APPLIED PUBLIC HEALTH STATISTICS SECTION

HUDSON COLLEGE OF PUBLIC HEALTH
The UNIVERSITY of OKLAHOMA HEALTH SCIENCES CENTER

2018 APHS Breakfast Workshop, Supplemental Materials

Summer G. Frank-Pearce & Trent L. Lalonde

Research Goals, Questions, and Hypotheses for Intensive Longitudinal Studies

What is This About? The goals, hypotheses, research questions, or general intentions help guide a project from design to data collection to analysis to conclusions. While Intensive Longitudinal Data (ILD) may have inherent characteristics that cause the datasets to differ from those of traditional longitudinal studies, it is the research goals that truly differentiate ILD studies from those of more traditional longitudinal methods.

Why do we Care? The specific goals of a study will help determine the way the data are collected, the types of statistical analyses applied, and even the language that is appropriate when writing conclusions. Therefore it is essential to have a clear idea of the types of research questions that are appropriate for ILD and that can be of specific interest with ILD methods.

What Should we Know? While traditional longitudinal studies often focus on changes or growth over time, ILD studies tend *not* to have a focus on individual changes over long periods of time. Instead, ILD studies often take advantage of the high-intensity of observations to address the following types of research goals.

- Periodic Patterns: ILD studies can effectively address questions about recurring or short-term periodic patterns. For example, an ILD study could be used to determine the nature of the daily patterns of tobacco craving among adults who are heavy users of tobacco.

  EXAMPLE: How can we describe the daily pattern of tobacco craving for heavy users?

  Hypothesis: The daily pattern of tobacco craving for heavy users will be cyclic, with peaks in the early morning and mid-afternoon, and low points late-morning and evening.

- Synchronicity: ILD studies can be used to assess questions about in-the-moment associations. For example, an ILD study could be used

to quantify the momentary associations between immediate desire for caffeine and mood.

EXAMPLE: Is there an association between in-the-moment desire for caffeine and in-the-moment mood?

Hypothesis: There is a negative association between in-the-moment desire for caffeine and in-the-moment mood.

- Sequentiality: ILD studies are often used to answer questions about short-term antecedents or sequelae. For example, an ILD study could be used to determine whether the social context of one moment is associated with greater marijuana use four hours later.

  EXAMPLE: Can an individual's report of social context (number of other other people) at one moment effectively predict the frequency of marijuana use at a follow-up prompt four hours later?

  Hypothesis: There is a positive association between the number of people present at one moment and the frequency of marijuana use reported at the following prompt four hours later.

- Variability: ILD studies have sufficiently dense collections of observations to address interests about the volatility of an outcome of interest. For example, an ILD study could examine whether the fluctuations in alcohol cravings differ between sub-populations of men and women.

  EXAMPLE: Is there a difference in the fluctuations of alcohol craving between male and female populations?

  Hypothesis: The female population tends to show greater fluctuation in alcohol craving than the male population.

Selected References:

- Bolger, N. & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research.* New York, N.Y.: The Guilford Press.

- Walls, T. A. & Schafer J. L. (2006). *Models for intensive longitudinal data*. New York, N.Y.: Oxford University Press.
- Stone, A. A., Shiffman, S., Atienza, A. A., & Nebeling, L. (2007). *The science of real-time data capture: Self-reports in health research*. New York, N.Y.: Oxford University Press.

Data Collection and Interventions for Intensive Longitudinal Studies

What is this About? It is important to connect the goals of any research study to the methods of data collection. For Intensive Longitudinal Data (ILD) studies there are a number of aspects of data collection to consider that are not usually of concern with traditional longitudinal studies.

Why do we Care? Collecting data in a way that is not consistent with study goals can result in an inability to make the desired conclusions.

What Should we Know? Beyond standard issues of sampling and experimental design, it is important to consider the following factors in data collection for ILD studies.

