Long Longitudinal Data Modeling

Bengt Muthén
bmuthen@statmodel.com
www.statmodel.com

Papers at www.statmodel.com/papers

SMEP Sells Award presentation, October 12, 2018

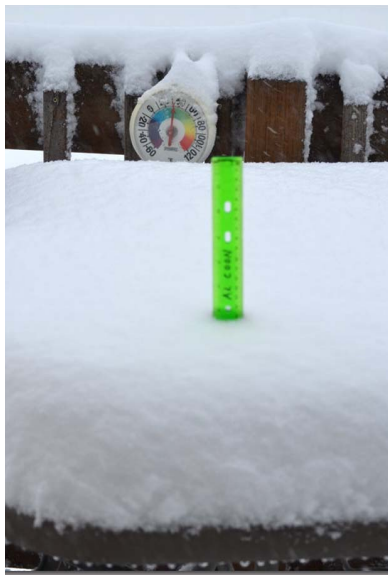
I thank Tihomir Asparouhov, Ellen Hamaker, and Marten Schultzberg for helpful comments and Noah Hastings for excellent assistance

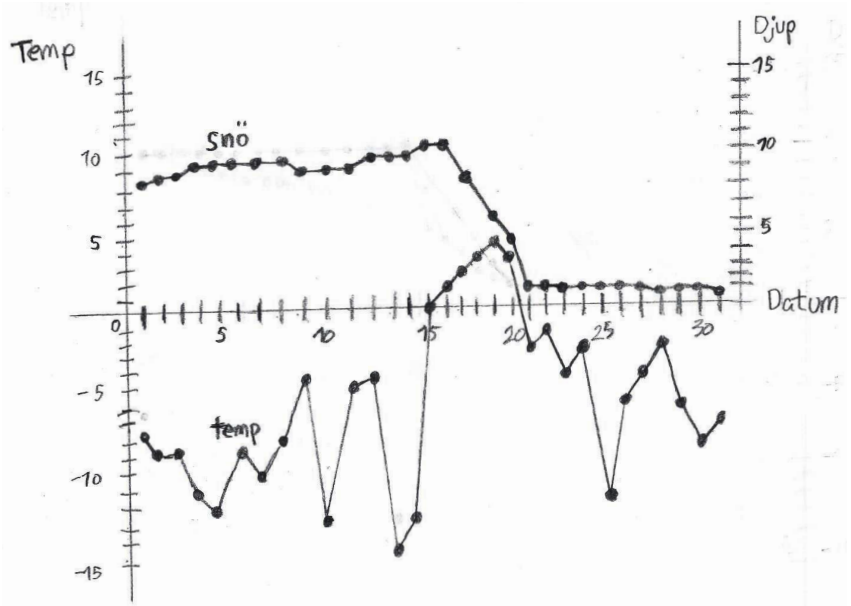
Outline

- A bit of history
- Features of long longitudinal data modeling
- Regression analysis: A smoking cessation example
- Growth/trend analysis
- Longitudinal factor analysis
- Current activities
  - Two-part, two-level longitudinal analysis
  - Modeling cycles by sine-cosine
  - Very long longitudinal data

This page intentionally left blank.

Temperature and Snow Depth
Bivariate Time-Series Data with a Lagged Effect
- Implications for Sledding





This page intentionally left blank.

Department of Statistics, Uppsala University, Early 70's

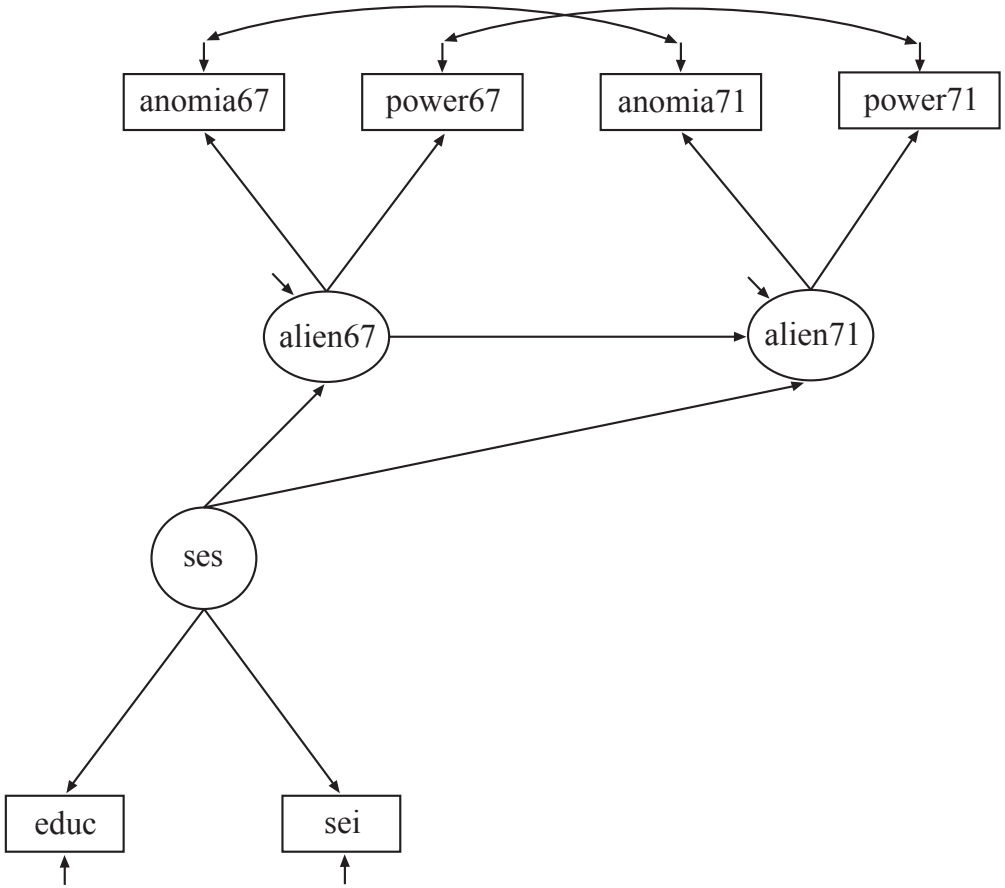
- Bengt's grad school term project related to time-series analysis:
  - Repeated measurements on respiratory problems of 7 dogs
  - Fortran program for ML estimation with autoregressive and heteroscedastic residuals

UPPSALA
UNIVERSITET

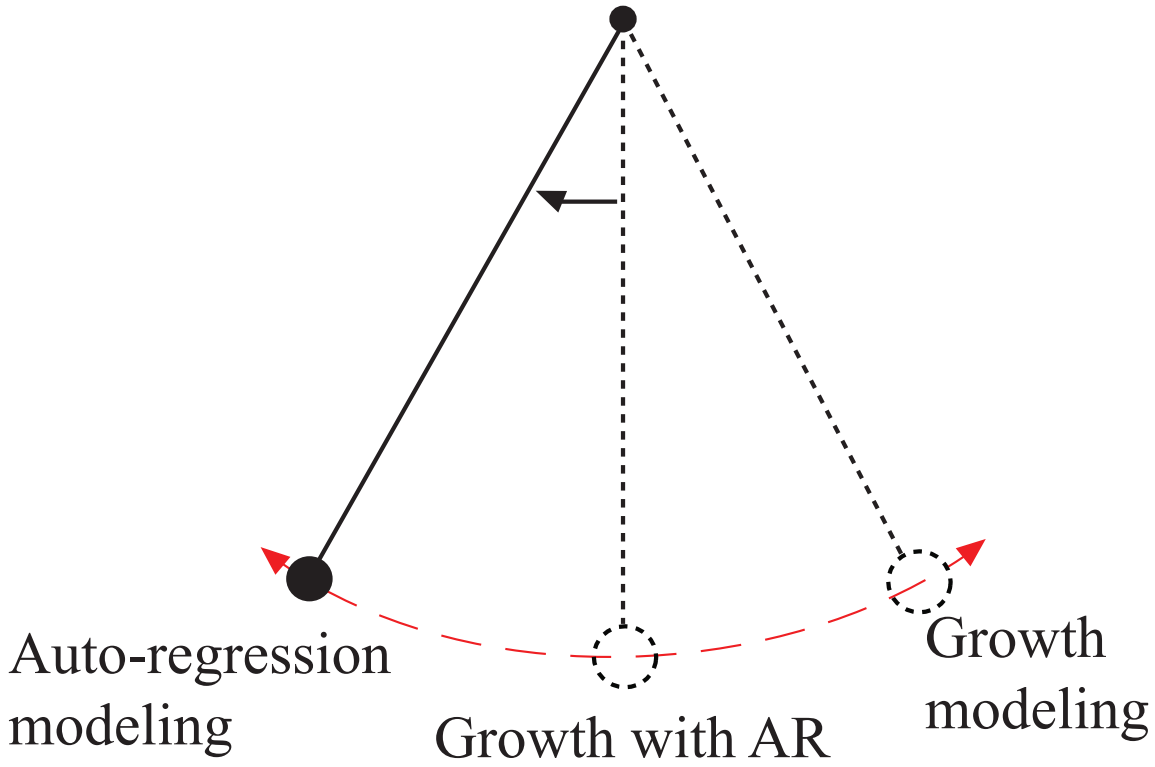
University of Wisconsin - Madison 1974-75

Wheaton, B., Muthén, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), Sociological Methodology 1977 (pp. 84 - 136). San Francisco: Jossey-Bass, Inc.

