MASSACHUSETTS GENERAL HOSPITAL

HARVARD MEDICAL SCHOOL

BROAD INSTITUTE

Longitudinal data analysis for observational cohort studies

Didactic Workshop

Kenny Westerman

June 2, 2023

Analysis Therapy

What is a longitudinal dataset?

Dataset containing repeated measurements over time

Observational cohorts

Table 1 Child-based questionnaires.

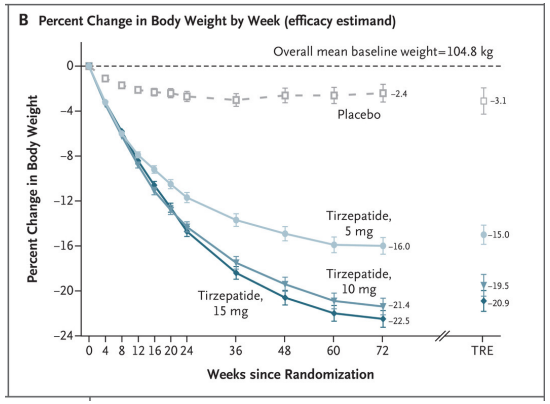
Age (months)	Eligible	Refused and other[a] (%)	Sent	Response (%)
1	14 010	2.5	13 659	90.4
6	13 994	3.6	13 491	85.2
15	13 983	5.5	13 217	83.8
18	13 982	6.7	13 040	85.3
24	13 980	8.3	12 826	81.3
30	13 977	9.3	12 683	81.4
38	13 977	10.5	12 510	79.7
42	13 974	11.0	12 434	80.9

[a] Includes lost to follow-up and mothers who requested no questionnaires for the moment.

ALSPAC cohort study of children

Golding et al. 2004, *Eur. J. Endocrinol.*

Randomized studies



Tirzepatide randomized clinical trial

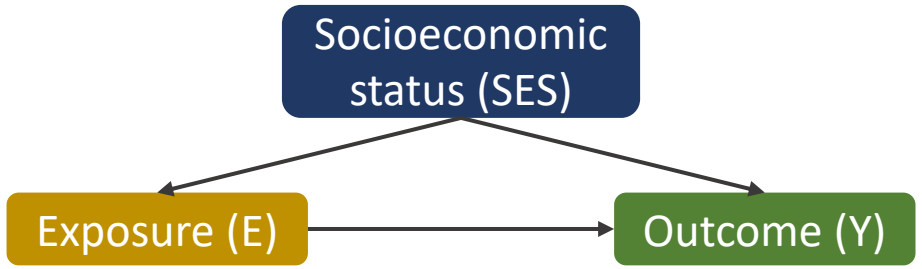
Jastreboff et al. 2022, *N. Engl. J. Med.*

Longitudinal data: advantages

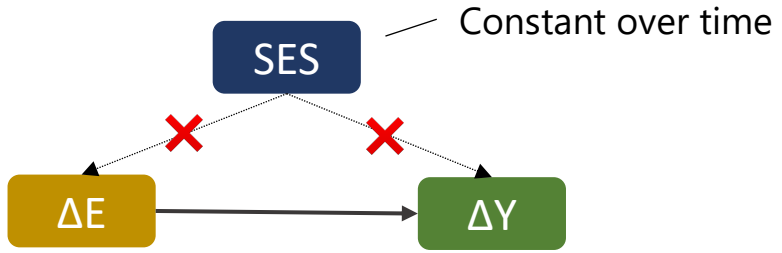
- Greater statistical power from more data points
- Biomarker/risk factor trajectories
  - Can be more complex than just mean differences
  - Time-to-event analysis
- **Within-person comparisons**
  - Decreases the influence of between-person confounding

Within-person comparisons are helpful

- Intuition is similar to using change scores (ΔY ~ ΔE)
- Example: confounding by socioeconomic status



Cross-sectional relationship



Within-person relationship

- Within-person comparisons:
  - Reduce false positives due to true underlying SES-Y relationship
  - Higher statistical power by removing excess variability in Y due to SES

Longitudinal data: challenges

1. Repeated measures are correlated with each other
2. Variance may not be constant across time

Standard regression assumptions are invalid

Problems with statistical inference
(false positives or false negatives)

Agenda

Objective: Understand how longitudinal data can leverage within-person comparisons to improve the estimation of effects in complex trait epidemiology

I. Introduce the linear mixed model (LMM) statistical framework
II. Demonstrate how to fit an LMM and interpret the results using simulated longitudinal data

Linear mixed model framework

Reminder: what do we need to account for beyond a standard regression model?

- Correlated observations
  - Repeated measurements from the same person
- Non-constant variance across timepoints (heteroscedasticity)
  - This is more relevant for randomized settings (not the focus here)

Linear mixed model: concept

- Accounts for expected similarity/correlation between specific observations in a dataset
  - Patients from the same hospital
  - Subjects from the same family (possibly defined by genetics)
  - **Repeated longitudinal measurements from the same subject**

