.
- Only longitudinal data from same individuals over time can be used to study how and why individuals change over time
- Not all panel data are created equal for examining the longitudinal relations among variables.
  - How long were respondents followed?
  - How many times were respondents followed?
  - What time interval at which were they followed?
  - Do the data have all the independent, dependent, and control variables measured at all measurement occasions?
  - Do the data have other methodological issues such as missing data, sample attrition, and inconsistent measurements over time?
- Panel data are probably the most common type of longitudinal data used in social science research

BGSU
Center for Family and Demographic Research

What Are Longitudinal Data? (Continued)

An example of panel data



What Are Longitudinal Data? (Continued)

Information about time, respondents, variables, and temporal order of variables in the longitudinal data shown above.

- Respondents were followed for 9 years
- Data were collected from individuals 4 times at an interval of 3 years
- Income and depression are continuous variables, and marital status is a categorical variable
- With each variable having 4 measures, researchers can construct it differently. For example, these variables can be conceptualized as the trajectory of the variable, the memberships of group having different trajectory of the variable, or simply used as a variable measured at different time points
- The temporal order of these variables at each time points is well-established, but less so when they are conceptualized as trajectories
- There may be reciprocal relations among these variables across time, and these reciprocal relations may differ across time
- The graph does not have other variables that contribute to concurrent and/or temporal associations among variables

Causal Inferences

- The ultimate goal for social science research is to empirically establish the causal relation between independent and dependent variables
- Three conditions of causality:
  - **Temporal order of X and Y** — X must come before Y
  - **Significant relation between X and Y** — The relation between X and Y cannot occur by chance alone
  - **The X-Y relation is not spurious** — There are no other intervening or unaccounted for variable that is responsible for the relation between X and Y
- Challenges:
  - Are these conditions automatically met in longitudinal data?
  - Are there additional conditions that need to be made to establish causality?

Panel Data and Causal Inference

Use panel data to draw causal inference on the X-Y relation

Advantages:

- Temporal sequence of variables is generally clearer (vs. cross-sectional data)
- Variables are measured at the individual level, which allows for analyzing within-individual changes and between-individual differences
- Information on the passage of time is available and can be incorporated into some SEM models to gain additional insights about individual changes

Disadvantages:

- The more waves of data are available, the more comprehensive the statistical models can be, and the more difficult for the model to converge
- The availability of panel data allows for testing various different SEM models, but researchers need to choose, specify, and justify their SEM models
- Models become more complex when they involves latent variables and/or control variables.
- Little is know about the actual functional form of change in the variables, and researchers often need to explore different functional forms in their analyses.
- When variables represent parallel processes over time, establishing temporal precedence between them may be challenging
- The cross-lagged effects of X on Y may vary over time, yet the duration required for X to exert its influence on Y remains uncertain
- Other methodological issues such as missing data, sample attrition, and measurement invariance, add additional analytic complexities

A Brief Introduction to SEM

- Structural Equation Modeling (SEM) is a statistical method used to simultaneously test and estimate relations among variables, both observed and latent, by incorporating aspects of measurement models and path analysis
- SEM allows researchers to test whether their hypothesized relations among observed variables are supported by the empirically associations among these variables in the data
- Researchers can continue modifying and testing their hypothetical models against the observed variables until the fit between the proposed model and the observed data is satisfactory and the proposed model is parsimonious
- SEM can be expressed in equations or path diagrams. We will mainly focus on path diagrams today

A Brief Introduction to SEM (Continued)

Variables:

- **Observed variables:** Represented by rectangles or squares
- **Latent variables:** Represented by circles or ovals without indicators inside. Sometimes, when the focus is on theoretical relations or when the model is relatively simple, the error terms of latent or observed variable may not be explicitly shown in circles to simplify the path diagrams.

Paths:

- **Regression paths:** Represented by arrows indicating the direction of influence between variables. These arrows typically point from predictor variables to outcome variables in path models or from latent variables to their indicators.
- **Covariance paths:** Represented by curved lines or double-headed arrows connecting error terms. These paths indicate the covariances of variables.

Steps of Conducting SEM Analysis

- Develop a theoretically based model
- Construct the SEM diagram
- Convert the SEM diagram into a set of structural equations
- Clean data and decide the input data type
- Determine the estimation method
- Run the model and evaluate goodness of fit of the model
- Modify the model
- Compare two models and decide if additional modification is needed