- Baseline Data: Baseline data can be collected to account for potential person-level confounders in statistical modeling of ILD. Measures collected at baseline are often time-independent, meaning that the value will not change over the course of the study (e.g., race/ethnicity). These variables can be used to identify specific sub-populations of individuals.
- Event-Based Data: Within ILD studies, individuals are often asked to initiate a survey or data-collection process when they experience a specific event. For example, individuals who smoke are often asked to initiate and complete a series of questions each time a cigarette is used. In this situation it is important for the researchers to have a *very clear* definition of the "event" and to be able to train individuals to identify this event. In addition, researchers should be prepared for individuals to "miss" or fail to report events during data collection. Thus, compliance with respect to event recording cannot be assessed.
- Signal-Based Data: Within ILD studies, researchers often prompt individuals to provide responses, either according to a pre-set schedule (fixed-interval) or randomly within "windows" of time. For example, participants may be randomly prompted three times per day to report on workplace atmosphere and mood. Fixed-interval schedules can lead to bias when a participant becomes accustomed to the assessment schedule

and begins anticipating assessments. The collection of data at random times within a day is intended to "surprise" an individual and to reduce the potential for bias. Questions for signal-based data collection may be different from those of event-based data collection.

- Automated Data Collection: Certain types of data can be collected automatically, without initiation from researchers or participants. For example, pharmaceutical containers can be equipped with RFID (Radio Frequency IDentification) technology to record each time a pill bottle is opened, thus providing timestamps of medication use.

- Daily Diary: While daily diaries represent one of the original types of ecological momentary assessment data, many studies still involve a single "end-of-day" report on variables of interest. It is important to recognize that daily diaries require participants to aggregate their own data by asking them to summarize quantitative and qualitative aspects of their experience over a day. For instance, an individual may be asked to report the number of times in the day an event happened, to characterize the intensity of the event, or to report both frequency and intensity. Reporting of less salient, i.e. more ordinary, events may be remembered less accurately, which may lead to bias.

- Combination: It is common for ILD studies to combine both event-based and signal-based data collection. In this case there will usually be a combination of questions identical for both event- and signal-based surveys, and a combination of questions that differ. For example, both event- and signal-based surveys may ask about momentary mood, while only signal-based prompts may include questions about tobacco craving.

- Just-in-Time Adaptive Interventions: A more recent development in ILD methods, the Just-in-Time Adaptive Intervention (JITAI) provides a process for responding to participants' responses with targeted, individualized intervention strategies that may improve outcomes over time. JITAI aim to provide the right amount of intervention or support at the right time and to reduce participant burden or fatigue by not providing support when it is not needed. Through consistent monitoring, they use dynamically changing information about an individual to identify opportunities to provide intervention in the natural environment.

Selected References:

- Stone, A. A., Shiffman, S., Atienza, A. A., & Nebeling, L. (2007). *The science of real-time data capture: Self-reports in health research.* New York, N.Y.: Oxford University Press.

- Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA (2018). Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine*, 52(6):446-462. doi: *10.1007/s12160-016-9830-8.* PubMed PMID: 27663578; PubMed Central PMCID: PMC5364076.

- Goldstein SP, Evans BC, Flack D, Juarascio A, Manasse S, Zhang F, Forman EM (2017). Return of the JITAI: Applying a Just-in-Time Adaptive Intervention Framework to the Development of m-Health Solutions for Addictive Behaviors. *International Journal of Behavioral Medicine*, 24(5):673-682. doi: *10.1007/s12529-016-9627-y.* PubMed PMID: 28083725; PubMed Central PMCID: PMC5870794.

- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1-32.

- Connor, T. S. & Lehman, B. J. (2014). Getting started: Launching a study in daily life. In M. R. Mehl & T. S. Connor (Eds.), *Handbook of research methods for studying daily life.* New York, N.Y.: Guilford Press.

Data Visualization and Exploration for Intensive Longitudinal Studies

What is This About? Exploring and visualizing data is an important first step in any data analysis. Intensive Longitudinal Data (ILD) studies are no different.

Why do we Care? Data exploration and visualization can provide important context for model results and interpretations, and can help identify issues with the data that may cause model failure. But exploration **should not** be used to change the goals or hypotheses of the ILD study.

What Should we Know? Many of the descriptive methods used for ILD studies are also applied in traditional longitudinal data analyses. The following methods are quite common.