Wheaton et al. 1977

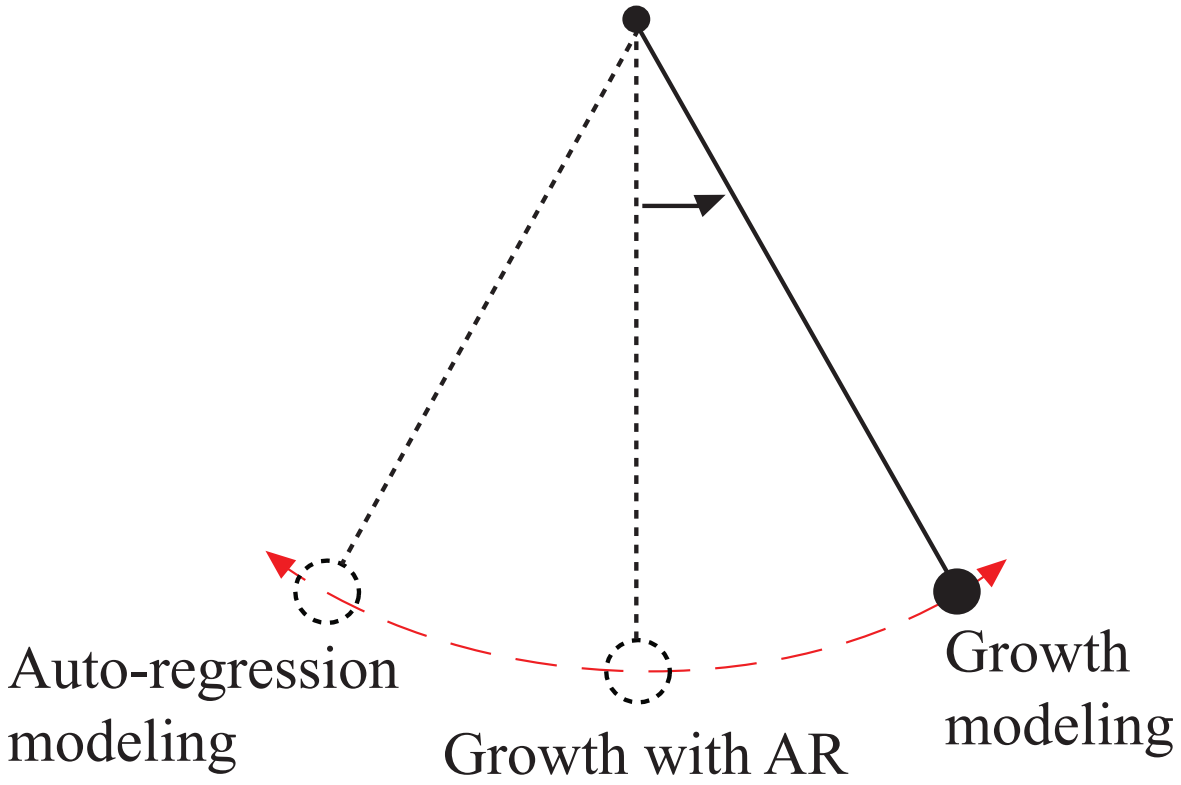


The Pendulum of Longitudinal Modeling

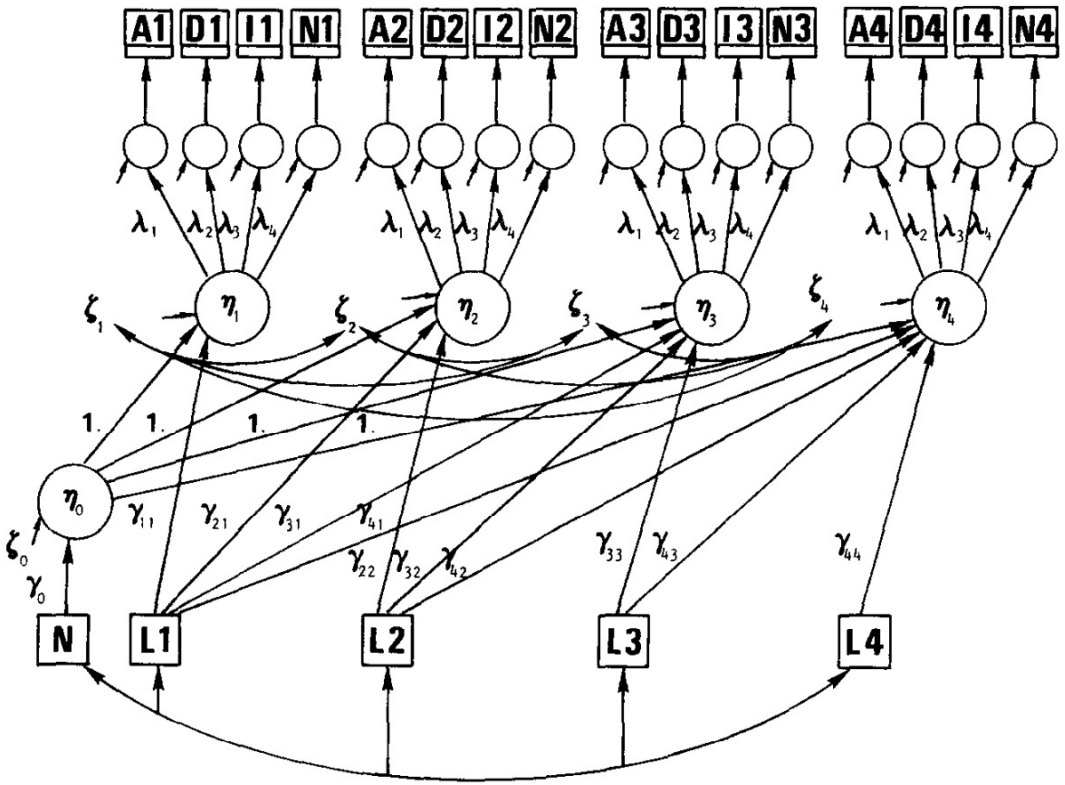


This page intentionally left blank.

The Pendulum of Longitudinal Modeling



UCLA Early 80's



Muthén (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 43-65.

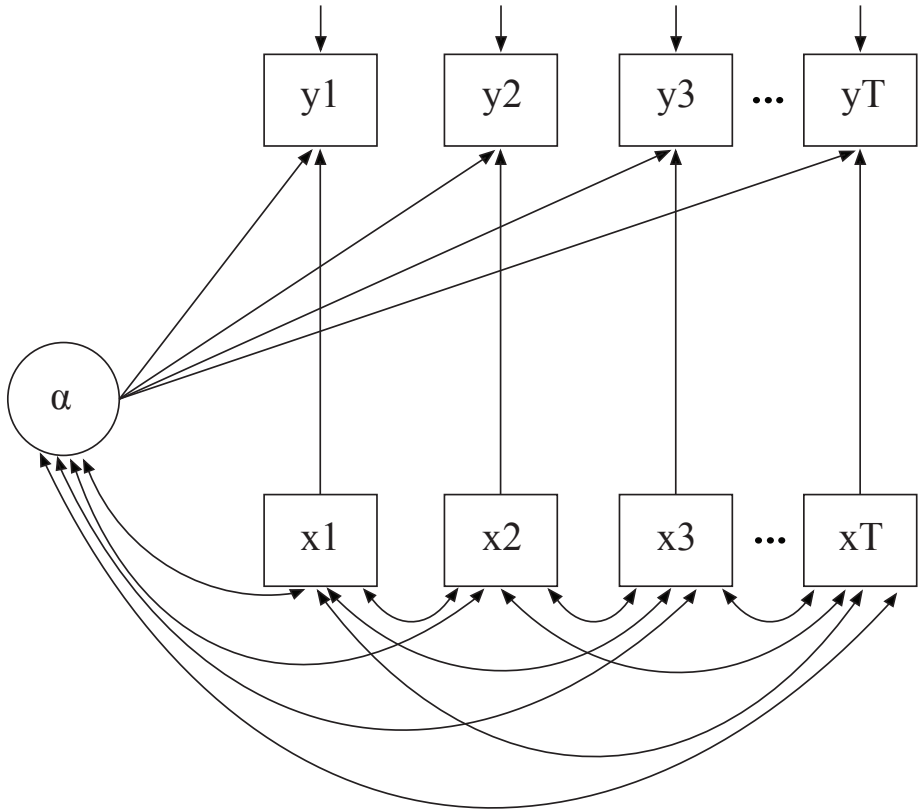
Muthén (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Longitudinal IRT Contributions 30+ Years Later

- Muthén & Asparouhov (2016). Multi-dimensional, multi-level, and multi-timepoint item response modeling. In van der Linden, Handbook of Item Response Theory. Volume One. Models, pp. 527-539. Boca Raton: CRC Press
- FAQ at www.statmodel.com: "Estimator choices with categorical outcomes": WLSMV, ML, and Bayes
- Asparouhov & Muthén (2016). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In Harring, Stapleton, & Beretvas, (Eds.), Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications (pp. 163-192). Charlotte, NC: Information Age Publishing, Inc

Fixed Effect vs Random Effect Debate for Panel Data

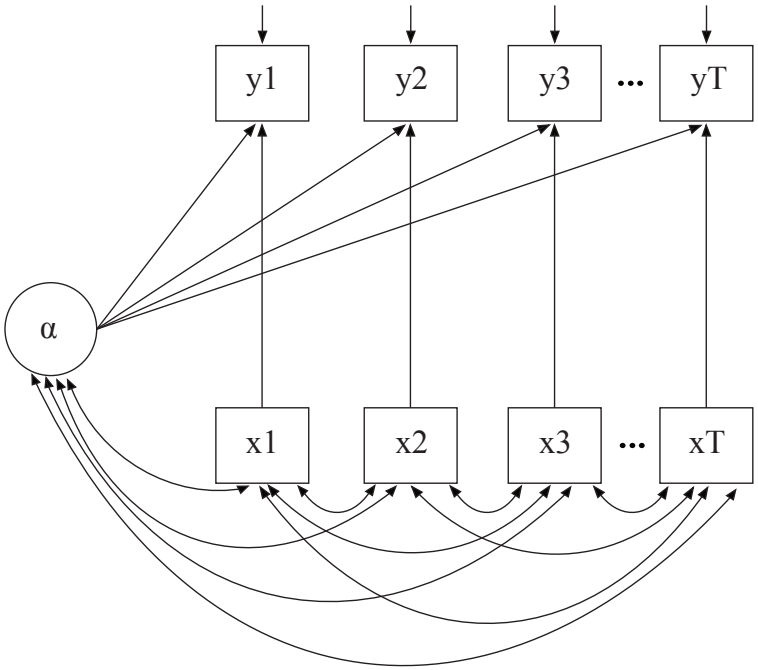
$$y_{it} = \beta_0 + \alpha_i + \beta x_{it} + e_{it}$$



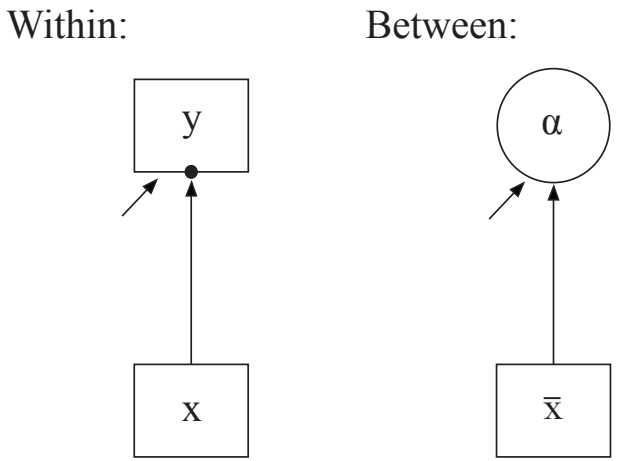
Fixed Effect vs Random Effect Debate: Equivalent Models

$$y_{it} = \beta_0 + \alpha_i + \beta x_{it} + e_{it}$$

Single-level, wide representation:

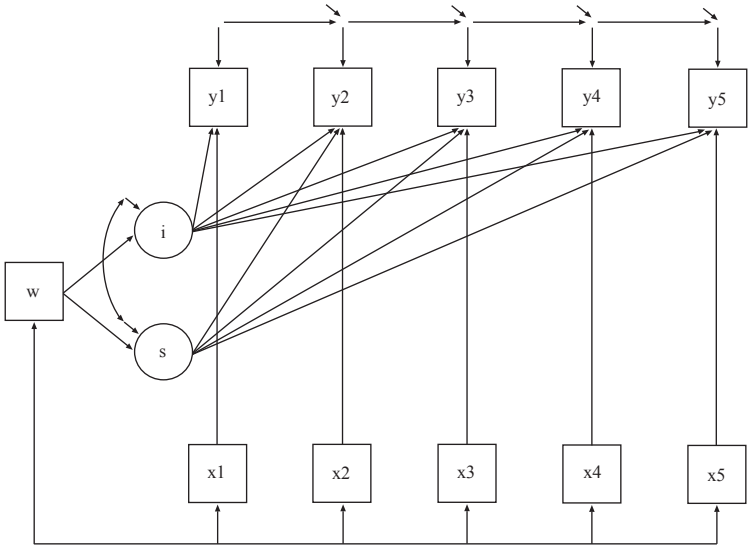


Two-level, long representation:

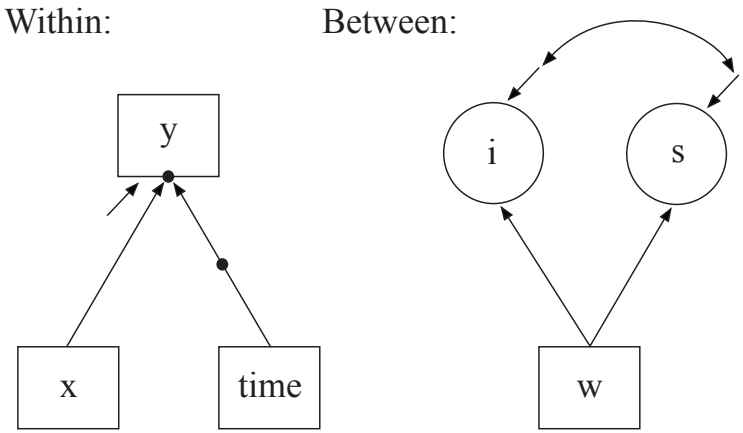


Hamaker & Muthén (2018). The fixed versus random effects debate and how it relates to centering in multilevel modeling. Submitted.

Single-level, wide representation:



Two-level, long representation:



Growth Mixture Modeling for Short Longitudinal Data

- Muthén & Shedden (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics, 55, 463-469.
- Muthén, Asparouhov, Hunter & Leuchter (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. Psychological Methods, 16, 17 - 33
- Muthén & Asparouhov (2015). Growth mixture modeling with non-normal distributions. Statistics in Medicine, 34:6, 1041 1058.
  - Allows for a non-normal within-class distribution using skew-t (using 2 more parameters than the normal distribution). BMI data

Survival Mixture Analysis

- Discrete-time survival:
  - Muthén & Masyn (2005). Discrete-time survival mixture analysis. Journal of Educational and Behavioral Statistics, 30, 27-58
- Continuous-time survival:
  - Asparouhov, Masyn & Muthén (2006). Continuous time survival in latent variable models. Proceedings of the Joint Statistical Meeting in Seattle, August 2006. ASA section on Biometrics, 180-187
  - Muthén, Asparouhov, Boye, Hackshaw & Naegeli (2009). Applications of continuous-time survival in latent variable models for the analysis of oncology randomized clinical trial data using Mplus

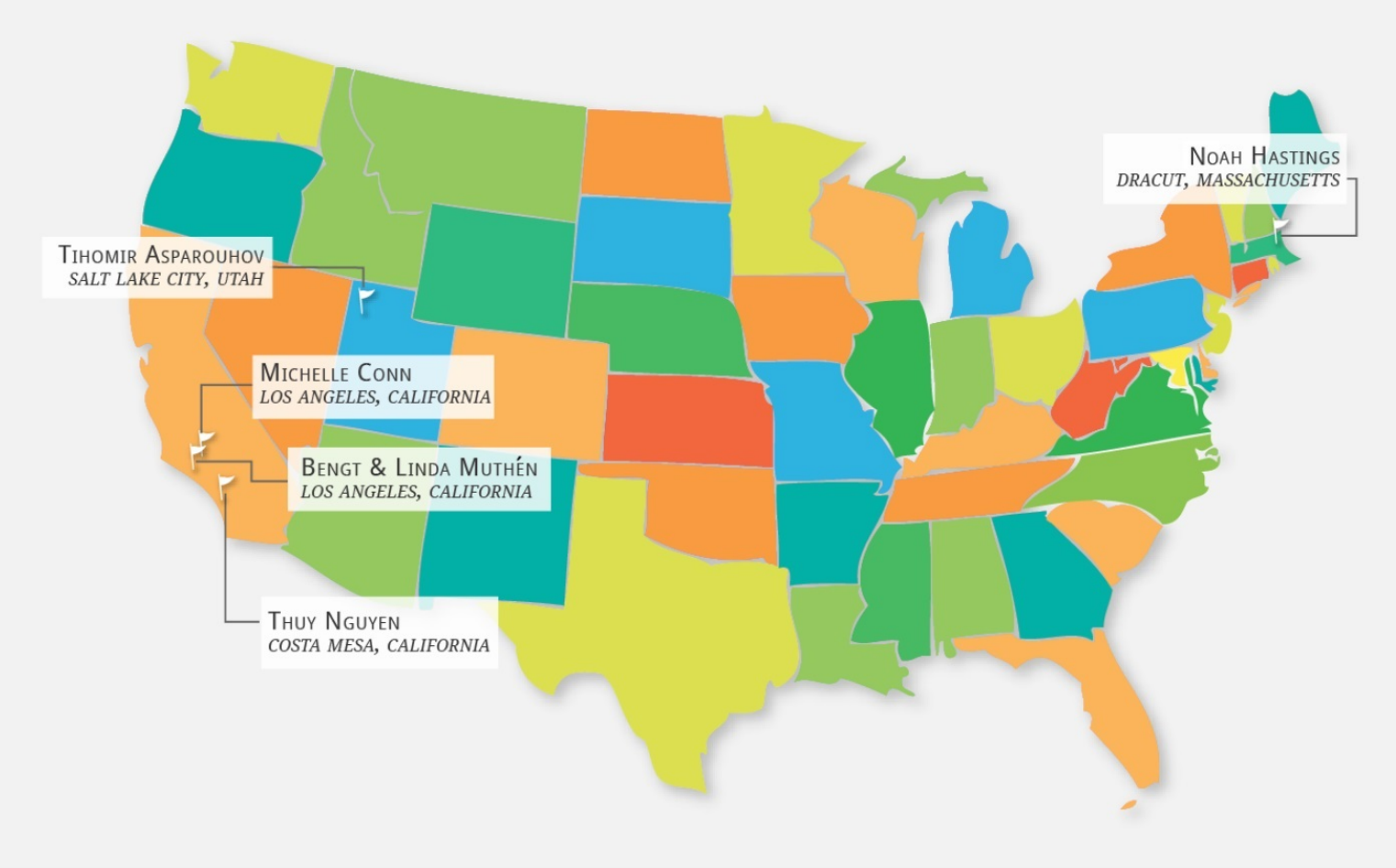
The General Latent Variable Framework of Mplus

- 1995 NIH SBIR: 1998 launch of Mplus
- Merging continuous and categorical latent variables
  - Continuous latent variables:
    - Factors measured by multiple indicators, random effects, frailties, liabilities, latent response variables with missing
  - Categorical latent variables:
    - Latent classes, finite mixtures, latent response variable categories with missing data
- General SEM structure on each of multiple levels for continuous, categorical, count, and censored observed variables
- Estimation by WLS, ML, and Bayes
- Different model types freely combined

General Latent Variable Modeling: Integration of a Multitude of Analyses

- Exploratory factor analysis
- Structural equation modeling
- Item response theory analysis
- Growth modeling
- Latent class analysis
- Latent transition analysis (Hidden Markov modeling)
- Growth mixture modeling
- Survival analysis
- Missing data modeling
- Multilevel analysis
- Complex survey data analysis
- Causal inference
- Time series analysis

The Mplus Team: 4 + 2



Overview papers

- Muthén (2002). Beyond SEM: General latent variable modeling. Behaviormetrika, 29, 81-117
- Muthén (2008). Latent variable hybrids: Overview of old and new models. In Hancock & Samuelsen (Eds.), Advances in latent variable mixture models, pp. 1-24. Information Age Publishing, Inc
- Muthén & Asparouhov (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, Davidian, Verbeke & Molenberghs, G. (eds.), Longitudinal Data Analysis, pp. 143-165. Boca Raton: Chapman & Hall/CRC Press.