BGSU
Center for Family and Demographic Research

Select SEM Models in Longitudinal Data Analyses

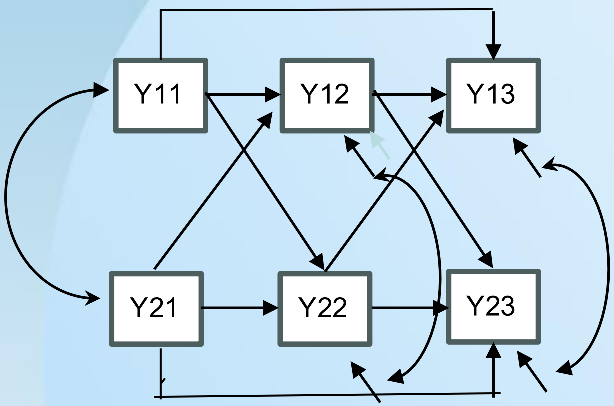
- Cross-lagged panel model examines the directional relations between variables measured at multiple time points, emphasizing how variables at an earlier time point predict another variable at a later time point

- Latent transition model analyzes transitions between categorical latent variables over time, capturing shifts between latent states across different time points

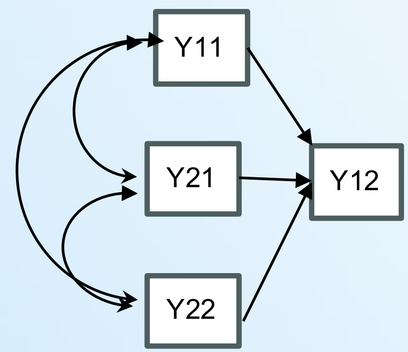
- Growth curve model examines changes in continuous latent variables over time, typically using mathematical functions to represent growth trajectories

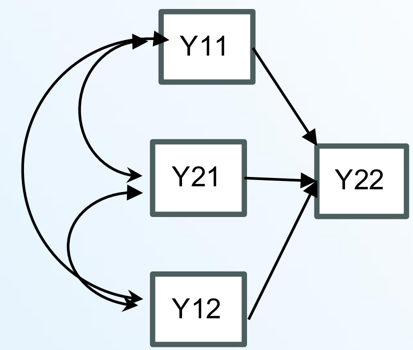
Cross-lagged Panel Model

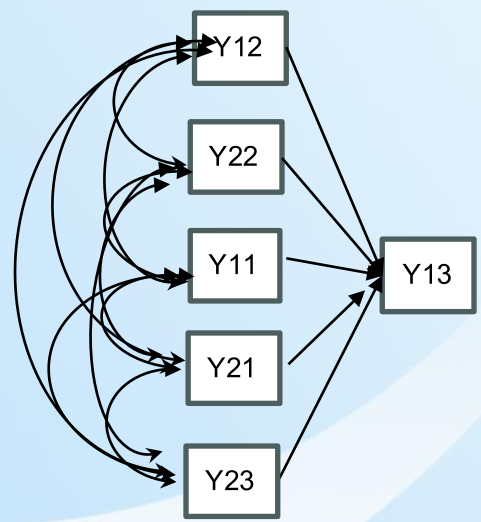
Cross-lagged Panel Models With Observed Variables

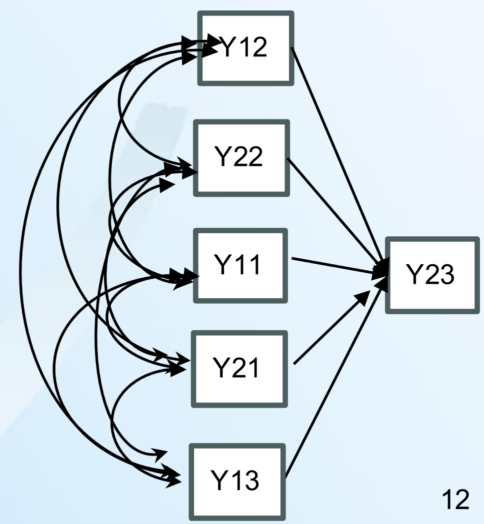


Cross-lagged Regression Models (i.e., not a good alternative to SEM models)