- Data Descriptives: For individual variables, provide measures of the central tendency (e.g., means and variances) for continuous data and measures of distribution (proportions and percentages) for categorical data. To illustrate pairwise relationships, provide scatter plots and cross-tabulated summaries. It is important to minimize the amount of averaging or aggregating *across* individuals. This can be misleading because it ignores the repeated observations of individuals.
- Distributions by Individuals: It is often helpful to summarize data aggregated by individual. For example, exploring the distribution of the *total* or *mean* number of times individuals used cigarettes over the course of an ILD study can be investigated through means, histograms, box plots, etc.
- Time Plots: Time plots generally show all observations either over *calendar time* (actual dates and times) or over *gap times* (the amount of time that has passed in the study). In both cases the data analyst is generally looking for changes in typical values over time.
- Spaghetti Plots: Spaghetti plots are constructed as time plots, except observations are connected by subject. Each participant in the study will

have a line representing the trend of outcomes over the course of time. These plots can be extremely helpful in identifying trends over time, variation over the course of the study, and whether there are consistent versus changing patterns of variation for different sub-populations. For example, males may show a great amount of volatility in desire to smoke cigarettes, while females show a more consistent trend in desire.

- Variograms: A variogram very generally plots residuals from a non-correlated model versus gap times in a study. Because the inherent auto-correlation in the data is ignored, patterns in variograms can give an indication of the nature of the auto-correlation in the data. For example, a variogram showing a generally "flat" trend can indicate an exchangeable or compound symmetry type of correlation in which all outcomes are correlated equally for any individual.

Selected References:

- Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. *Analysis of longitudinal data*. New York, N. Y.: Oxford University Press.
- Hedeker, D. & Gibbons, R. D. *Longitudinal data analysis*. New York, N.Y.: Wiley-Interscience.
- Bolger, N. & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, N.Y.: The Guilford Press.
- Walls, T. A. & Schafer J. L. (2006). *Models for intensive longitudinal data*. New York, N.Y.: Oxford University Press.

Continuous Outcome Multilevel Model: Model Formulation and Parameter Interpretations

What is This About? Intensive Longitudinal Data (ILD) are often analyzed using multilevel models, in which both variation within individuals (over time) and between individuals (across different sub-populations) are accounted for in the model. Such models have also been referred to as hierarchical models, HLM, random-effects models, and mixed-effects models.

Why do we Care? Multilevel modeling is extremely common in the ILD literature. It is important to be able to recognize common language used to reference model terms and conclusions, and to be able to make appropriate types of interpretations with our own ILD studies.

What Should we Know? Multilevel modeling is used commonly for traditional longitudinal data analysis, but the following components are common across both types of studies. In general, we are interested in modeling the mean of a continuous outcome in terms of baseline covariates (independent of time), ILD covariates (changing over time), time, a term that accounts for the fluctuation of (or variability in) outcomes among the population of interest, and a term that accounts for the fluctuation of outcomes over time for individuals.

Dependent Variable	~	Baseline Variables	+	ILD Variables	+	Time Trends	+	Variation *Across* Individuals	+	Variation *Within* Individuals

More specifically, baseline variables are included as in a typical regression model. On the other hand, ILD variables are often included as a decomposition of "between" effects (differences across individuals) and "within" effects (differences within individuals over time). Time trends can be included as a simple time term, a polynomial time trend with multiple terms, or possibly as a periodic time term represented by sine and cosine terms (depending on researcher interest and intuition).

Variation across individuals can be represented by a random subject effect,

called a *random intercept*, which allows for variation in the intercept term among the population of individuals represented in the study. Variation across individuals can be further represented by interactions of additional random effects with independent variables, called *random slopes*, which allow for variation in predictor effects among the population of individuals represented in the study. Finally, variation within individuals is represented by a standard error term. Consider the following normal multilevel model equation,

$$Y_{it} = \beta_0 + \beta_1 z_i + \beta_{2B}\bar{x}_{i.} + \beta_{2W}(\bar{x}_{i.} - x_{it}) + \beta_t t_{it} + u_i + \epsilon_{it},$$

where:

- Y_{it} represents the continuous outcome of interest for individual i at time t (conditional on the random effect u_i),
- β_0 is the intercept term,
- z_i is a baseline (time-independent) predictor with effect β_1 (also commonly referred to as the coefficient, parameter estimate, or slope),
- x_{it} is an ILD (time-dependent) predictor with between effect β_{2B} and within effect β_{2W},
- t_{it} represents the time in the study with effect β_t,
- u_i is the random individual effect, and
- ϵ_{it} is the random variation over time.