Outline

- A bit of history
- **Features of long longitudinal data modeling**
- Regression analysis: A smoking cessation example
- Growth/trend analysis
- Longitudinal factor analysis
- Current activities

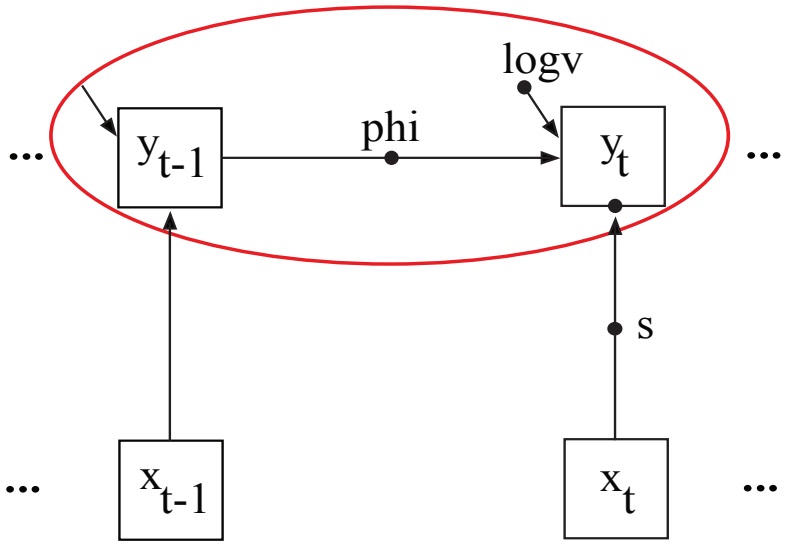
Intensive Longitudinal Data: EMA, ESM, AA, Diary

- Many time points: T = 50 - 1000
  - Wide format analysis not feasible/practical
  - Long format analysis needed
- Measurements are typically close in time
  - Random effects are not sufficient to represent correlation across time for subjects
  - Auto-regression needed as well
- Two-level and cross-classified time series analysis (Dynamic SEM; DSEM, RDSEM)
  - General SEM structure on each level
  - Random effects for intercepts, slopes, ARs, and variances
  - As T and N increase, increasingly more flexible models can be estimated
  - Bayesian estimation needed

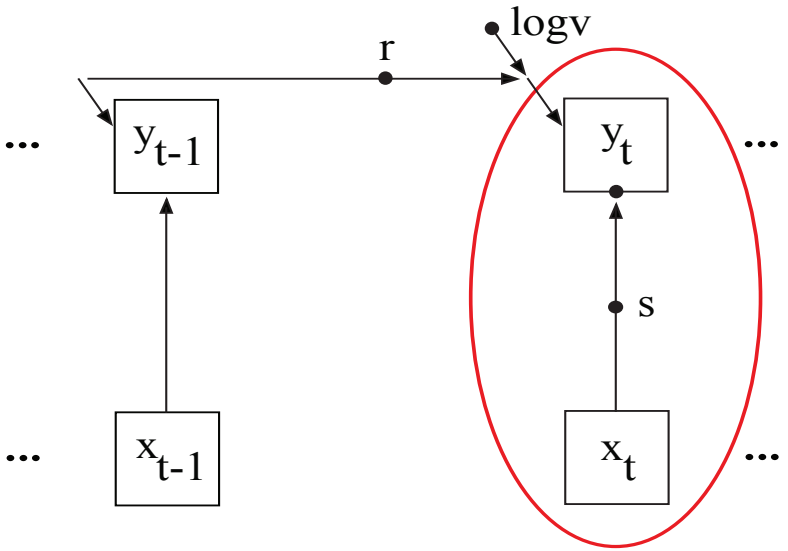
Our Recent Papers on Long Longitudinal Data Analysis: Posted at statmodel.com/TimeSeries with Topic 12-13 Videos

- Asparouhov, Hamaker & Muthén (2017). Dynamic latent class analysis. Structural Equation Modeling, 24:2, 257-269
- Asparouhov, Hamaker & Muthén (2018). Dynamic structural equation models. Structural Equation Modeling, 25:3, 359-388
- Hamaker, Asparouhov, Brose, Schmiedek & Muthén (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. Multivariate Behavioral Research
- Schultzberg & Muthén (2018). Number of subjects and time points needed for multilevel time series analysis: A simulation study of dynamic structural equation modeling. Structural Equation Modeling, 25:4, 495-515
- Asparouhov & Muthén (2018). Latent variable centering of predictors and mediators in multilevel and time-series models. Accepted for publication in Structural Equation Modeling
- Asparouhov & Muthén (2018). Comparison of models for the analysis of intensive longitudinal data. Submitted for publication

Auto-Regression for the Outcome or the Residual?



- **Focus on y_t regressed on y_{t-1}**
- Hamaker et al., MBR (2018): Daily measurements of negative and positive affect over 100 days
- Autoregressive parameter indicating "how quickly a person restores equilibrium after being perturbed": inertia
- Time series tradition (our term: DSEM)



- **Focus on y_t regressed on x_t**
- Liu & West, J. Personality (2015): Daily diary study over 60 days
- Stress during the day influencing alcohol consumption that evening
- Multilevel tradition (our term: RDSEM)

Outline

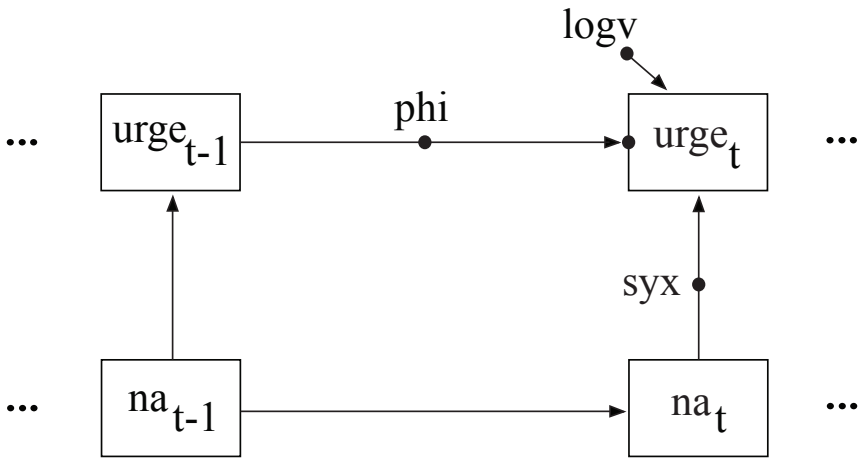
- A bit of history
- Features of long longitudinal data modeling
- **Regression analysis: A smoking cessation example**
- Growth/trend analysis
- Longitudinal factor analysis
- Current activities

EMA Example: Smoking Urge Data

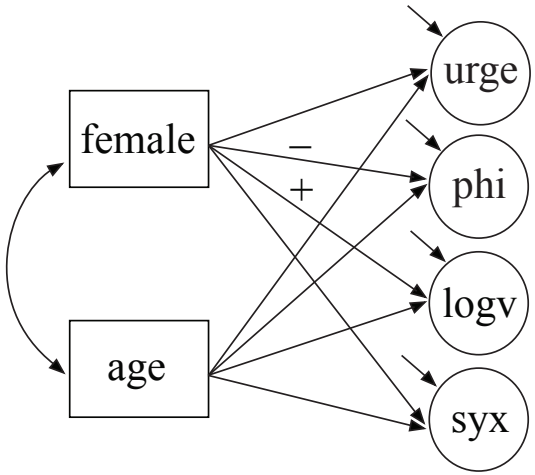
- Shiffman smoking cessation data
- N = 230, T $\approx$ 150: Random prompts from Personal Digital Assistant (hand held PC) approx. 5 times per day for a month
- Variables: Smoking urge (0-10 scale), negative affect (unhappy, irritable, miserable, tense, discontent, frustrated-angry, sad), gender, age, quit/relapse

Two-Level Time Series Analysis: Regression of Smoking Urge on Negative Affect (na) Using 4 Random Effects

Within:



Between:



Bayesian Analysis: Advantages over ML

- Bayes with non-informative priors - a powerful computing algorithm :
  - Analyses are often less computationally demanding, for example, when maximum-likelihood requires high-dimensional numerical integration due to many latent variables (factors, random effects)
  - In cases where maximum-likelihood computations are prohibitive, Bayes with non-informative priors can be viewed as a computing algorithm that would give essentially the same results as maximum-likelihood if maximum-likelihood estimation were computationally feasible
    - New types of models can be analyzed where the maximum-likelihood approach is not feasible (e.g. multilevel time series models with many random effects)
- Bayes with informative parameter priors - a better reflection of hypotheses based on previous studies

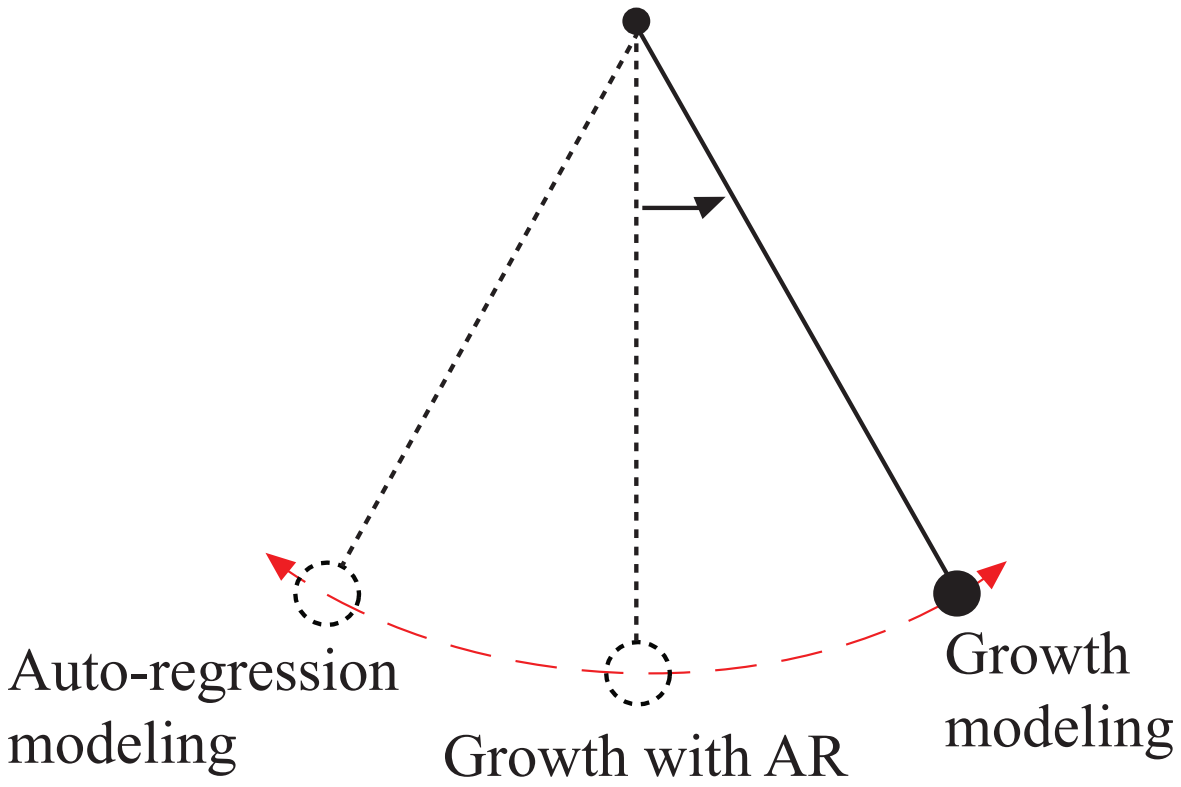
Bayesian Analysis

- Learning Bayesian analysis in the early 90's via mixture modeling using BUGS - too slow, switched to ML
- Arminger & Muthén (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. Psychometrika, 63, 271-300
- Muthén & Asparouhov (2012). Bayesian SEM: A more flexible representation of substantive theory. Psychological Methods, 17, 313-335
  - Zero-mean, small-variance priors instead of fixed zeros
  - Strictness of the factor model hypothesis: EFA $<$ BSEM $<$ CFA
  - Reduces the CFA risk of over-estimating factor correlations
  - Produces a counterpart to ML's modification indices (Lagrange multipliers)
- Bayes papers posted at statmodel.com/papers: Bayesian Analysis