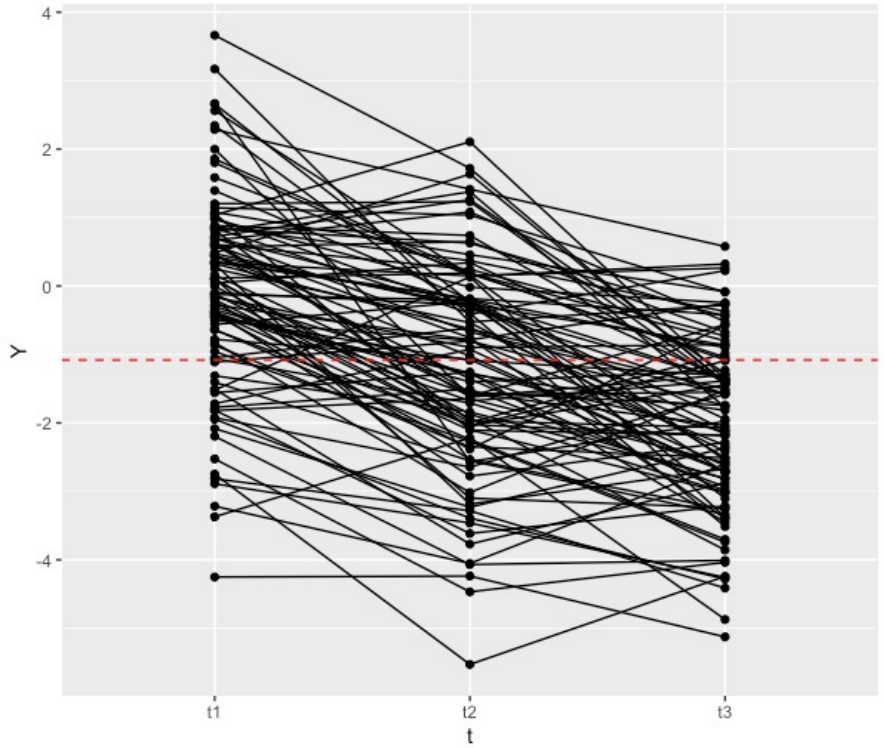
Full linear mixed model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 g_i + \beta_3 e_{ij} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_C + b_{1i} + \epsilon_{ij}$$

- Y_{ij}: Outcome measurement for subject i at timepoint j
- t_{ij}: Time of measurement
- g_i: Genotype
- e_{ij}: Exposure measurement (e.g., some dietary behavior)
- $\boldsymbol{X}_{ij}$: Vector of covariates
- b_{1i}: Random intercept
- ϵ_{ij}: Additional random error