In terms of interpretations, consider an interest in modeling the continuous outcome of alcohol craving.

- β_1 represents the expected difference in the mean outcome across populations that differ, on average, by a unit change in z_i (a *time-independent* predictor). For example, if z_i represents the ages of different individuals, β_1 represents the expected difference in alcohol craving across populations that differ (on average) by one year in age.

- β_{2B} represents the expected difference in the mean outcome across populations that differ in terms of average values of x_{it} (the average values of x_{it} are *time-independent*). For example, if x_{it} represents the momentary mood of individual i at time t, β_{2B} represents the expected difference in alcohol craving across populations that differ by one unit of average mood reported.

- β_{2W} represents the expected change over time of the mean outcome as individuals change in terms of momentary values of x_{it} (the momentary values give *time-dependent* predictors). For example, β_{2W} represents the expected change over time for an individual in their mean alcohol craving as their mood changes.

- u_i represents the variation in individual alcohol craving among the population of individuals and is assumed to be distributed as a random normal variable with mean 0 and variance σ_u^2.

In addition to a random intercept, any effects in the model can be included as random slopes:

$$Y_{it} = \beta_0 + (\beta_1 + v_i)z_i + \beta_{2B}\bar{x}_{i.} + \beta_{2W}(\bar{x}_{i.} - x_{it}) + \beta_t t_{it} + u_i + \epsilon_{it},$$

where:

- v_i represents the variation in the effect of z_i among the population of individuals and is assumed to be distributed as a random normal variable with mean 0 and variance σ_v^2.

<u>Selected References:</u>

- Bolger, N. & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research.* New York, N.Y.: The Guilford Press.

- Walls, T. A. & Schafer J. L. (2006). *Models for intensive longitudinal data.* New York, N.Y.: Oxford University Press.

- Hedeker, D. Mermelstein, R. J., Berbaum, M. L., & Campbell, R. T. (2009). Modeling mood variation associated with smoking : An application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data. *Addiction*, 104(2): 297-307. doi:*10.1111/j.1360-0443.2008.02435.x*.

- Hedeker, D. Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location-scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, 64, 627-634. DOI: *10.1111/j.l541-0420.2007.00924.x*.

- Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, 31: 3328-3336. DOI: *10.1002/sim.5338*.

- Coffman, C. J., Allen, C. D., & Woolson, R. F. (2012). Mixed-effects regression modeling of real-time momentary pain assessments osteoarthritic (OA) patients. *Health Services Outcomes Research Methods*, 12: 200-218.

- Neuhaus, J. M. & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54: 638-645.

Binary Outcome Multilevel Model: Model Formulation and Parameter Interpretations

What is This About? Intensive Longitudinal Data (ILD) with binary outcomes are often analyzed using logistic multilevel models, in which both variation within individuals (over time) and between individuals (across different sub-populations) are accounted for in the model, but the outcome of interest is the *probability* of a specific binary outcome. Such models have also been referred to as hierarchical generalized linear models, generalized linear mixed models, logistic random-effects models, and mixed-effects logistic models.

Why do we Care? Multilevel modeling is extremely common in the ILD literature. It is important to be able to recognize common language used to reference model terms and conclusions, and to be able to make appropriate types of interpretations with our own ILD studies, particularly in the case of binary outcomes which requires discussion of probabilities, odds, and odds ratios.