Outline

- A bit of history
- Features of long longitudinal data modeling
- Regression analysis: A smoking cessation example
- **Growth/trend analysis**
- Longitudinal factor analysis
- Current activities

The Pendulum of Longitudinal Modeling



Cross-Classified Longitudinal Analysis

- Two between-level cluster variables: subject crossed with time (one observation for a given subject at a given time point)
- Generalization of the two-level model providing more flexibility: random effects can vary across not only subject but also across time

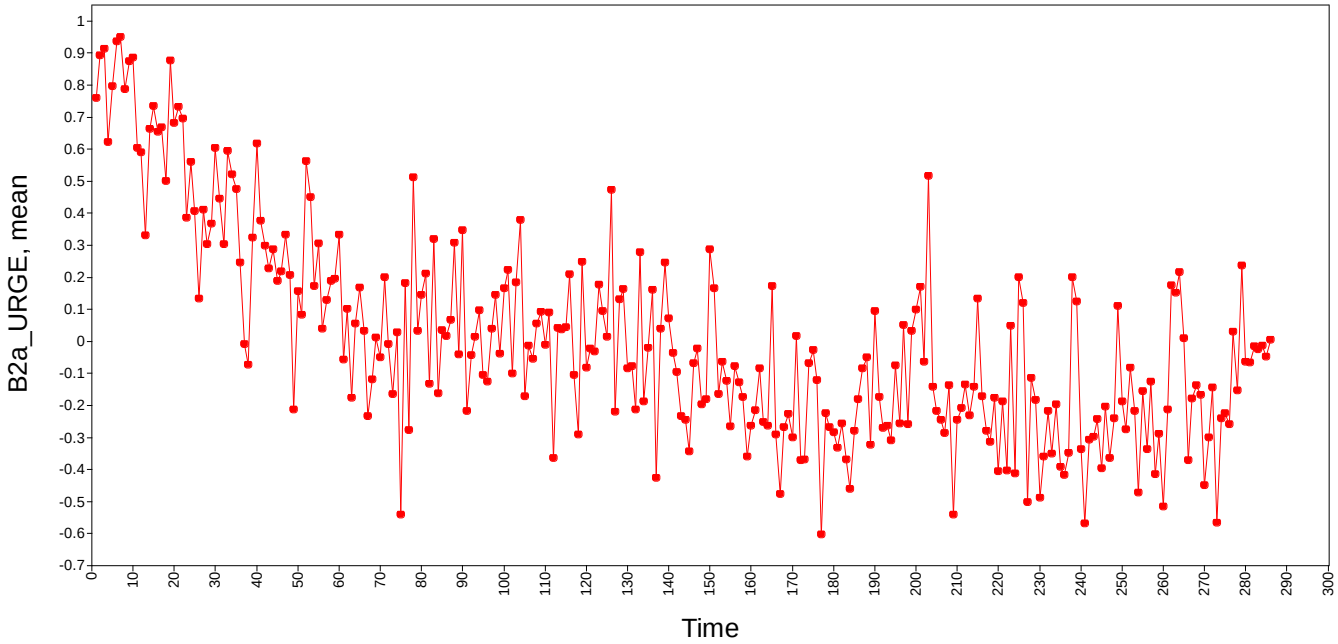
Consider the two-level model with a random intercept/mean:

$$y_{it} = \underbrace{\alpha + \alpha_i}_{\text{Between subject}} + \underbrace{\beta\, y_{w,it-1} + \varepsilon_{it}}_{\text{Within subject}}. \quad (1)$$

The corresponding cross-classified model is:

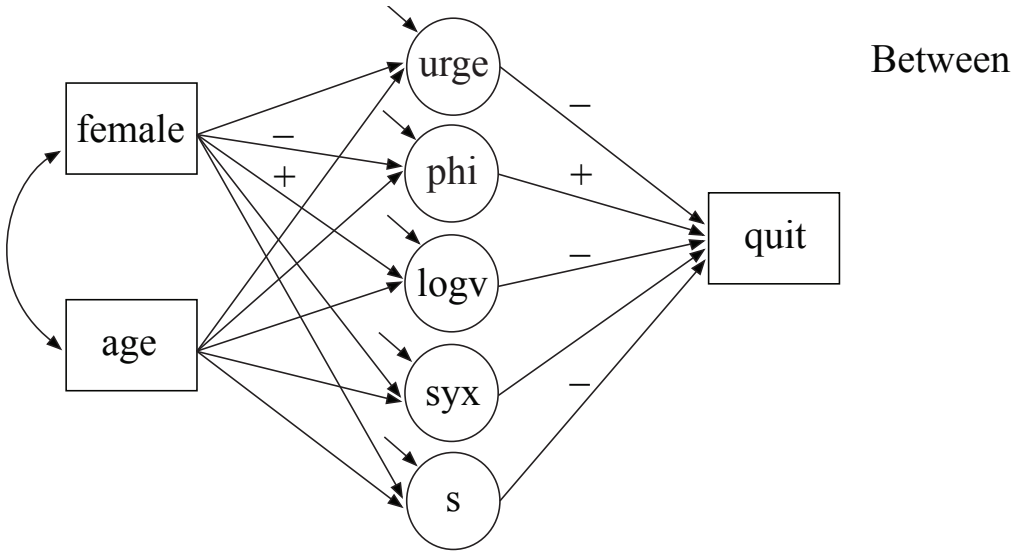
$$y_{it} = \underbrace{\alpha + \alpha_i}_{\text{Between subject}} + \underbrace{\alpha_t}_{\text{Between time}} + \underbrace{\beta\, y_{w,it-1} + \varepsilon_{it}}_{\text{Within subject}}. \quad (2)$$

Cross-Classified Analysis: Finding a Trend
Plot of Time-Varying Random Effects for Smoking Urge



- Analysis used cross-classified modeling
- The trend can be modeled according to some functional form
  - In a cross-classified analysis
  - In a two-level analysis

Smoking Urge Data: Two-Level Analysis Adding a Trend for Urge and a Binary Dependent Variable on Between



Quit (binary) regressed on random effects:

- higher urge gives lower quit probability
- higher autocorrelation gives higher quit probability
- higher residual variance gives lower quit probability
- higher trend slope gives lower quit probability

Counterfactually-Defined Indirect Effects

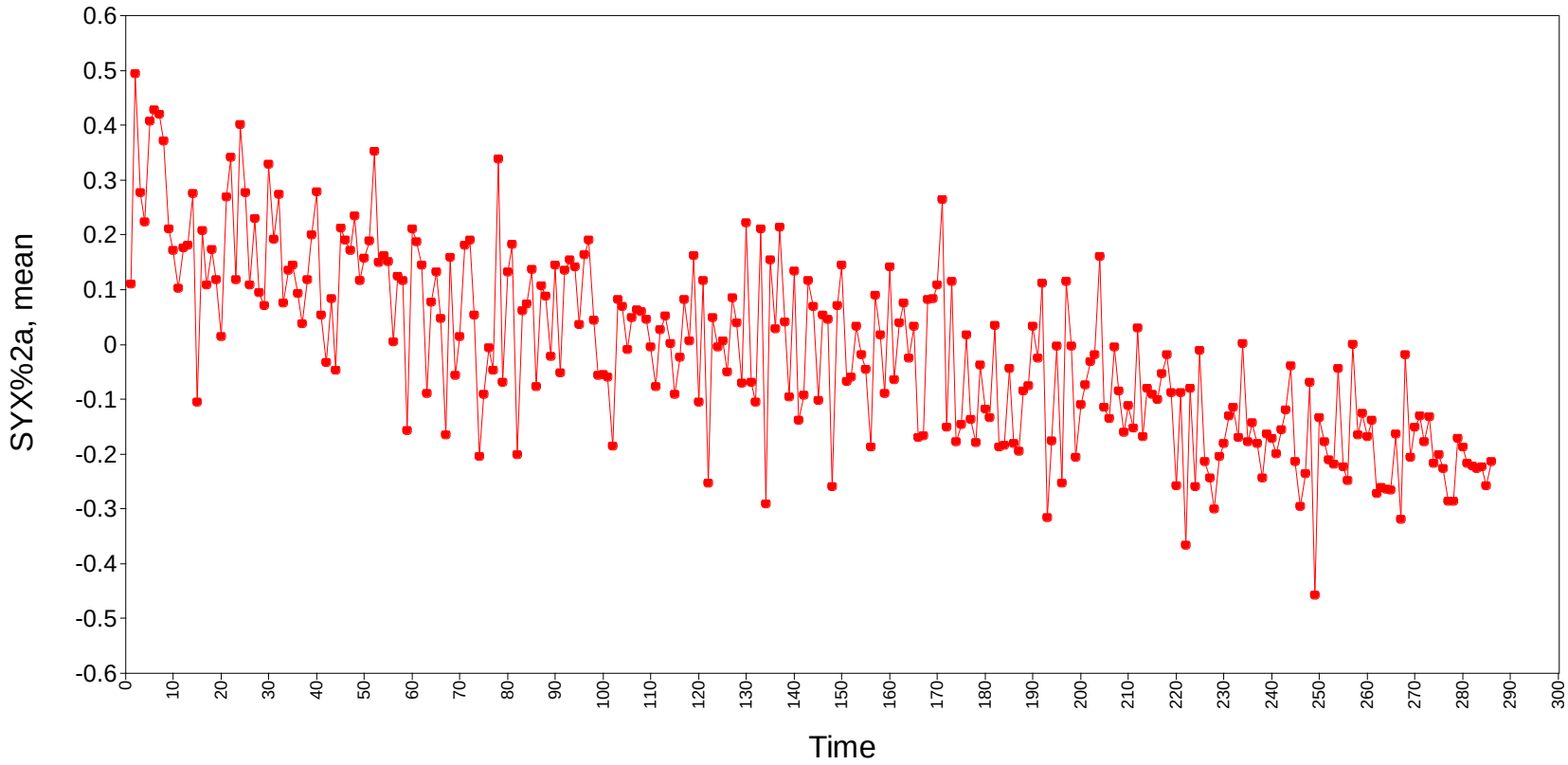
- The binary quit outcome calls for counterfactually-defined (causal) indirect and direct effects
- Muthén & Asparouhov (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. Structural Equation Modeling: A Multidisciplinary Journal, 22(1), 12-23
- Muthén, B., Muthén, L. & Asparouhov, T. (2016). Regression And Mediation Analysis Using Mplus. - Chapters 4 and 8
- In this example, mediation is on level 2 with multiple latent mediators

Cross-Classified Analysis: Time-Varying Effect Modeling (TVEM)

- Cross-classified modeling allows parameters to change over time
- An example is a regression slope
  - Does the influence of negative affect on smoking urge show a decline over time?