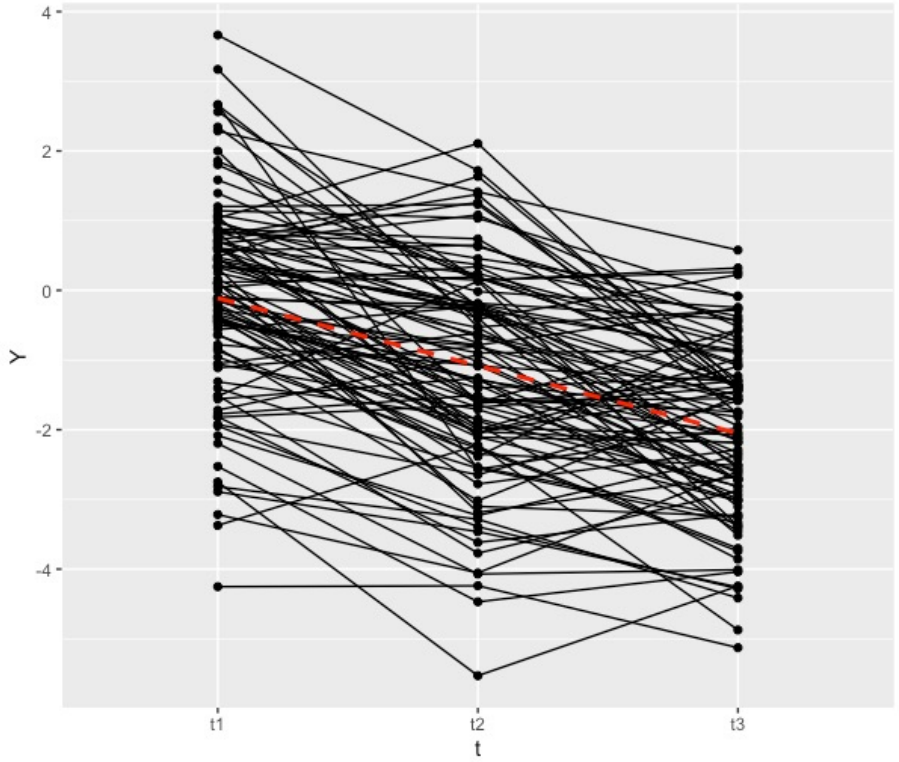
Intercept-only

$$Y_{ij} = \beta_0 + \epsilon_{ij}$$



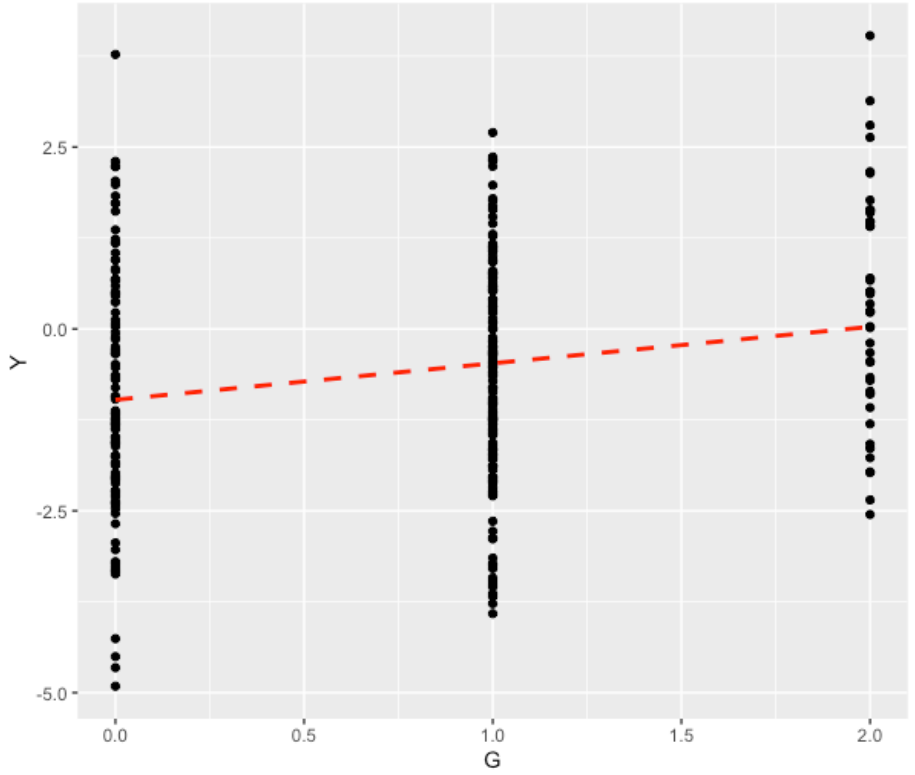
Add a time slope

$$Y_{ij} = \beta_0 + \boldsymbol{\beta_1} t_{ij} + \epsilon_{ij}$$



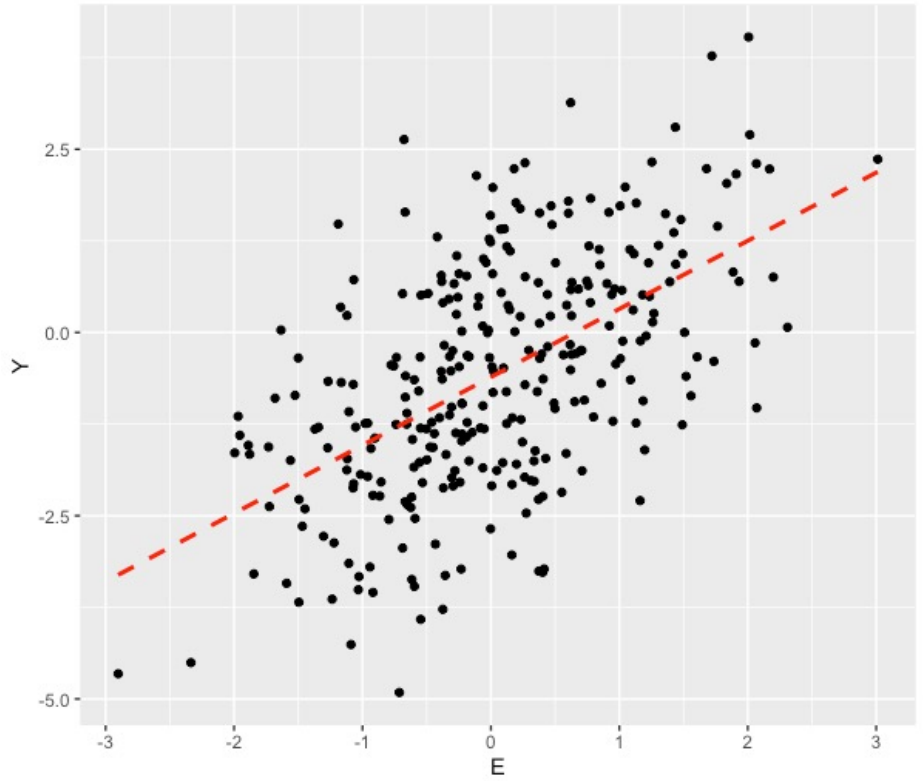
Add a genotype effect

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \boldsymbol{\beta_2} g_i + \epsilon_{ij}$$



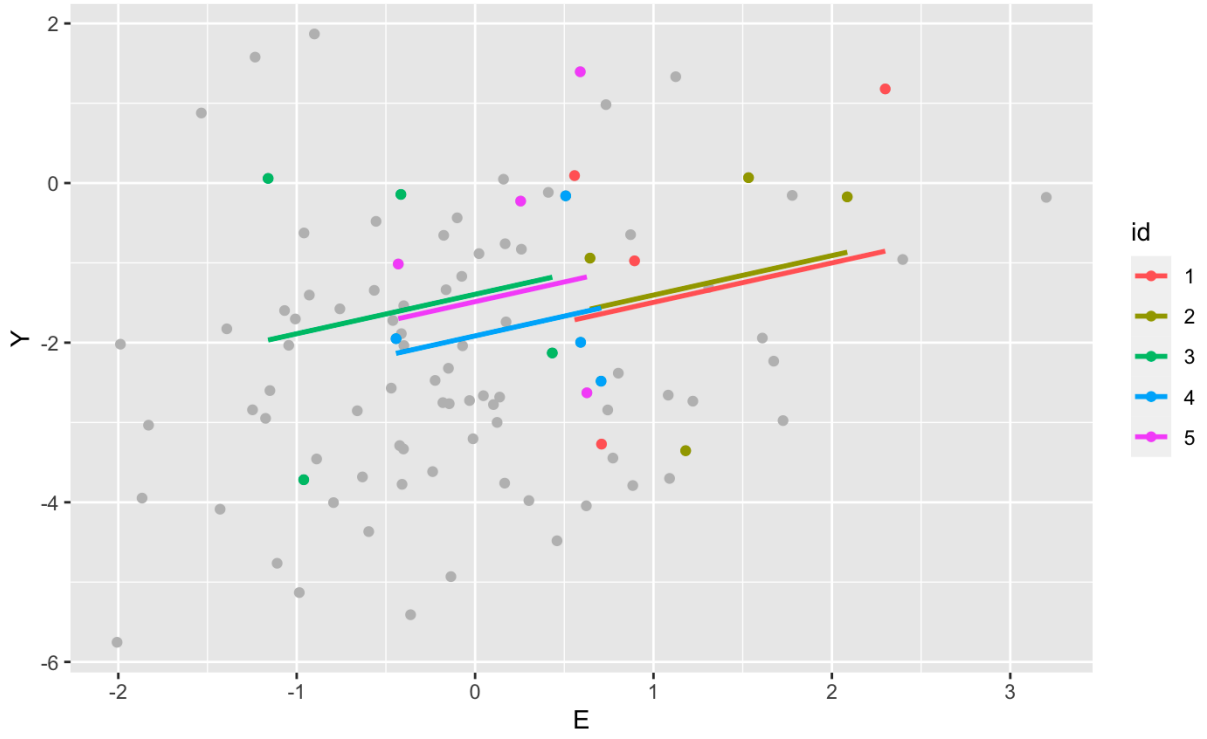
Add an exposure effect

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 g_i + \boldsymbol{\beta_3} e_{ij} + \epsilon_{ij}$$



Add a random intercept

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 g_i + \beta_3 e_{ij} + \boldsymbol{b_{1i}} + \epsilon_{ij}$$



This is the key element of the LMM that accounts for correlated observations from a given individual!