What Should we Know? Logistic multilevel modeling is used commonly for traditional longitudinal data analysis, but the following components are common across both types of studies. In general, we are interested in modeling the probability associated with one of two outcomes in terms of baseline covariates (independent of time), ILD covariates (changing over time), time, and a term that accounts for the fluctuation of (or variability in) outcomes among the population of interest. The model does not show the random error term accounting for the variability in individual responses over time, as it is written in terms of the mean instead of the raw response.

$$\begin{array}{lcccccccc} \text{Probability} & & \text{Baseline} & + & \text{ILD} & + & \text{Time} & + & \text{Variation } \textit{Across} \\ \text{of Outcome} & \sim & \text{Variables} & & \text{Variables} & & \text{Trends} & & \text{Individuals} \end{array}$$

More specifically, baseline variables are included as in a typical regression model. On the other hand, ILD variables are often included as a decomposition of "between" effects (differences across individuals) and "within" effects (differences within individuals over time). Time trends can be included as a

simple time term, a polynomial time trend with multiple terms, or possibly as a periodic time term represented by sine and cosine terms (depending on researcher interest and intuition).

Variation across individuals can be represented by a random subject effect, called a *random intercept*, which allows for variation in the intercept term among the population of individuals represented in the study. Variation across individuals can be further represented by interactions of additional random effects with independent variables, called *random slopes*, which allow for variation in predictor effects among the population of individuals represented in the study. Finally, variation within individuals is not shown explicitly in the model, as the probability is used to represent the outcome instead of the raw response. In general, this can be written,

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 z_i + \beta_{2B}\bar{x}_{i.} + \beta_{2W}(\bar{x}_{i.} - x_{it}) + \beta_t t_{it} + u_i,$$

where:

- π_{it} represents the probability of one (of two) outcomes for individual i at time t (conditional on the random effect u_i), (The entire quantity on the left-hand-side of the equation is referred to as the *logit of the success probability*, or the *log of the odds*. This transformation allows a probability, which is restricted to the range of $[0, 1]$, to be equated to any predictors that can take any value.)
- β_0 is the intercept term,
- z_i is a baseline (time-independent) predictor with effect β_1 (also commonly referred to as the coefficient, parameter estimate, or slope),
- x_{it} is an ILD (time-dependent) predictor with between effect β_{2B} and within effect β_{2W},
- t_{it} represents the time in the study with effect β_t, and
- u_i is the random individual effect.

In terms of interpretations, consider an interest in modeling the probability of using a cigarette at any moment in an ILD study.

- β_1 represents the expected difference in the log-odds across populations that differ by a unit change of z_i (a *time-independent* predictor). These are typically interpreted in terms of odds ratios by exponentiating the effect. For example, if z_i represents age, then $exp(\beta_1)$ represents the expected multiplicative change in the odds of cigarette use for populations that differ, on average, by one year of age.
- β_{2B} represents the expected difference in the log-odds across populations that differ in terms of average values of x_{it} (the average values of x_{it} are *time-independent*). Interpretations are again made using odds ratios. For example, if x_{it} represents the momentary mood of individual i at time t, $exp(\beta_{2B})$ represents the expected multiplicative change in the odds of cigarette use across populations that differ by one unit of average mood reported.
- β_{2W} represents the expected change over time of the log-odds as individuals change in terms of momentary values of x_{it} (the momentary values give *time-dependent* predictors). Interpretations are again made using odds ratios. For example, $exp(\beta_{2W})$ represents the expected multiplicative change in the odds of cigarette use for an individual (over time) as mood changes.
- u_i represents the variation in individual log-odds of cigarette use among the population of individuals and is assumed to be distributed as a random normal variable with mean 0 and variance σ_u^2.

Selected References:

- Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. *Analysis of longitudinal data*. New York, N. Y.: Oxford University Press.
- Lalonde, T. L., Wilson, J. R., & Yin, J. (2014). GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Statistics in Medicine*, 33(27): 4756-4769.

- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. *Applied longitudinal analysis*. New York, N.Y.: Wiley.
- Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. *Longitudinal data analysis*. New York, N.Y.: Chapman & Hall / CRC.
- Walls, T. A. & Schafer J. L. (2006). *Models for intensive longitudinal data*. New York, N.Y.: Oxford University Press.

Count Outcome Multilevel Model: Model Formulation and Parameter Interpretations

What is This About? Intensive Longitudinal Data (ILD) with count outcomes are often analyzed using count multilevel models, in which both variation within individuals (over time) and between individuals (across different sub-populations) are accounted for in the model, but the outcome of interest is the *frequency* or *rate* of a specific count outcome. Such models have also been referred to as hierarchical generalized linear models, generalized linear mixed models, count random-effects models, overdispersed count models, and mixed-effects count models.