Trend in Slope for Urge Regressed on Negative Affect

- syx is the slope of urge regressed on negative affect



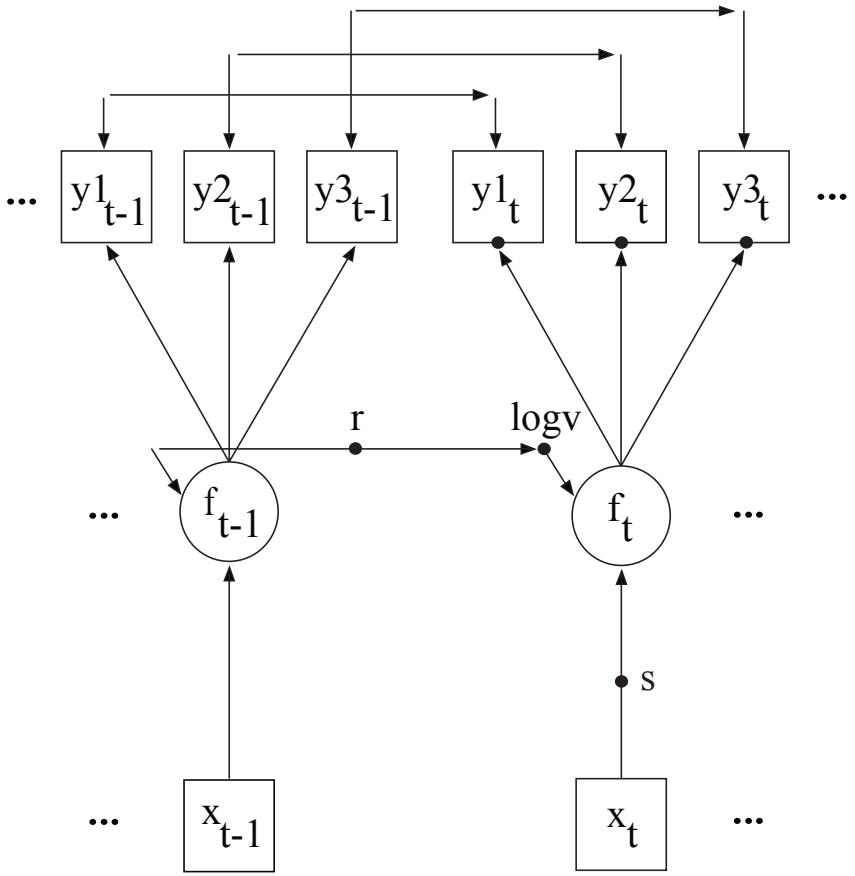
- syx regressed on time gives a significant negative slope: The effect of negative affect on smoking urge is reduced over time

Outline

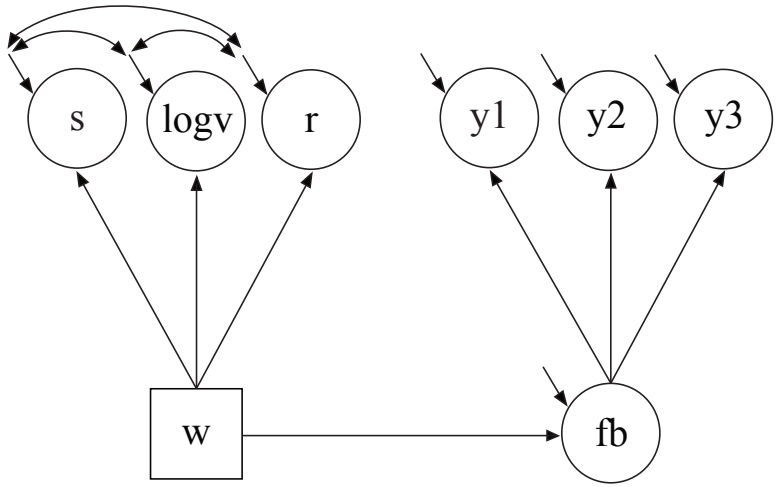
- A bit of history
- Features of long longitudinal data modeling
- Regression analysis: A smoking cessation example
- Growth/trend analysis
- **Longitudinal factor analysis**
- Current activities

Two-Level Time Series Factor Analysis

Within:



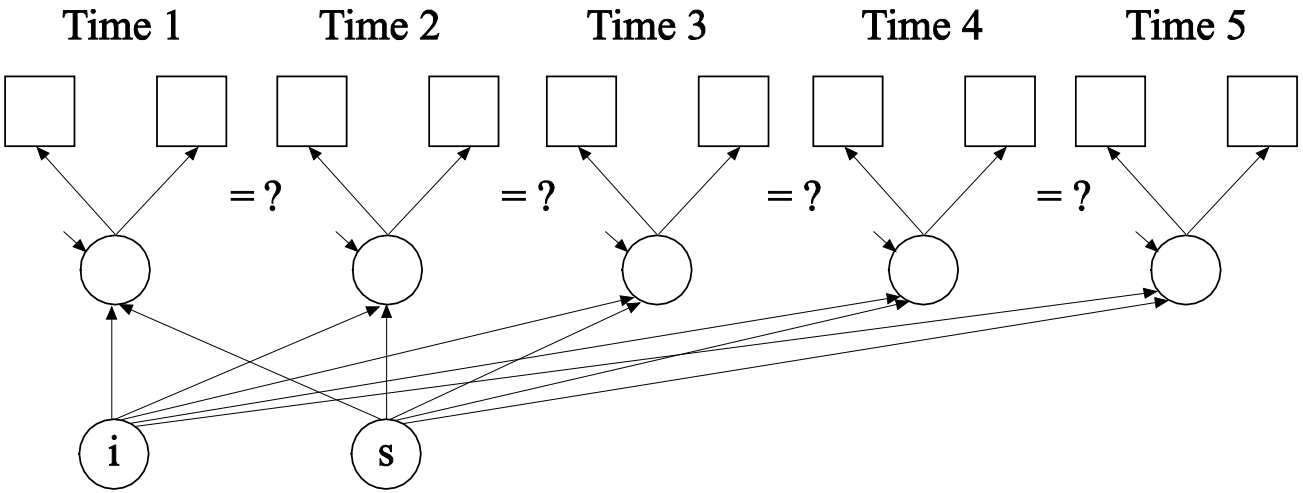
Between:



Measurement Invariance in Longitudinal Factor Analysis

- An old dilemma
- A new solution

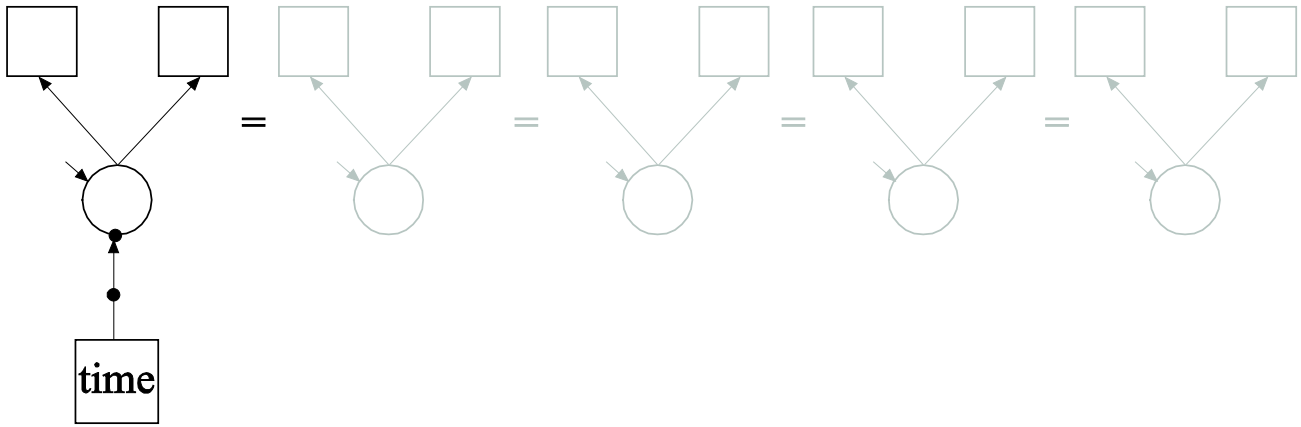
Categorical Items, Single-Level, Wide Format Approach



Single-level analysis with $p \times T = 2 \times 5 = 10$ variables, $T = 5$ factors.