Full linear mixed model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 g_i + \beta_3 e_{ij} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_C + b_{1i} + \epsilon_{ij}$$

- Y_{ij}: Outcome measurement for subject i at timepoint j
- t_{ij}: Time of measurement
- g_i: Genotype
- e_{ij}: Exposure measurement (e.g., some dietary behavior)
- $\boldsymbol{X}_{ij}$: Vector of covariates
- b_{1i}: Random intercept
- ϵ_{ij}: Additional random error

Working through an example using simulated data

Simulated dataset

```
generate_dataset <- function(
  N = 1000,  # Number of individuals
  K = 4,  # Number of timepoints
  maf = 0.25,  # Minor allele frequency
  beta_tY = -1,  # Slope with respect to time
  beta_GY = 0.5,  # Genotype effect size
  icc_E = 0.2,  # ICC for exposure
  var_e_E = 1,  # Error variance for exposure
  beta_EY = 1,  # E-Y effect size
  icc_Y = 0.8,  # ICC for outcome (across repeated measures)
  var_e_Y = 1  # Error variance for outcome
) {
```

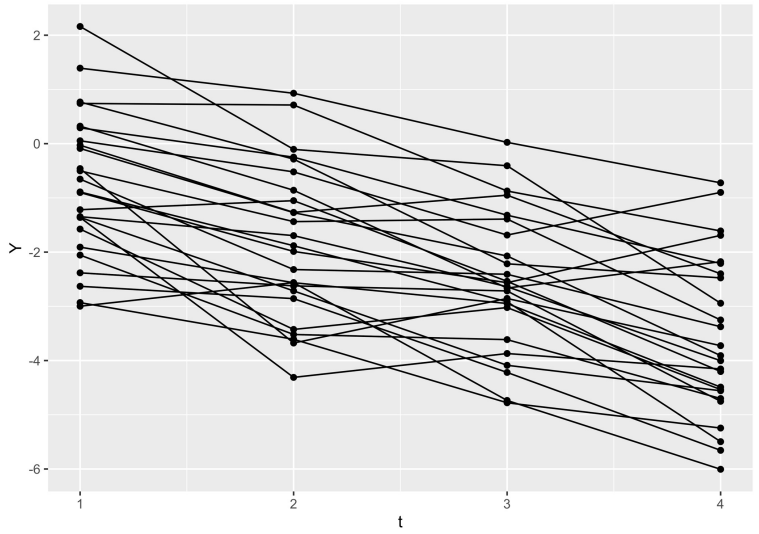
- 1000 individuals
- 4 timepoints
- Random genotype per person (MAF=0.25; constant over time)
- Random continuous exposure (time-varying but with some within-person clustering)
- Outcome simulated as a function of time, genotype, and exposure

longsim: my first R package!

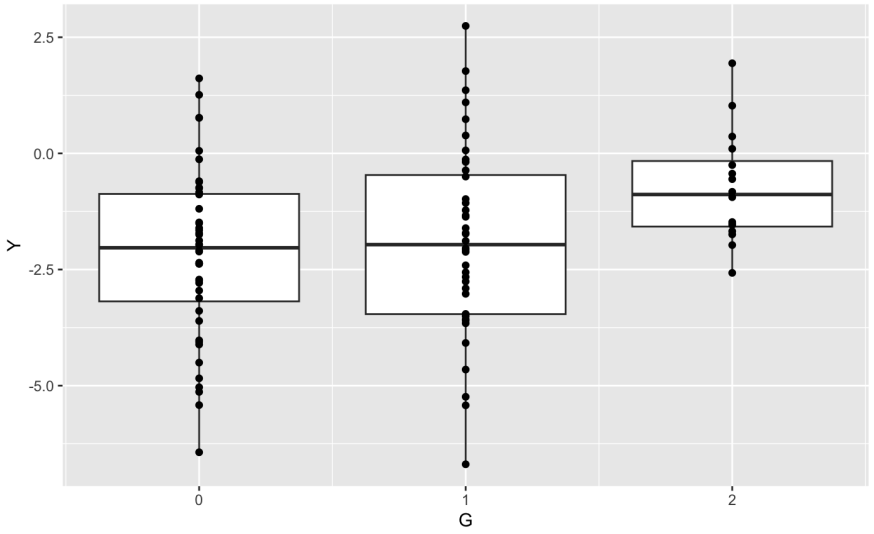
Simulated dataset

```r
generate_dataset <- function(
  N = 1000,  # Number of individuals
  K = 4,  # Number of timepoints
  maf = 0.25,  # Minor allele frequency
  beta_tY = -1,  # Slope with respect to time
  beta_GY = 0.5,  # Genotype effect size
  icc_E = 0.2,  # ICC for exposure
  var_e_E = 1,  # Error variance for exposure
  beta_EY = 1,  # E-Y effect size
  icc_Y = 0.8,  # ICC for outcome (across repeated measures)
  var_e_Y = 1  # Error variance for outcome
) {
```