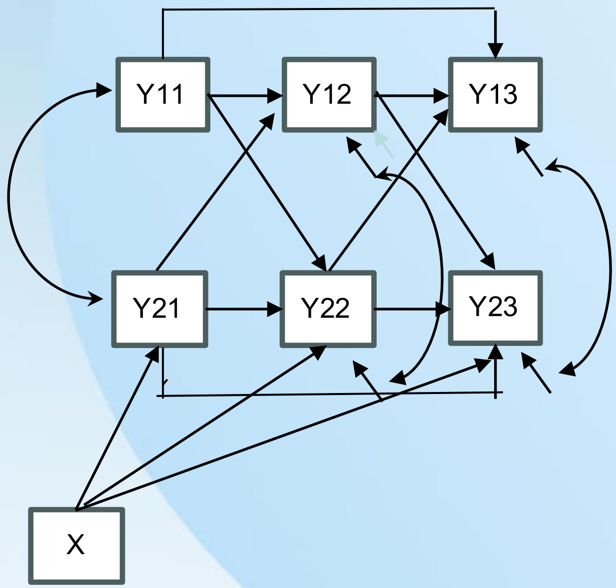




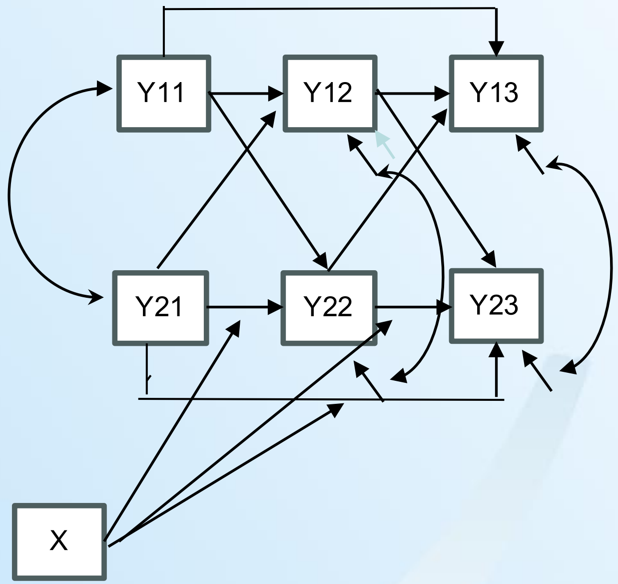


Cross-lagged Panel Model (Continued)

Cross-lagged Panel Models With Covariates



Cross-lagged Panel Models With Moderators



Strengths and Weakness of Cross-lagged Panel Model

Strengths:

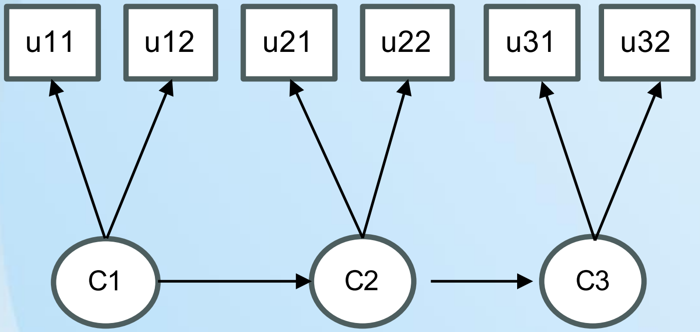
- Enables causal inference by assessing the ranking of one variable at an earlier time is related to that of another variable at a later time, that is, the directional relations among between-individual differences of variables across different time points
- Suitable for analyzing longitudinal data, allowing ab examination of changes and stability of between-individual differences.
- Accounts for autocorrelation and concurrent correlations by including lagged variables and correlated variances of variables.
- Facilitates testing of reciprocal relations between variables

Weaknesses:

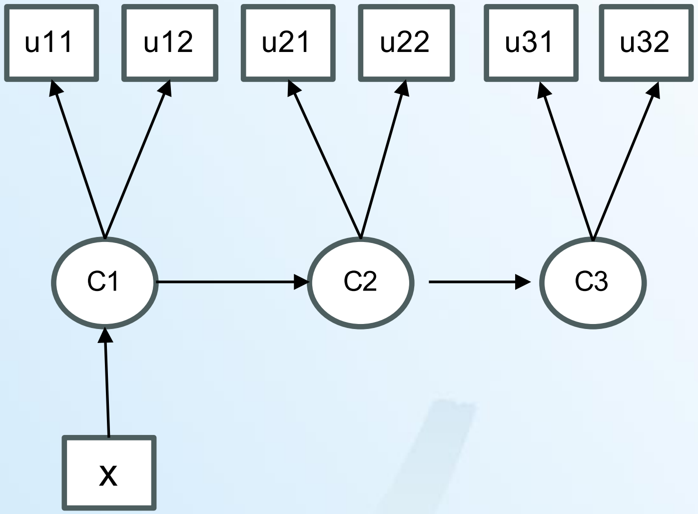
- The model focuses on between-individual differences at certain time points and does not examine within-individual changes over time.
- Establishing causality can be difficult due to confounding factors without knowing whether individuals experience increase or decreases in these variables.
- Measurement errors and measurement invariances are not considered
- Complexity may arise quickly, especially with multiple variables, potentially leading to overfitting

Latent Transition Analysis (LTA)