- ML hard and impossible as T increases (numerical integration)
- WLSMV possible but hard when $p \times T$ increases and biased unless attrition is MCAR or multiple imputation is done first
- Bayes possible
- Searching for partial measurement invariance is cumbersome

Categorical Items, Two-Level, Long Format Approach



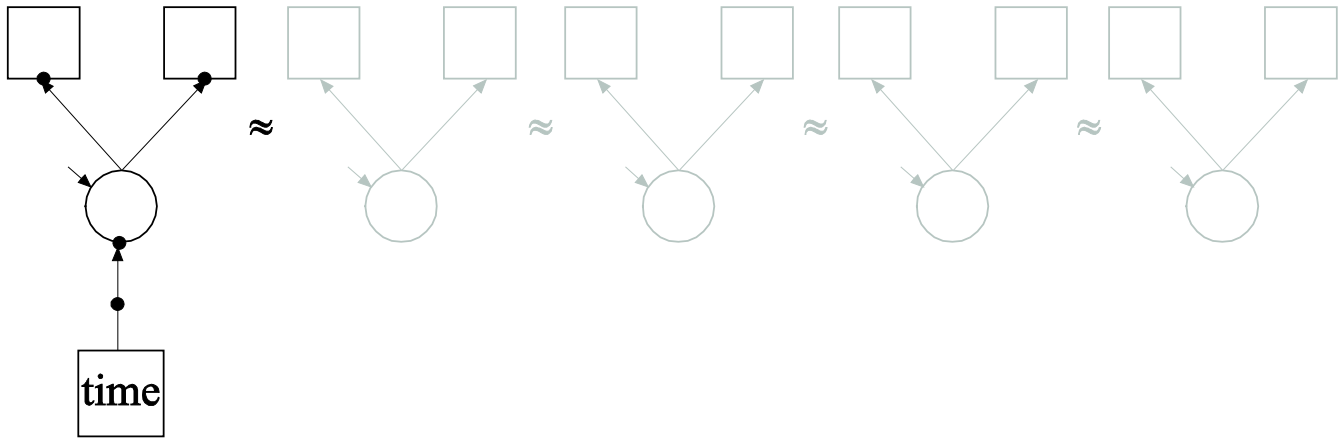
Two-level analysis with $p = 2$ variables, 1 within-factor, 2-between factors, **assuming full measurement invariance across time**.

- ML feasible
- WLSMV feasible (2-level WLSMV)
- Bayes feasible

Measurement Invariance Across Time

- New solution, time is a random mode
- A long format, cross-classified approach
  - Best of both worlds: Keeping the limited number of variables of the two-level approach without having to assume full measurement invariance across time

Cross-Classified, Long Format Approach



- Clusters are person and time
- Bayes cross-classified random effects analysis with random measurement intercepts varying across person **and time**

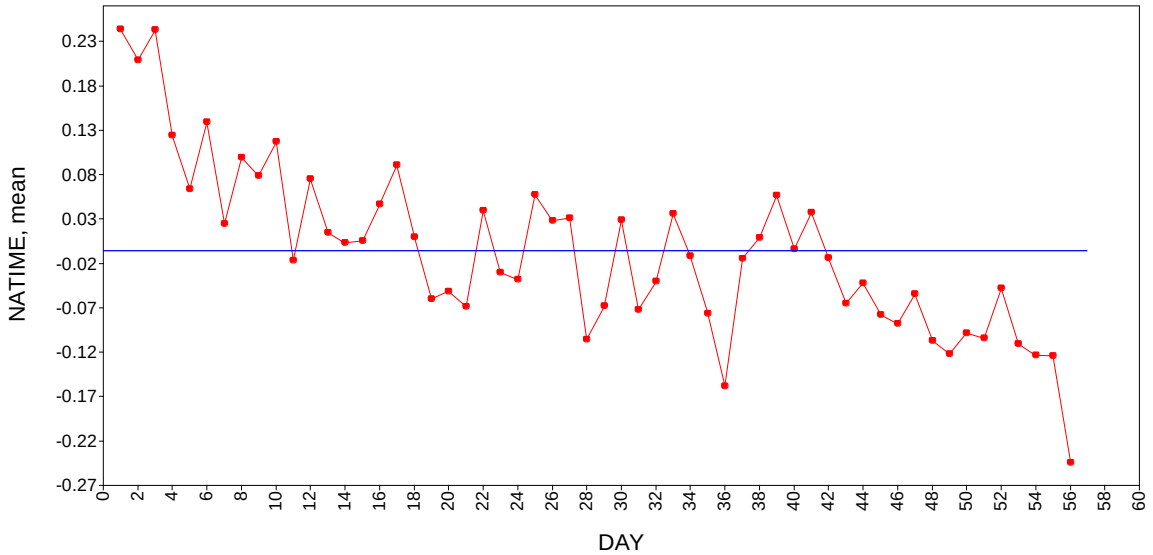
Example: Item Factor Analysis (IRT) Using 10 Negative Affect Items

- Data from the older cohort of the Notre Dame Study of Health & Well-being (Bergeman): N = 270, T = 56 (daily measures on consecutive days)
- Wang, Hamaker, Bergeman (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*
- Predictors and distal outcomes of negative affect development over the 56 days
- 10 NA items (5-cat scale): afraid, ashamed, guilty, hostile, scared, upset, irritable, jittery, nervous, distressed (average score used in article). Wide format would have 56*10 variables
- Question format: Today I felt... (1 = Not at all, ..., 5 = Extremely)
- 1-factor DAFS model with ordinal factor indicators

Results of Cross-Classified Factor Analysis with One NA Factor for 10 Ordinal Items

$$na\,factor_{it} = \underbrace{\alpha + \alpha_i}_{\text{Between subject}} + \underbrace{\alpha_t}_{\text{Between time}} + \underbrace{\beta\, y_{w,it-1} + \varepsilon_{it}}_{\text{Within subject}}.$$

- The factor score plot for the na_time factor (on the between day level) shows a drop of 40% of the total factor SD over the 56 days:



Outline

- A bit of history
- Features of long longitudinal data modeling
- Regression analysis: A smoking cessation example
- Growth/trend analysis
- Longitudinal factor analysis
- **Current activities**

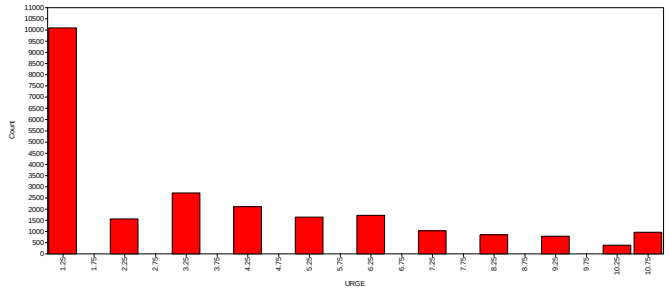
This page intentionally left blank

Current Activities

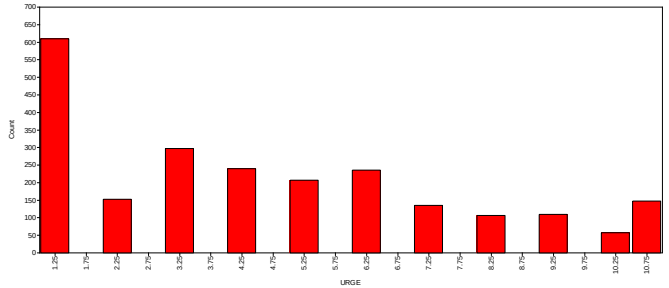
- Two-part, two-level longitudinal analysis
- Modeling cycles by sine-cosine
- Very long longitudinal data

Long Longitudinal Analysis with Strong Floor Effects

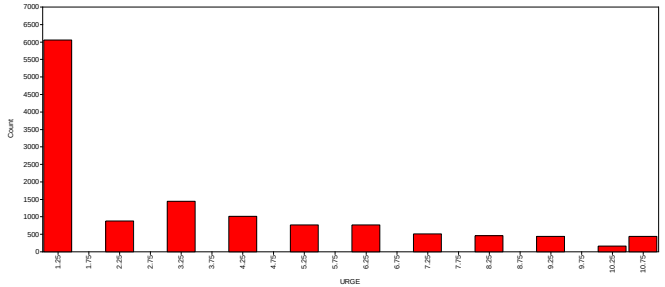
Overall: 42% at the floor value (smoking urge in cessation study)



Early: 27% at the floor value



Late: 47% at the floor value



Two-Part Modeling of Floor Effects

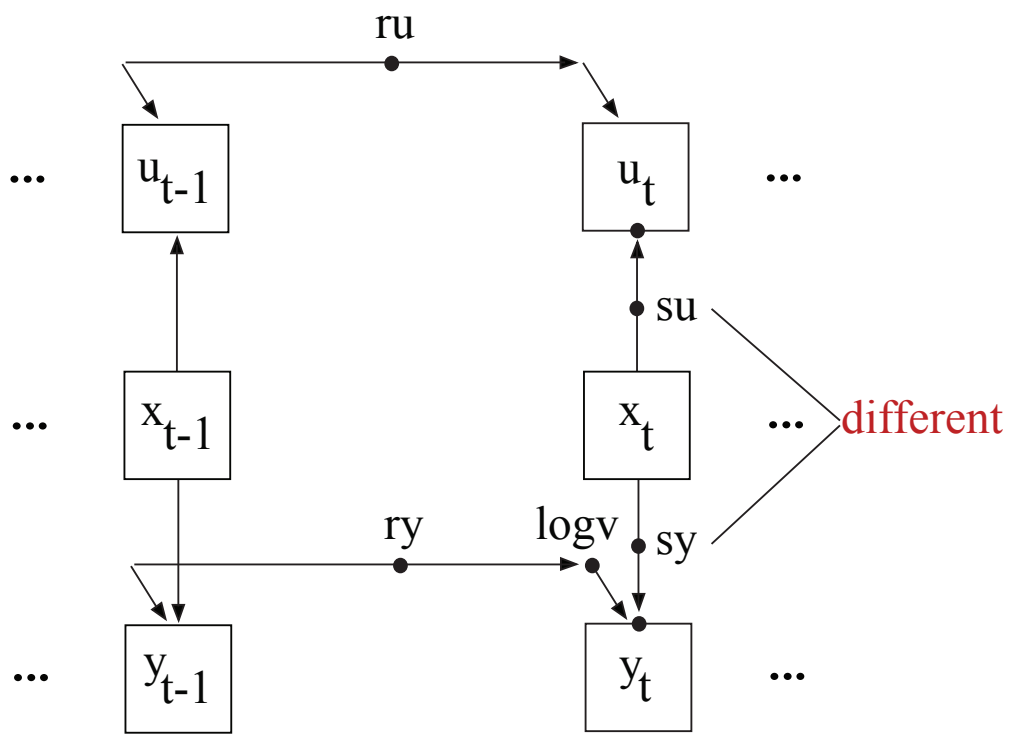
- Olsen & Schafer (2001). Two-part random-effects model for semicontinuous longitudinal data. JASA
- Kim & Muthén (2009). Two-part factor mixture modeling: Application to an aggressive behavior measurement instrument. Structural Equation Modeling, 16, 602-624
- Categorical DSEM and RDSEM: Asparouhov, Hamaker & Muthén (2018). Dynamic structural equation models. Structural Equation Modeling, 25:3, 359-388