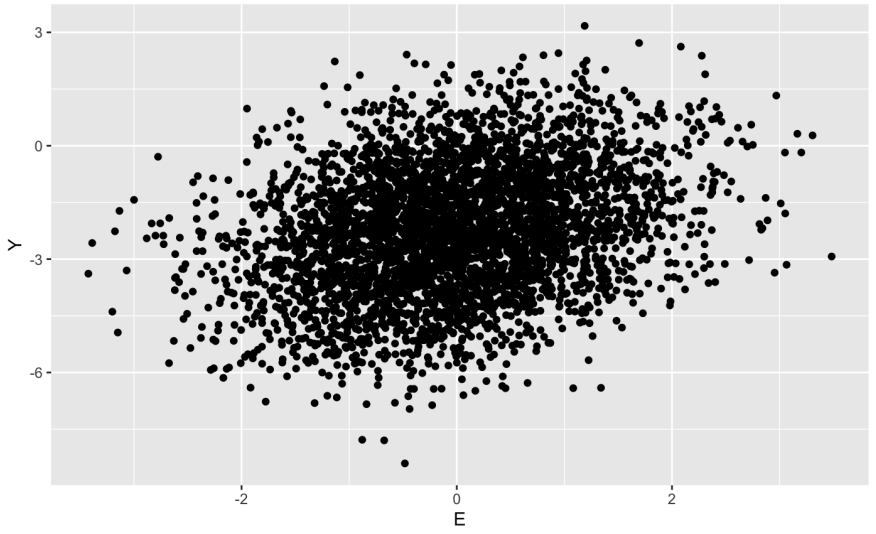
Time effect



Genotype effect



Exposure effect



Simulated dataset

```r
generate_dataset <- function(
  N = 1000,  # Number of individuals
  K = 4,  # Number of timepoints
  maf = 0.25,  # Minor allele frequency
  beta_tY = -1,  # Slope with respect to time
  beta_GY = 0.5,  # Genotype effect size
  icc_E = 0.2,  # ICC for exposure
  var_e_E = 1,  # Error variance for exposure
  beta_EY = 1,  # E-Y effect size
  icc_Y = 0.8,  # ICC for outcome (across repeated measures)
  var_e_Y = 1  # Error variance for outcome
) {
```

Prepare dataset in “long” format

id	timept	Y	t	G	E	C
<fct>	<fct>	<dbl>	<int>	<int>	<dbl>	<dbl>
1	t1	-1.61	1	0	-0.0269	-1.35
1	t2	-1.72	2	0	0.317	-0.127
1	t3	-1.75	3	0	1.57	1.23
1	t4	-4.02	4	0	0.0440	-0.929
2	t1	-0.185	1	1	0.597	-0.408
2	t2	-0.500	2	1	1.19	0.0441

Fitting an LMM using R

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 g_i + \beta_3 e_{ij} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_C + b_{1i} + \epsilon_{ij}$$

```
lmm1_fit <- lmer(Y ~ t + G + E + (1|id), data=long_df)
```

Fixed effects

Random intercept

From the lme4 package for mixed modeling

Output from lme4::lmer() mixed model fit

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Y ~ t + G + E + (1 | id)
   Data: long_df

REML criterion at convergence: 7799.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7113 -0.5944 -0.0096  0.5957  3.3004

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.8369   0.9148
 Residual             0.1993   0.4464
Number of obs: 4000, groups:  id, 1000