LTA

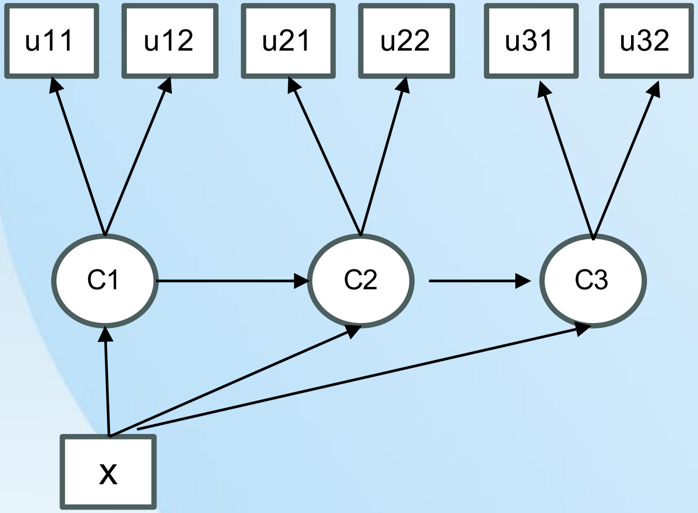


LTA with covariates

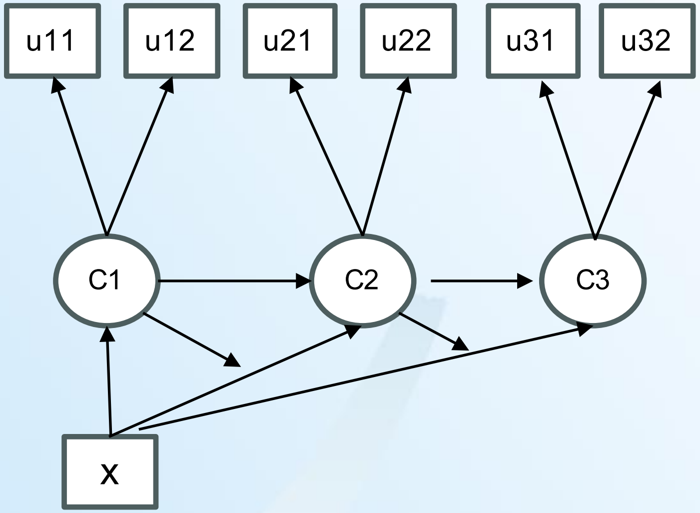


Latent Transition Analysis (Continued)

LTA with covariates

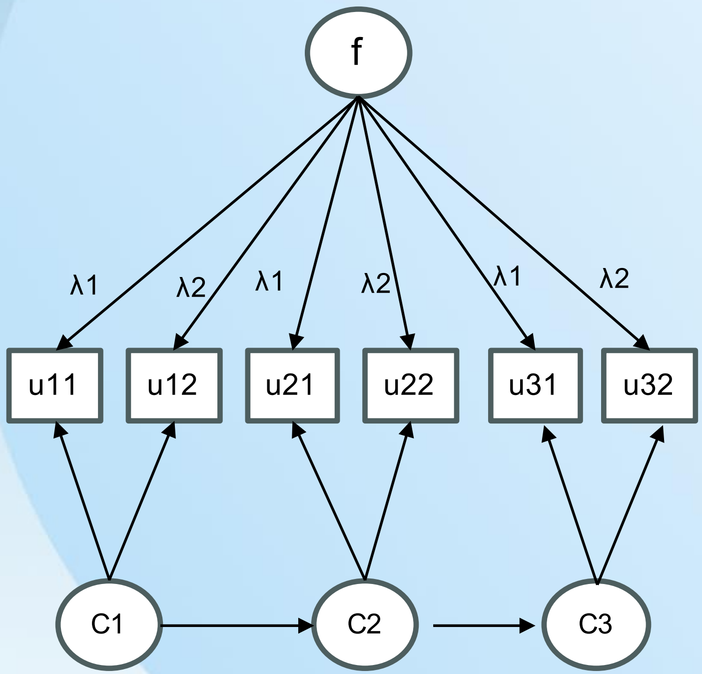


LTA with covariates

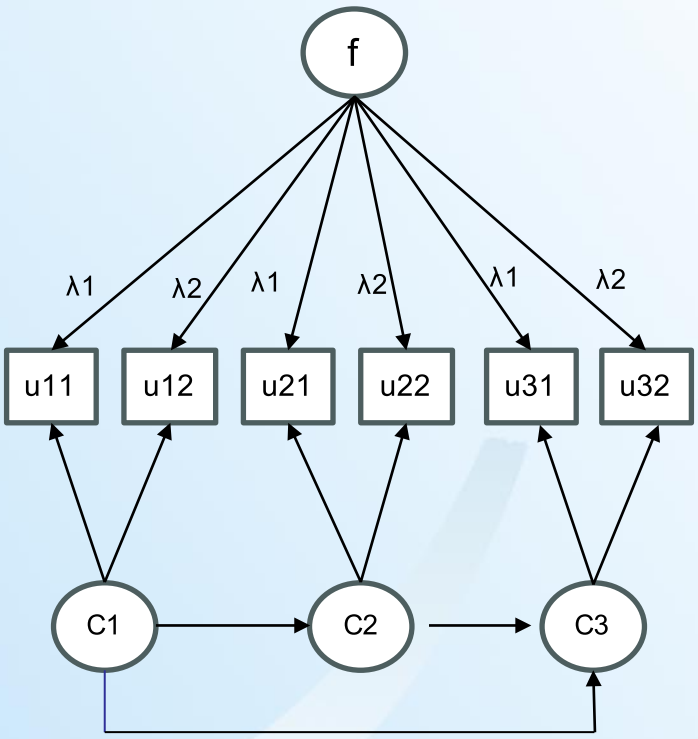


Latent Transition Analysis (Continued)