Why do we Care? Multilevel modeling is extremely common in the ILD literature. It is important to be able to recognize common language used to reference model terms and conclusions, and to be able to make appropriate types of interpretations with our own ILD studies, particularly in the case of count outcomes which requires discussion of mean counts, rates, and overdispersion.

What Should we Know? Count multilevel modeling is used commonly for traditional longitudinal data analysis, but the following components are common across both types of studies. In general, we are interested in modeling the rate or mean count associated with an observed frequency of outcomes in terms of baseline covariates (independent of time), ILD covariates (changing over time), time, and a term that accounts for the fluctuation of (or variability in) outcomes among the population of interest. The model does not show the random error term accounting for the variability in individual responses over time, as it is written in terms of the mean instead of the raw response.

$$\begin{array}{ccccccccc} \text{Mean or Rate} & & \text{Baseline} & & \text{ILD} & & \text{Time} & & \text{Variation } \textit{Across} \\ \text{of Outcome} & \sim & \text{Variables} & + & \text{Variables} & + & \text{Trends} & + & \text{Individuals} \end{array}$$

More specifically, baseline variables are included as in a typical regression model. On the other hand, ILD variables are often included as a decomposition of "between" effects (differences across individuals) and "within" effects

(differences within individuals over time). Time trends can be included as a simple time term, a polynomial time trend with multiple terms, or possibly as a periodic time term represented by sine and cosine terms (depending on researcher interest and intuition).

Variation across individuals can be represented by a random subject effect, called a *random intercept*, which allows for variation in the intercept term among the population of individuals represented in the study. Variation across individuals can be further represented by interactions of additional random effects with independent variables, called *random slopes*, which allow for variation in predictor effects among the population of individuals represented in the study. Finally, variation within individuals is not shown explicitly in the model, as the mean count or rate is used to represent the outcome instead of the raw rsponse. In general, this can be written,

$$\ln(\lambda_{it}) = \beta_0 + \beta_1 z_i + \beta_{2B}\bar{x}_{i.} + \beta_{2W}(\bar{x}_{i.} - x_{it}) + \beta_t t_{it} + u_i,$$

where:

- λ_{it} represents the mean count or rate of outcomes for individual i at time t (conditional on the random effect u_i),
- β_0 is the intercept term,
- z_i is a baseline (time-independent) predictor with effect β_1 (also commonly referred to as the coefficient, parameter estimate, or slope),
- x_{it} is an ILD (time-dependent) predictor with between effect β_{2B} and within effect β_{2W},
- t_{it} represents the time in the study with effect β_t, and
- u_i is the random individual effect.

In terms of interpretations, consider an interest in modeling the mean count or rate of the number of alcoholic drinks consumed at any moment in an ILD study.

- β_1 represents the expected difference in the log-rate across populations that differ by a unit change of z_i (a *time-independent* predictor). These are typically interpreted in terms of multiplicative changes by exponentiating the effect. For example, if z_i represents age, then $exp(\beta_1)$ represents the expected multiplicative change in the mean number of alcoholic drinks for populations that differ, on average, by one year of age.

- β_{2B} represents the expected difference in the log-rate across populations that differ in terms of average values of x_{it} (the average values of x_{it} are *time-independent*). Interpretations are again made using multiplicative chance. For example, if x_{it} represents the momentary mood of individual i at time t, $exp(\beta_{2B})$ represents the expected multiplicative change in the mean number of alcoholic drinks across populations that differ by one unit of average mood reported.

- β_{2W} represents the expected change over time of the log-rate as individuals change in terms of momentary values of x_{it} (the momentary values give *time-dependent* predictors). Interpretations are again made using multiplicative change. For example, $exp(\beta_{2W})$ represents the expected multiplicative change in the mean number of alcoholic drinks for an individual (over time) as mood changes.

- u_i represents the variation in individual log-rate of alcoholic drinks among the population of individuals and is assumed to be distributed as a random normal variable with mean 0 and variance σ_u^2.