- Transform the variable into 2 variables:
  - - A binary u and a continuous y (DATA TWOPART)

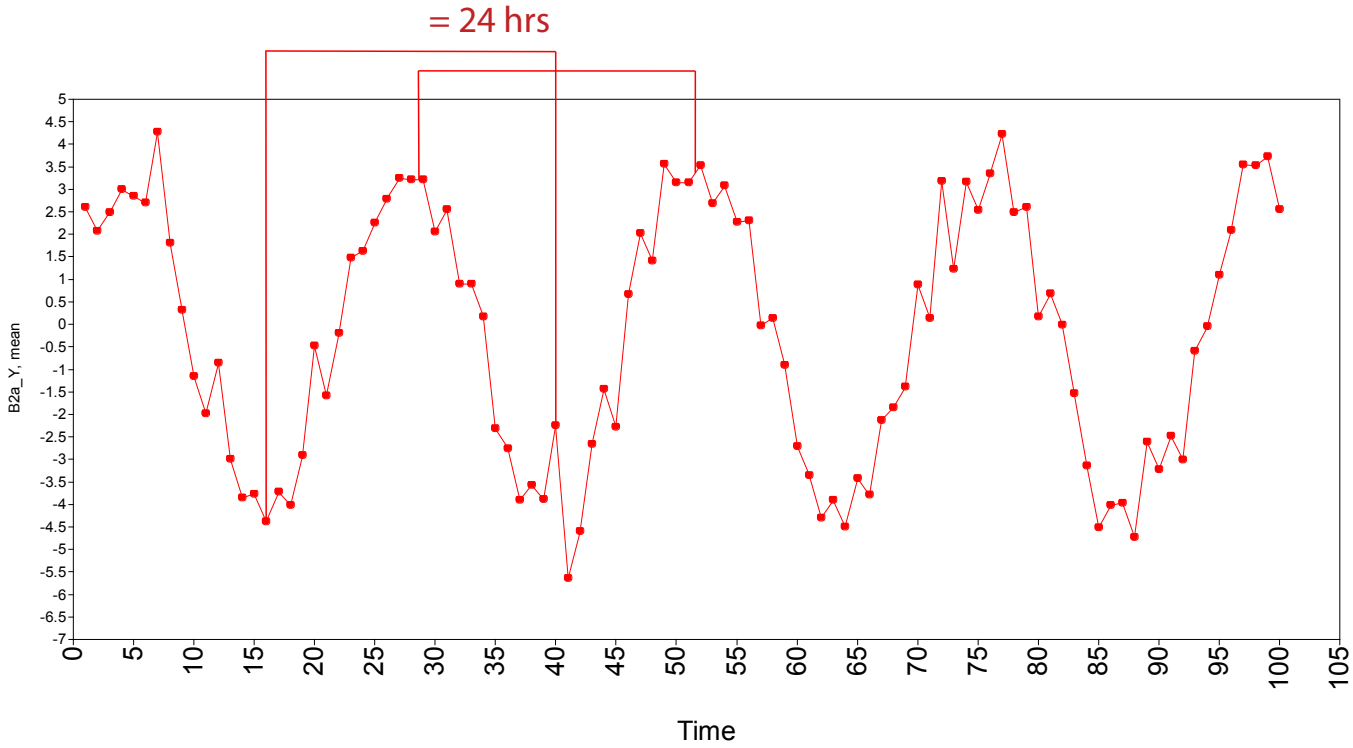
- u = 0 if at the floor: y is missing
- u = 1 if not at the floor: y is observed

- Probit model for u
- Log normal model for y

Two-Part, Two-Level Regression Modeling Binary and Continuous Outcome



Modeling Cycles: Dummies, Splines, Sine-Cos, Free Form



- Biological cycles
  - 24-hour cycles: Circadian rhythm such as heart rate
- Behavioral cycles
  - Weekly drinking pattern
- Environmental cycles
  - Monthly temperature fluctuations

Cyclic Formulas

$$f(t) = A\ cos\ (2\pi\ \omega\ t + \phi)$$
$$= \underbrace{-A\ sin\ \phi}_{\beta_1}\ \underbrace{sin\ (2\pi\ \omega\ t)}_{x_1} + \underbrace{A\ cos\ \phi}_{\beta_2}\ \underbrace{cos\ (2\pi\ \omega\ t)}_{x_2}$$

$$Amplitude = A = \sqrt{\beta_1^2 + \beta_2^2}$$
$$Phase = \phi = tan^{-1}(-\beta_1/\beta_2)$$

- ω is a frequency index defined as cycles per unit. Using $\omega = 1/24 = 0.04167$ gives 24-hour cycles
- Multiple $f(t)$ components with different cycles per unit can be used. Spectral analysis finds the components of the cycles
- Two-level or cross-classified analysis with random effects

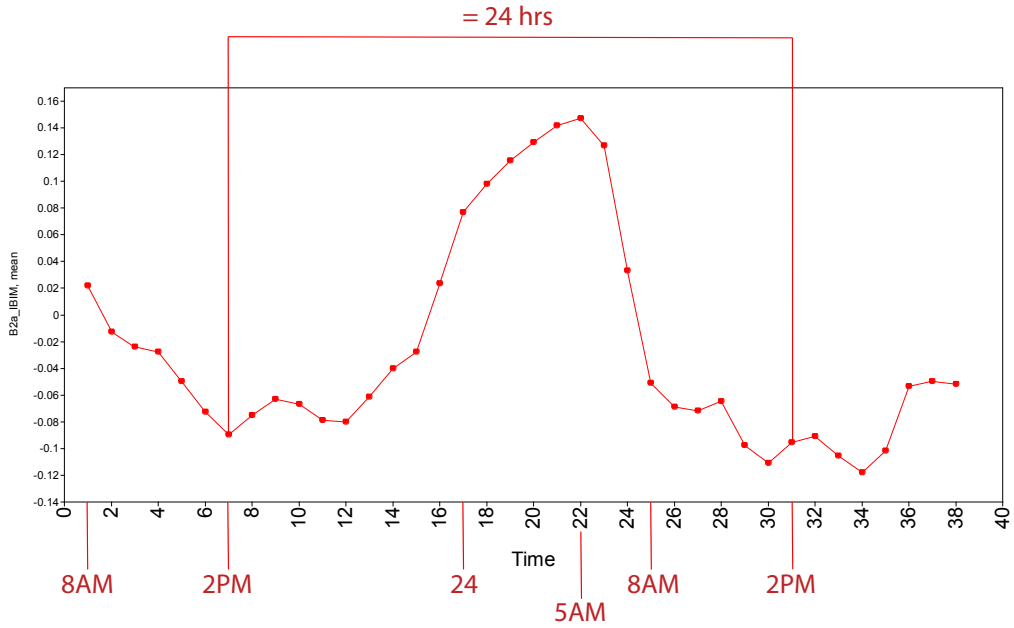
Example: Circadian Analysis of Heart Beat Data

- Data:
  - N = 162, T = 38 (hourly measures used here)
  - Outcome: ibi (time in between heart beats - high is good)
  - Covariates: Gender, smoking, sports
- Model:
  - Single-component two-level with random effects

Background:

Houtveen, Hamaker, van Doornen (2010). Using multilevel path analysis in analyzing 24-h ambulatory physiological recordings applied to medically unexplained symptoms. Psychophysiology, 47, 570-578.

Estimated In-Between Heart Beats Random Effects



- Slopes for sine and cosine both significant with little variation across subjects
- Random effects across subjects for residual variance and auto-regression:
  - Exercise increases time in between heart beats and increases residual variance
  - Smoking decreases time in between heart beats, decreases residual variance, and increases auto-regression

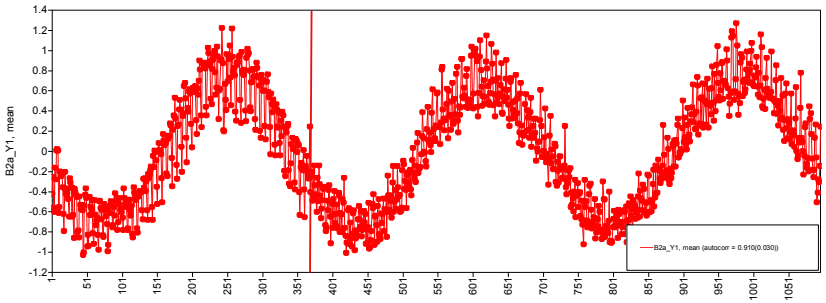
Cyclic Modeling: Two-Level Model as a Measurement Instrument for N=1 Analysis

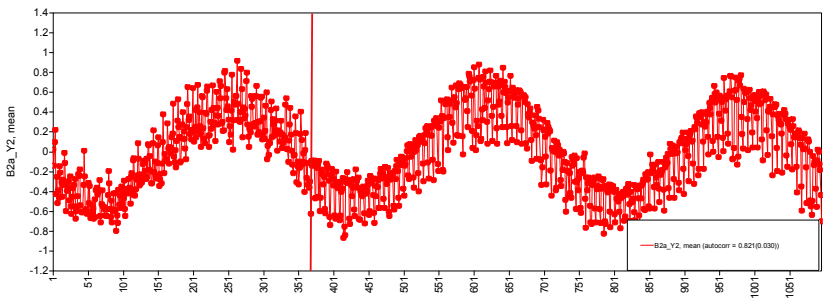
- Ambulatory measurement of blood pressure
- T = 48: Every 30 minutes for 24 hours
- N = 886
- 2-component sine-cosine model with 4 random slopes
- Madden et al. (2018). Morning surge in blood pressure using a random-effects multiple-component cosinor model. Statistics in Medicine.

- Problem: How do you estimate an individual's curve for this complex model from only T = 48?
- Potential solution: Do a two-level analysis with N=1

Very Long Longitudinal Data: T= 1096

- Electricity consumption of firms measured daily (and hourly) over 3 years: T=1096 (Schultzberg, 2018)
  - Intervention: change in tariff
  - N=184 intervention group (N= 800 Control group; not used here)
  - Pre-intervention data for 1 year, post-intervention data for 2 years
  - Sine-cosine cross-classified model





- Significant drop in amplitude after the intervention (marked by a vertical line)
- In the left part of the figure, the curve after the intervention shows the predicted development in the absence of the intervention

Very Long Longitudinal Data

Schultzberg & Muthén (2018). Number of subjects and time points needed for multilevel time series analysis: A simulation study of dynamic structural equation modeling. Structural Equation Modeling. 25:4, 495-515

This page intentionally left blank.