Random Intercept- LTA



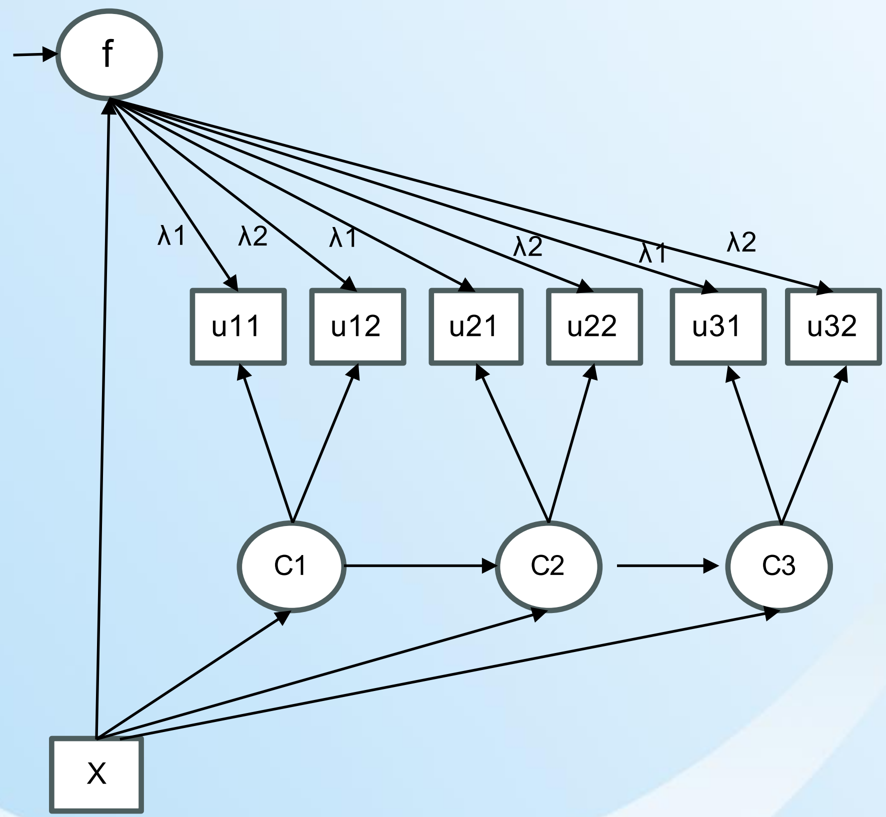
Random Intercept- LTA with lag-2 effect



BGSU
Center for Family and Demographic Research

Latent Transition Analysis (Continued)

Random intercept- LTA with lag-2 effect
and covariates



BGSU
Center for Family and Demographic Research

Strengths and Weakness of Latent Transition Model

Strengths:

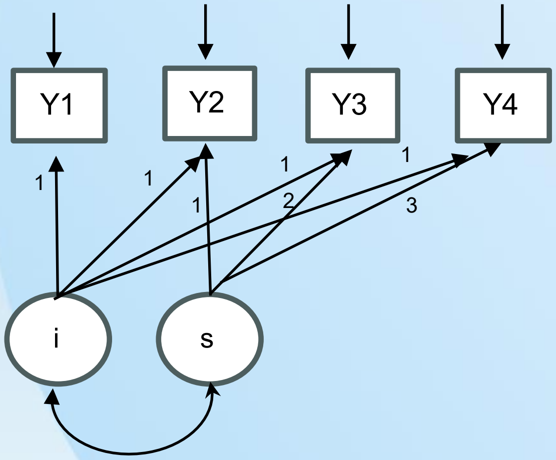
- Provides a framework to analyze transitions between latent states over time
- Offers flexibility in modeling categorical latent variables and their transitions
- Allows for the examination of individual-level changes across time points.
- Incorporates uncertainty in latent class assignment, providing more accurate estimates

Weaknesses:

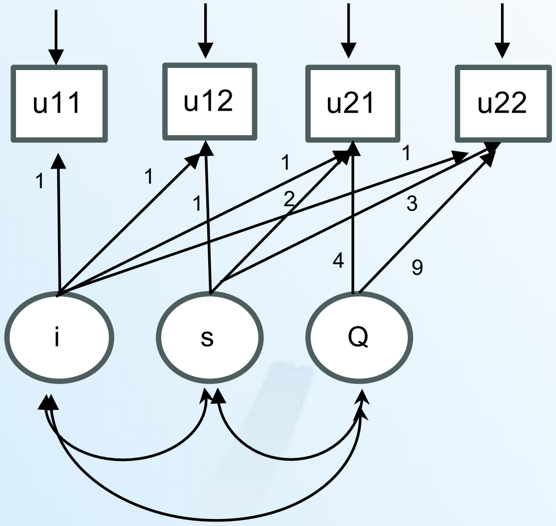
- Requires assumptions about the number and nature of latent states
- May be computationally intensive, especially with complex models or large datasets
- Interpretation of latent classes and transitions can be challenging
- Sensitivity to model misspecification, particularly in determining the appropriate number of latent states

Growth Curve Model

Linear Growth Model



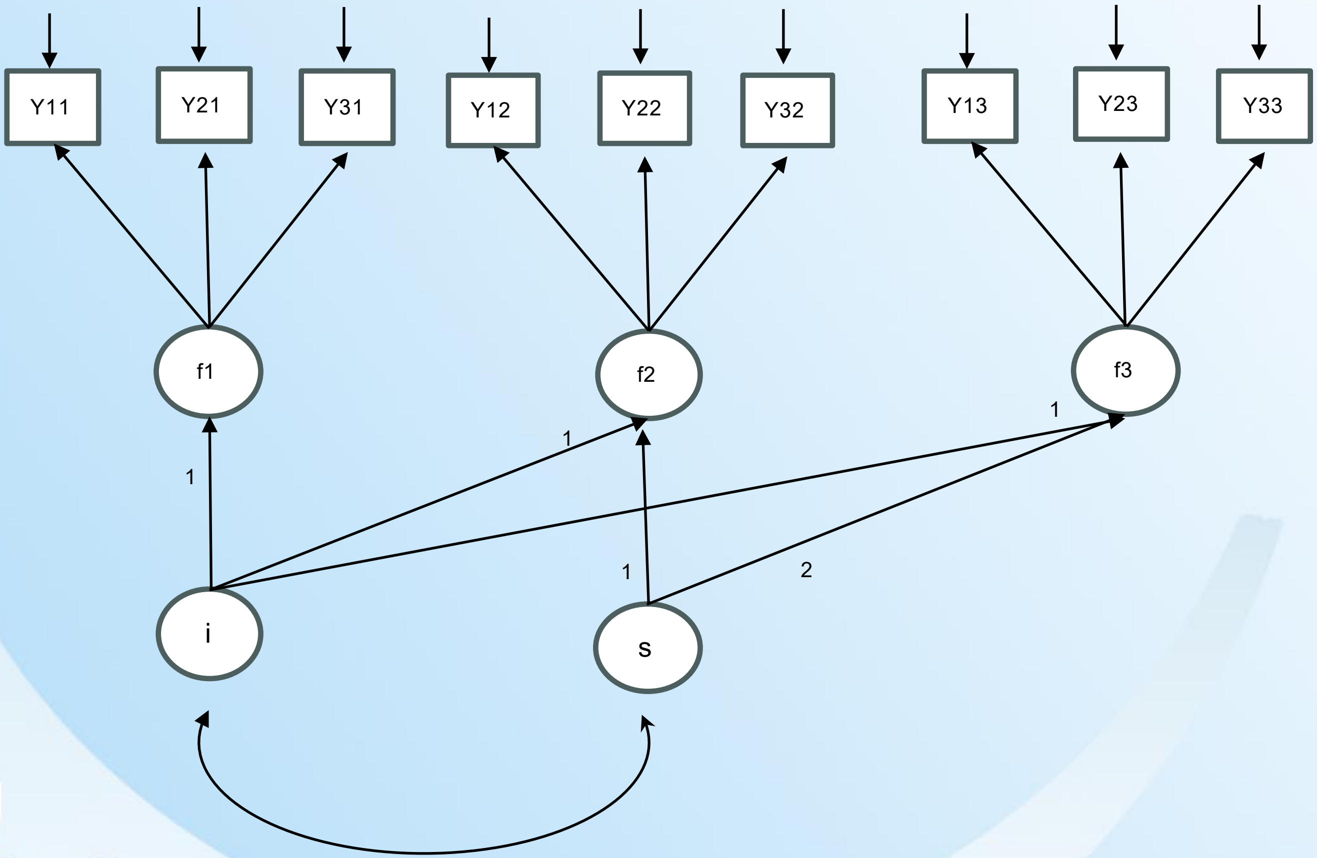
Quadratic Growth Model



BGSU
Center for Family and Demographic Research

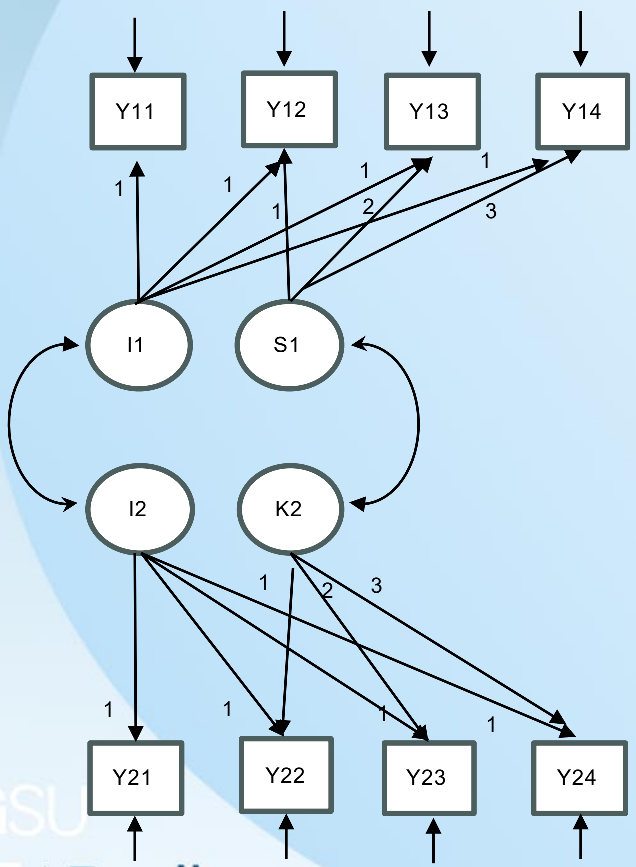
Growth Curve Model (Continued)

Multiple Indicator Linear Growth Model

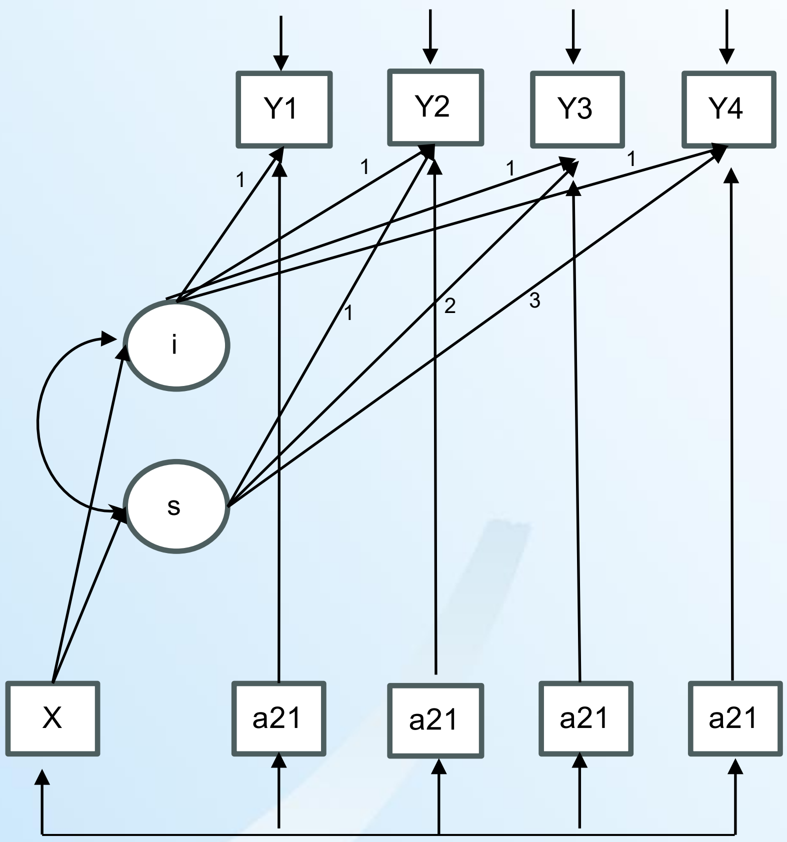


Growth Curve Model (Continued)