Fixed effects:
             Estimate Std. Error         df  t value Pr(>|t|)
(Intercept) -8.726e-03  4.143e-02  1.352e+03   -0.211    0.833
t           -9.905e-01  6.313e-03  2.998e+03 -156.905   <2e-16 ***
G            5.211e-01  4.870e-02  9.979e+02   10.702   <2e-16 ***
E            1.004e+00  1.107e-02  3.499e+03   90.707   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
  (Intr) t      G
t -0.381
G -0.582  0.000
E  0.007  0.001 -0.009
```

Mixed models are all about explaining the outcome variability! After accounting for the fixed effects, how much variance is "assigned" to subject-specific means vs. random error?

Fixed effects estimates look similar to typical linear regression output

How we will visualize LMM results for the next few slides

Beta (SE) taken from fixed effects table

Variance estimates taken from random effects table

Term	LMM
(Intercept)	-0.01 (0.04)
t	-0.99 (0.01)
G	0.52 (0.05)
E	1 (0.01)
Variance - id	0.84
Variance - Residual	0.2

What if we didn't use a linear mixed model?

Use 1st timepoint and use linear regression

Term	LMM	1st timept
(Intercept)	-0.01 (0.04)	-1.01 (0.04)
t	-0.99 (0.01)	NA
G	0.52 (0.05)	0.54 (0.05)
E	1 (0.01)	0.97 (0.03)
Variance - id	0.84	NA
Variance - Residual	0.2	1

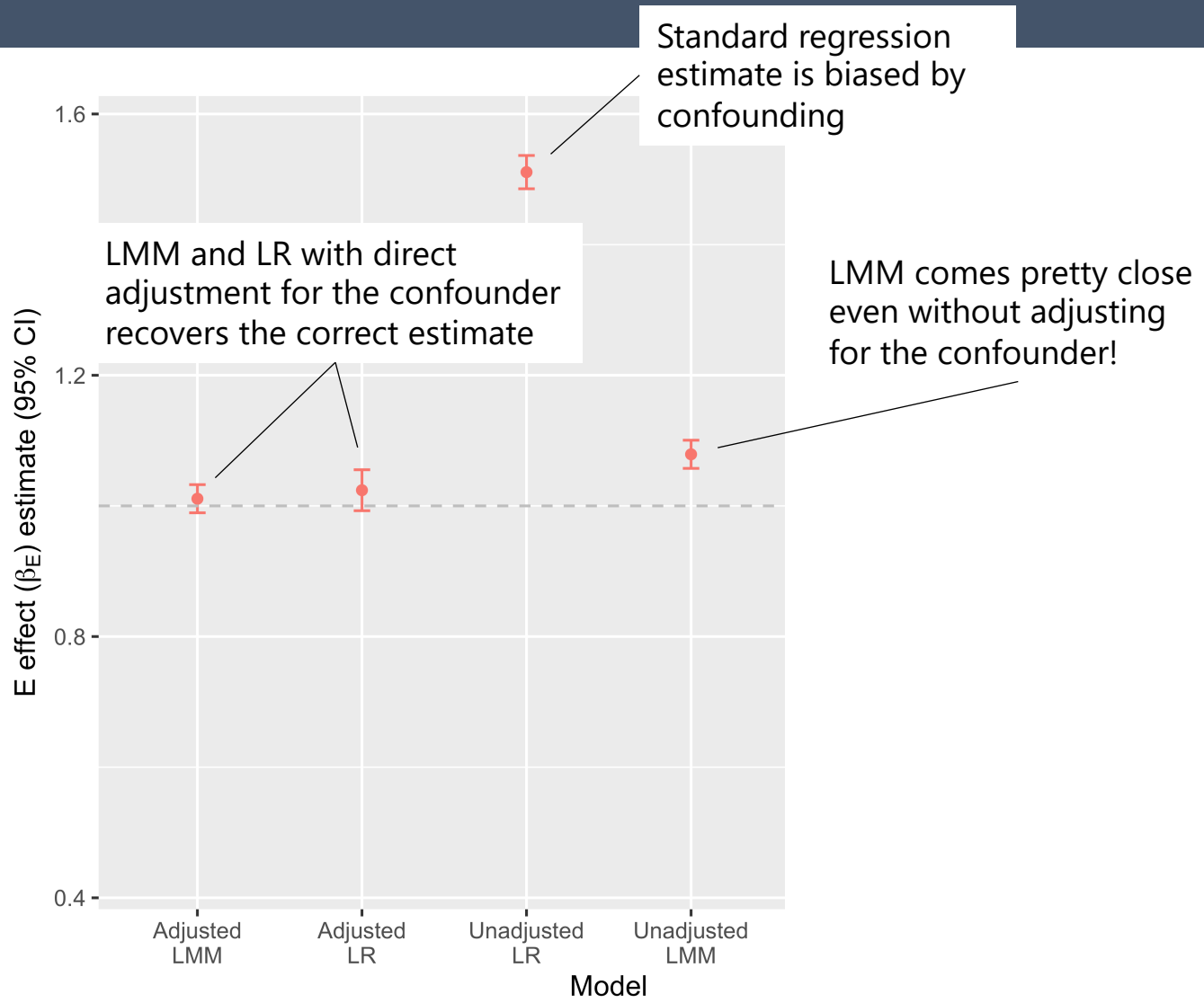
Standard errors are larger due to less information being used

Better than collapsing to 1st/mean, but SE still larger

When does the “within-person comparison” advantage help us?

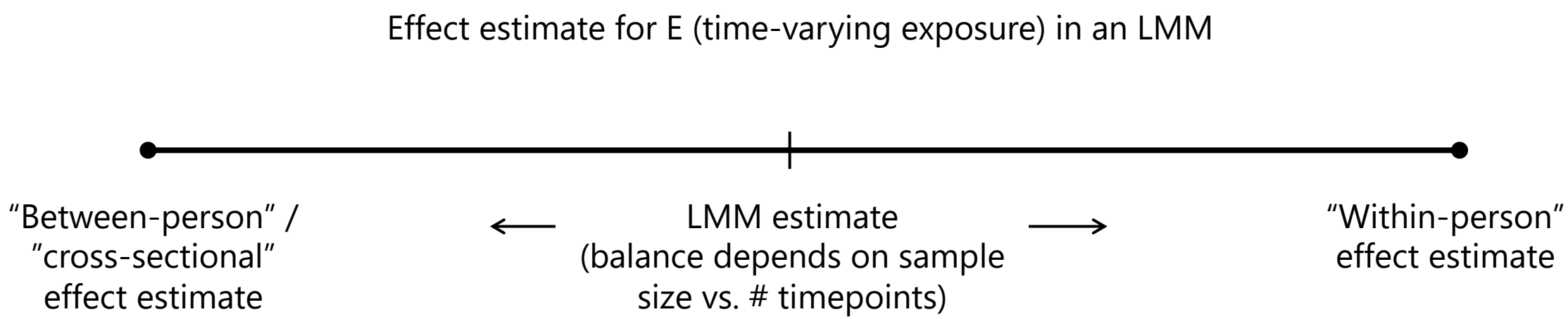
- Investigate empirically:
  - Simulate a time-constant confounder (correlated with each person’s average E and with Y; e.g., socioeconomic status)
  - Important: this type of confounder is not always known or well-measured!

When does the “within-person comparison” advantage help us?



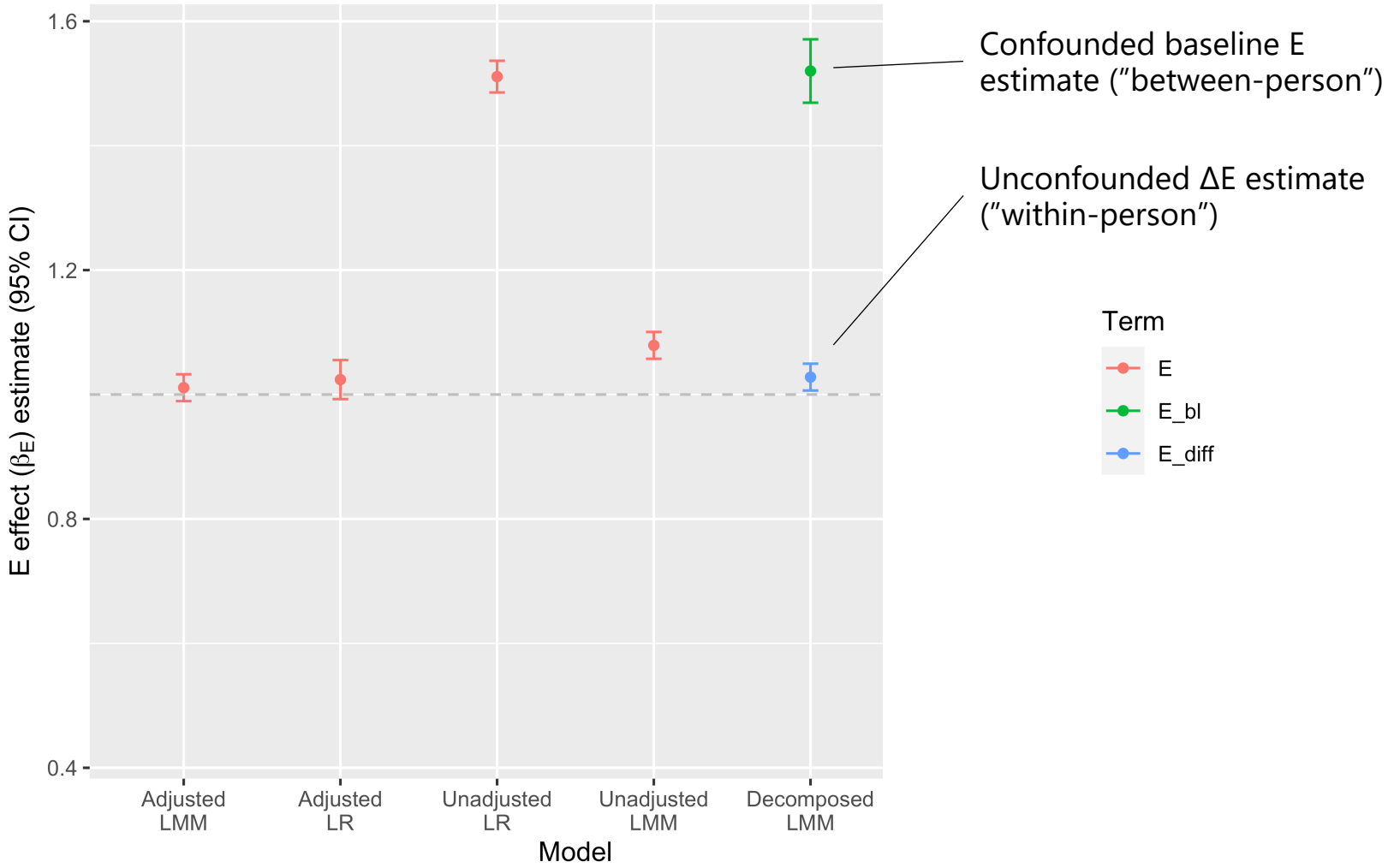
Why is the unadjusted LMM estimate so much better than standard linear regression?

LMM effect estimate is a balance between between- and within-person estimates

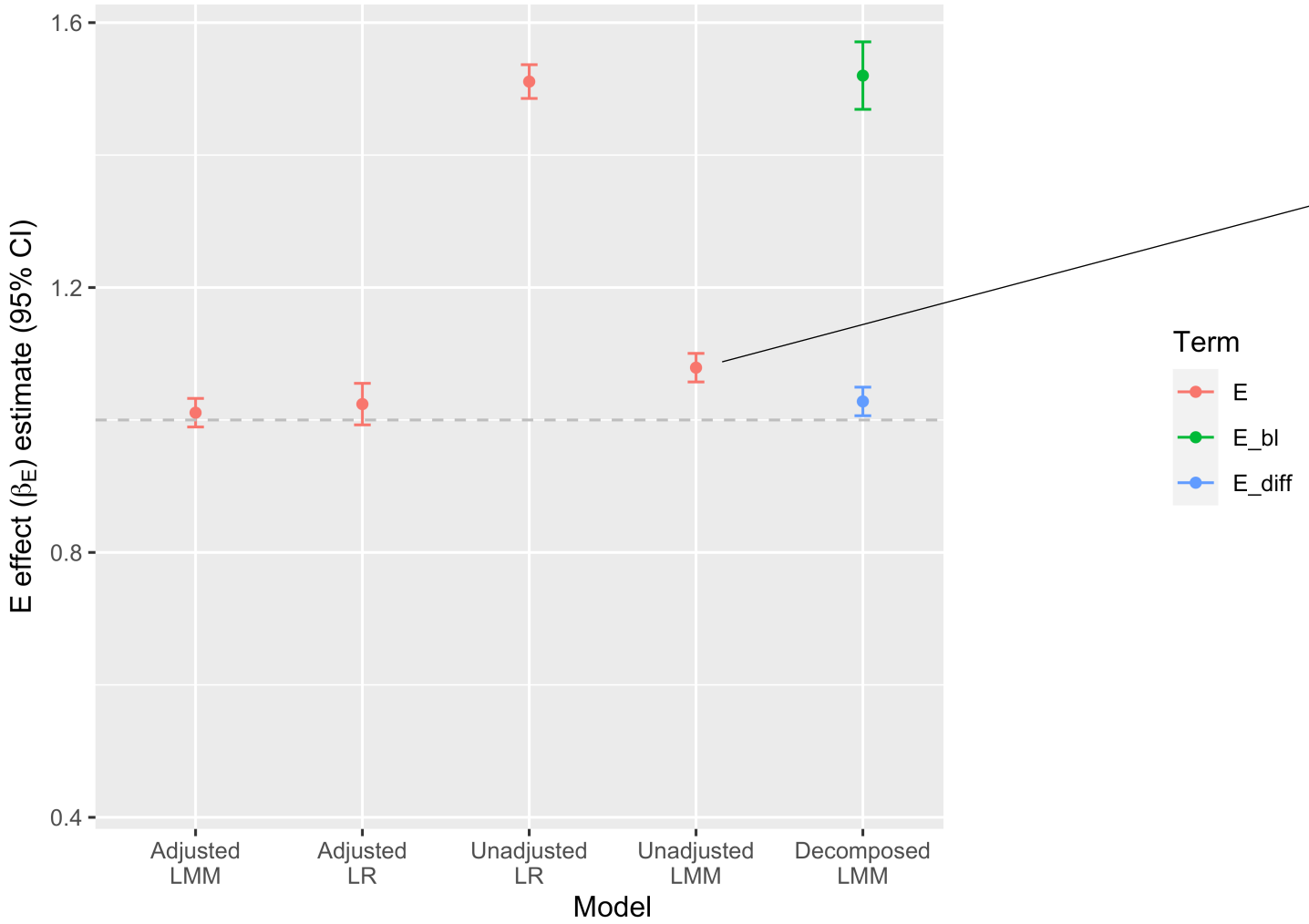


- We can model this explicitly! Imagine "decomposing" the LMM-estimated E effect into 2 parts:
  - An effect of baseline E
  - An effect of subsequent person-specific deviations from that baseline
- Estimate both of these effects in the same LMM and compare them