Count multilevel models can be adjusted according to different properties of the counts under consideration. For example, standard counts can be modeled using a Poisson MLM, while overdispersed data often require a Negative Binomial MLM, excess-zero counts are better suited for Zero-Inflated MLM or Hurdle MLM, and counts without zeros would require a Zero-Truncated MLM.

Selected References:

- Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. *Analysis of longitudinal data*. New York, N. Y.: Oxford University Press.

- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. *Applied longitudinal analysis*. New York, N.Y.: Wiley.
- Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. *Longitudinal data analysis*. New York, N.Y.: Chapman & Hall / CRC.

Joint Modeling of Mean and Dispersion: Model Formulation and Parameter Interpretations

What is This About? Intensive Longitudinal Data (ILD) are often employed with an interest in making conclusions about fluctuation, volatility, or variation in an outcome. Because of the high intensity of data collection, ILD are uniquely suited to making statements about outcome variances.

Why do we Care? One of the research interests most suited for ILD are those about comparisons of variation. Because of the way the data are collected, it is reasonable to collect data with the goal of comparing variances across populations, or evaluating changes in variation over time.

What Should we Know? Variation in an outcome of interest cannot be analyzed on its own. Instead, researchers must consider the *variation around the mean* of an outcome. Thus, this means that the mean must be modeled as usual, combined with a separate but connected model for the variation. This leads to what are commonly called *joint models of mean and variation.* In such models the mean of the outcome is modeled using the typical multilevel model, while the mean model residuals are used as "data" in a model for the variance in the outcome *around the mean* of the outcome. In general, the two models consider similar predictors as the previous multilevel models.

$$\begin{array}{c} \text{Dependent} \\ \text{Mean} \end{array} \sim \begin{array}{c} \text{Baseline} \\ \text{Variables} \end{array} + \begin{array}{c} \text{ILD} \\ \text{Variables} \end{array} + \begin{array}{c} \text{Time} \\ \text{Trends} \end{array} + \begin{array}{c} \text{Variation } \textit{Across} \\ \text{Individuals} \end{array} + \begin{array}{c} \text{Variation } \textit{Within} \\ \text{Individuals} \end{array}$$

$$\begin{array}{c} \text{Dependent} \\ \text{Variation} \end{array} \sim \begin{array}{c} \text{Baseline} \\ \text{Variables} \end{array} + \begin{array}{c} \text{ILD} \\ \text{Variables} \end{array} + \begin{array}{c} \text{Time} \\ \text{Trends} \end{array} + \begin{array}{c} \text{Variation } \textit{Across} \\ \text{Individuals} \end{array}$$

More specifically, for both mean and variance models, baseline variables are included as in a typical regression model. On the other hand, ILD variables are often included as a decomposition of "between" effects (differences across individuals) and "within" effects (differences within individuals over time). Time trends can be included as a simple time term, a polynomial time trend

with multiple terms, or possibly as a periodic time term represented by sine and cosine terms (depending on researcher interest and intuition).

Variation across individuals can be represented by a random subject effect, called a *random intercept*, which allows for variation in the intercept term among the population of individuals represented in the study. Variation across individuals can be further represented by interactions of additional random effects with independent variables, called *random slopes*, which allow for variation in predictor effects among the population of individuals represented in the study. These random variation terms can be included in the variance model in addition to the mean model. Consider the following normal multilevel model equation,

$$\mu_{it} = \beta_0 + \beta_1 z_i + \beta_{2B}\bar{x}_{i.} + \beta_{2W}(\bar{x}_{i.} - x_{it}) + \beta_t t_{it} + u_i,$$

$$\ln\left(\sigma^2_{it}\right) = \gamma_0 + \gamma_1 z_i + \gamma_{2B}\bar{x}_{i.} + \gamma_{2W}(\bar{x}_{i.} - x_{it}) + \gamma_t t_{it} + v_i,$$

where for the *mean model*:

- μ_{it} represents the mean of the continuous outcome of interest for individual i at time t (conditional on the random effect u_i),
- β_0 is the intercept term,
- z_i is a baseline (time-independent) predictor with effect β_1 (also commonly referred to as the coefficient, parameter estimate, or slope),
- x_{it} is an ILD (time-dependent) predictor with between effect β_{2B} and within effect β_{2W},
- t_{it} represents the time in the study with effect β_t,
- and u_i is the random individual effect.