Parallel Linear Growth Model

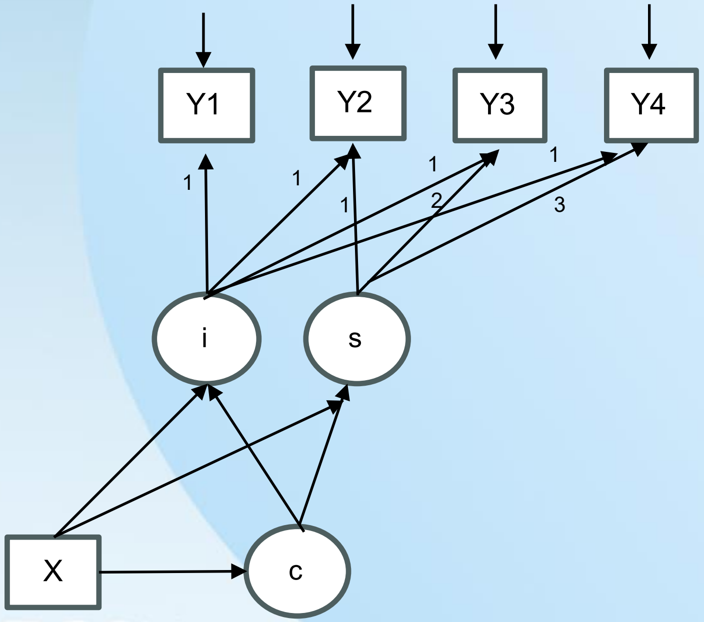


Linear Growth Model with Time-Varying and Time-Invariant Covariates

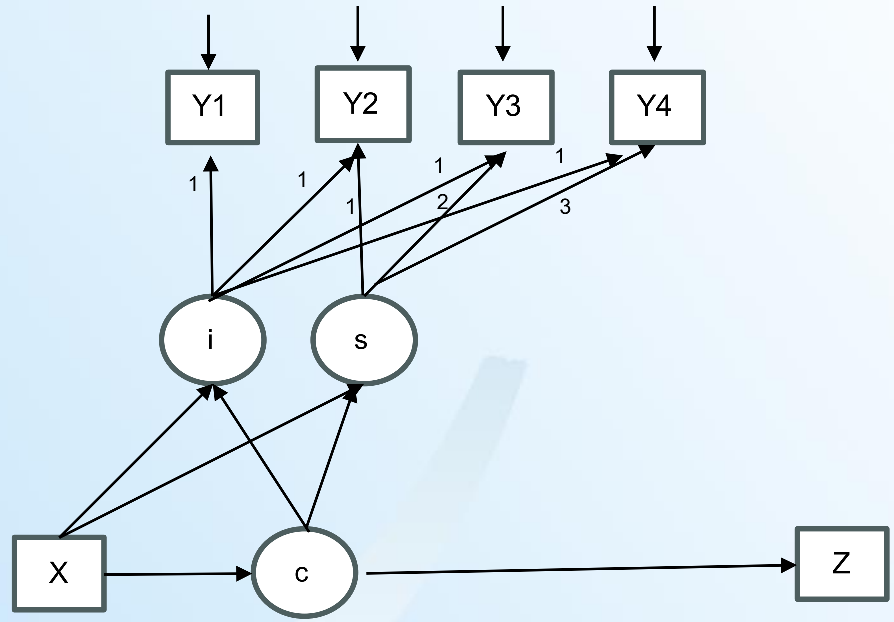


Growth Curve Model (continued)

Linear Growth Mixture Model
With a covariate



Linear Growth Mixture Model with a covariate and a distal outcome



BGSU
Center for Family and Demographic Research

Strengths and Weakness of Growth Curve Model

Strengths:

- Considers intra-individual change as a continuous trajectory and models change as a function of time
- Different growth factors (i.e., intercept, linear, and quadratic factors) depict the shape of the trajectory over time. The predictors of these factors delineate inter-individual differences on the trajectories
- Incorporate parallel trajectories, time-varying and time-invariant predictors, and latent class analyses to understand why and how individuals change over time

Weaknesses:

- Choosing the appropriate functional form for growth (e.g., linear, quadratic) and determining the number of latent factors can be subjective and may affect model fit and interpretation
- It can be difficult to interpret why factors may influence certain growth factors but not the others
- The models may not converge when too many covariates and control variables are added to the model

Conclusions

- SEM offers powerful capabilities to analyze longitudinal data, including modeling complex relations among variables, accounting for measurement error, testing causal pathways, estimating change over time, and comparing different hypothesized models
- Researchers should carefully consider the nature of the outcome variable, whether it involves directional associations among variables, transitions between latent states, or changes in continuous variable, as different SEM models are needed for testing different outcome variables.
- When fitting SEM models, researchers may want to start with a full model that includes all specified theoretical relations and later trim the model by constraining insignificant paths or parameters to zero, ensuring that the significant paths and parameters in the final model remains theory-driven rather than data-driven
- SEM analyzes the variance and covariance matrix of observed variables. Different syntax are used for analyzing variances and covariances of continuous variables than those of categorical variables. When the models contains both continuous and categorical variables, correct syntax should be used.
- SEM can be estimated using SAS, SPSS, Stata, or Mplus. However, they vary in their capacity in handling categorical variables, missing data, complex survey weights
- Although mastering SEM modeling might seem challenging at first, gaining proficiency in various SEM models and their respective syntaxes can be fulfilling, enabling researchers to confidently evaluate hypotheses using empirical data. If you encounter any difficulties while running SEM analyses, please don't hesitate to reach out to me at wuh@bgsu.edu.