LMM effect estimate is a balance between between- and within-person estimates



LMM effect estimate is a balance between between- and within-person estimates



Final quiz:
What happens to the LMM effect estimate if we reduce the number of timepoints (say, 2 instead of 4)?

Answer:
It will go up! With fewer timepoints per person, the LMM will give less "weight" to the within-person estimate.

Technical note: time-invariance of confounders

- Confounders don’t need to be strictly constant over time for LMMs to help – they just need some degree of within-person consistency
- Example: healthy lifestyle**
  - Healthy lifestyle driven by upbringing is constant over time (OR ZIP CODE)
  - Healthy lifestyle driven by a recent doctor’s visit is variable

Closing notes

Takeaways

- Conceptual:
  - Longitudinal data can add statistical power and improve effect estimates when correctly modeled (e.g., using LMMs)
  - One key advantage is in their ability to leverage within-person effects that avoid time-constant confounders
  - Effect estimates end up being a balance between (1) between-person effects and (2) within-person effects

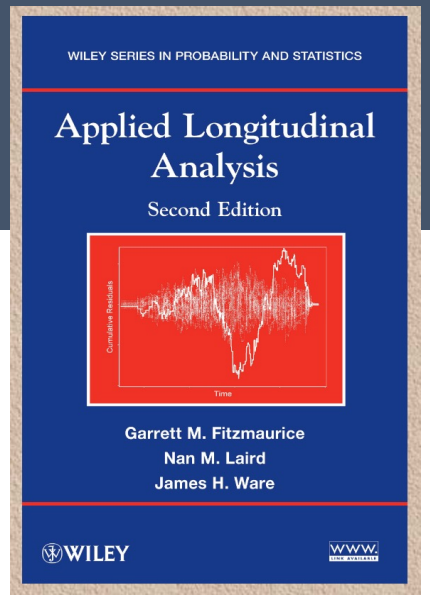
Takeaways

- Practical
  - Data prep:
    - "Long" data format
  - Model specification:
    - Key R packages: *lme4*, *nlme*
    - Include random intercept in model specification: Y ~ E + (1|id)
    - Still adjust for any measured confounders!
  - Model evaluation:
    - Opportunity to use "decomposed" model (baseline E *and* ΔE) to see whether between- and within-person estimates are substantially different

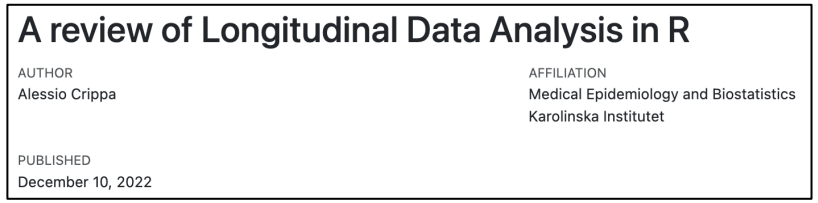
Source materials

- U. Bristol course module: “Introduction to Multilevel Modeling” (2023)
- Garrett Fitzmaurice lecture slides: “Longitudinal Data Analysis” (HSPH BIO 226; slides from 2007)

Further reading



- Textbook: "Applied Longitudinal Analysis" (Fitzmaurice, Laird, & Ware, 2011)
  - https://content.sph.harvard.edu/fitzmaur/ala2e/
  - Accompanying lecture slides (HSPH BIO 226)
- R tutorial using *lme4* R package
  - https://rpubs.com/alecri/review_longitudinal



- Discussion of time-varying confounders and within-person effect estimates
  - Rohrer & Murayama 2023, *Adv. Meth. Pract. Psych. Sci.*: https://doi.org/10.1177/25152459221140842

Extra notes if time

How does this fit in the context of a genetic epidemiology workflow?