For the *variance model*:

- σ^2_{it} represents the variation in the outcome for subject i at time t (conditional on the random effect v_i). The logarithmic transformation has been

applied to ensure a positive estimate for σ_{it}^2. (NOTE: The data for the variance model come from the individual deviance residual components d_{it}.)

- γ_0 is the intercept term,
- γ_1 is the effect of the time-independent predictor,
- γ_{2B} and γ_{2W} are the between and within effects, respectively, of the time-dependent predictor,
- γ_t is the time effect, and
- v_i is the random subject effect for the variance model.

Note that neither model shows a random error term such as ϵ_{it} because the left-hand-side is written in terms of the parameters μ_{it} and σ_{it}^2 instead of the raw response and deviance residuals.

In terms of interpretations, consider an interest in modeling the continuous outcome of alcohol craving.

Mean Model:

- β_1 represents the expected difference in the mean outcome across populations that differ, on average, by a unit change in z_i (a *time-independent* predictor). For example, if z_i represents the ages of different individuals, β_1 represents the expected difference in alcohol craving across populations that differ (on average) by one year in age.
- β_{2B} represents the expected difference in the mean outcome across populations that differ in terms of average values of x_{it} (the average values of x_{it} are *time-independent*). For example, if x_{it} represents the momentary mood of individual i at time t, β_{2B} represents the expected difference in alcohol craving across populations that differ by one unit of average mood reported.
- β_{2W} represents the expected change over time of the mean outcome as individuals change in terms of momentary values of x_{it} (the momentary values give *time-dependent* predictors). For example, β_{2W} represents the

expected change for an individual over time in mean alcohol craving as mood changes.

- u_i represents the variation in individual alcohol craving among the population of individuals and is assumed to be distributed as a random normal variable with mean 0 and variance σ_u^2.

Variance Model:

- γ_1 represents the expected difference in the outcome log-variance across populations that differ, on average, by a unit change in z_i (a *time-independent* predictor). For example, if z_i represents the ages of different individuals, γ_1 represents the expected difference in alcohol craving log-variance across populations that differ (on average) by one year in age.

- γ_{2B} represents the expected difference in the outcome log-variance across populations that differ in terms of average values of x_{it} (the average values of x_{it} are *time-independent*). For example, if x_{it} represents the momentary mood of individual i at time t, γ_{2B} represents the expected difference in alcohol craving log-variance across populations that differ by one unit of average mood reported.

- γ_{2W} represents the expected change over time of the outcome log-variance as individuals change in terms of momentary values of x_{it} (the momentary values give *time-dependent* predictors). For example, γ_{2W} represents the expected change for an individual over time in alcohol craving log-variance as mood changes.

- v_i is assumed the be distributed as a random normal variable with mean 0 and variance σ_u^2, which represents the variation in individual alcohol craving among the population of individuals. In many cases v_i is transformed similarly to the variance parameter σ_{it}^2 to ensure its relationship is meaningful on the same scale as the predictors in the model.

Selected References:

- Hedeker, D. Mermelstein, R. J., Berbaum, M. L., & Campbell, R. T. (2009). Modeling mood variation associated with smoking : An ap-

plication of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data. *Addiction*, 104(2): 297-307. doi:*10.1111/j.1360-0443.2008.02435.x*.

- Hedeker, D. Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location-scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, 64, 627-634. DOI: *10.1111/j.l541-0420.2007.00924.x*.

- Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, 31: 3328-3336. DOI: *10.1002/sim.5338*.

- Elliott, M. R., Sammel, M. D., & Faul, J. (2012). Associations between variability of risk factors and health outcomes in longitudinal studies. *Statistics in Medicine*, 31: 2745-2756. DOI: *10.1002/sim.5370*.

- Pugach, O., Hedeker, D., Richmond, M. J., Sokolovsky, A., & Mermelstein, R. (2014). Modeling mood variation and covariation among adolescent smokers: Application of a bivariate location-scale mixed-effects model. *Nicotine & Tobacco Research*, 16: S151-S158.