- Major obstacle: computational complexity of mixed model fits
- Efficient implementations
  - Ex. glmmkin::glmm() for null model fitting, which allows for longitudinal data
  - TrajGWAS software tool
- Follow-up or characterization of known loci

Drawbacks of non-LMM approaches

- Naïve OLS
  - Affects standard errors: can be either too aggressive or too conservative
- OLS with robust standard errors (also generalized estimating equations)
  - Doesn't produce estimates of individual effects (only populations)
  - Can't handle multiple sources of clustering
- Include fixed-effect intercepts for every person
  - Lose statistical power (degrees of freedom)
  - Can't estimate effects for constant person-level variables (e.g., genotype)
- Collapse data (e.g., take means) then use OLS
  - Lose information
  - Ecological fallacy

What about binary or categorical outcomes?

- Generalized linear mixed models extend standard LMMs
- Conceptually, similar extension that e.g. logistic regression is to ordinary linear regression
- Nuances in interpretation of random effects due to the nonlinear transformation between the linear prediction based on covariates and the quantity of interest (say, an outcome probability)

ML vs. REML

- Restricted maximum likelihood (REML) is commonly used to fit mixed models, especially when the number of clusters (here, individuals) is small relative to the number of observations
- Conceptually, ML underestimates the residual* variance, whereas REML is an unbiased alternative
- See other resources for in-depth discussion

Missing data

- More common in longitudinal data
- Depends on the “mechanism” assumed to generate the missingness (MCAR, MAR, NMAR) – these definitions are outside our scope here
- High-level: the more “non-random” the missingness pattern is assumed to be, the more complex the models need to be and the more care needs to be taken in modeling and interpretation

High-frequency data and monitoring

- Are useful patterns discoverable in high-frequency data?
- Data sources: accelerometer data, other wearables, CGM
- What is special about high-frequency data?
  - One timepoint → can only calculate/compare means
  - Two timepoints → mean or slope
  - Many timepoints → mean, slope, max/min, anomaly detection
- Example paper: Alavi et al. 2021 (Michael Snyder lab)