Simple methods for handling non-randomly missing data

Sophia Rabe-Hesketh

Education & Biostatistics
University of California, Berkeley
sophiarh@berkeley.edu

Anders Skrondal

Norwegian Institute of Public Health, Oslo

MMM Conference 2014, Connecticut

Outline

- Cross-sectional data
  - I. Regression models
- Longitudinal data
  - II. Multilevel linear models
  - III. Multilevel logistic models

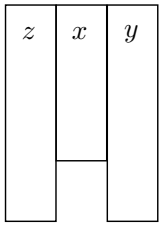
Missing (completely) at random

- Missing at random (MAR): probability of observing variables in the model cannot depend on unobserved values of these variables (given the observed values) [Rubin, 1976]
- Notation:
  - $\mathbf{u}_i$ is vector of variables in model for individual i, e.g. $\mathbf{u}_i = (z_i, x_i, y_i)'$
  - $\mathbf{s}_i$ is vector of **selection indicators**, equal to 1 if corresponding variable observed, e.g. $\mathbf{s}_i = (s_i^z, s_i^x, s_i^y)$
  - $\mathbf{u}_i^{\text{obs}}$ is subset of $\mathbf{u}_i$ that is observed
- MAR means $p(\mathbf{s}_i|\mathbf{u}_i) = p(\mathbf{s}_i|\mathbf{u}_i^{\text{obs}})$
- Missing completely at random (MCAR) means $p(\mathbf{s}_i|\mathbf{u}_i) = p(\mathbf{s}_i)$
- Note:
  - Requiring MAR for consistency assumes that $\mathbf{u}_i$ are **response** variables
  - Definitions often applied to each variable individually, e.g. MAR for x_i means $p(s_i^x|z_i, x_i, y_i) = p(s_i^x|z_i, y_i)$
  - Selection of one variable can depend on selection of other variables

- Cross-sectional data
  - I. Regression models

Motivating example

Missingness pattern:



- z_i always observed
- x_i **often missing**
- y_i always observed

- Model: $y_i = \beta_0 + \beta_z z_i + \beta_x x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$
- For example, y_i is life satisfaction, z_i is age, x_i is income
- Best way to handle missing values of x_i?
  - Complete-case analysis (CC)?
  - Multiple imputation (MI)?
  - Structural equation modeling (SEM)?
    – joint modeling of z_i, x_i, y_i

Complete-case analysis (CC)

- Analyze subsample of individuals with complete data, also called listwise analysis
- Here, analyze individuals for whom x_i is observed or **selected**, indicated by $s_i^x = 1$, with $s_i^x = 0$, otherwise
- What missingness assumption required for consistent estimation?
- MCAR often believed to be necessary. **Not true!**

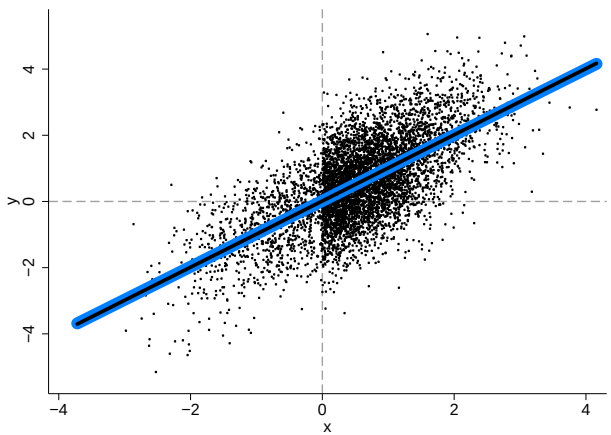
- Consistent estimation requires

$$p(s_i^x | z_i, x_i, y_i) = p(s_i^x | z_i, x_i)$$

  - Whether income is reported can depend on income (and age)!

[Little, 1992]

Illustration of requirement for CC to be consistent



- $p(s_i^x | x_i, y_i) = p(s_i^x | x_i)$
- Selection unlikely if x less than 0

Consistency of CC

- Investigate consistency by simulating huge sample ($N = 1,000,000$)
- Covariates: z_i, x_i bivariate normal, zero means, unit standard deviation, correlation 0.5
- Response: $y_i = \underbrace{1}_{\beta_0} + \underbrace{-1}_{\beta_z} z_i + \underbrace{1}_{\beta_x} x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \underbrace{\sigma^2}_{1})$
- Selection of x_i:

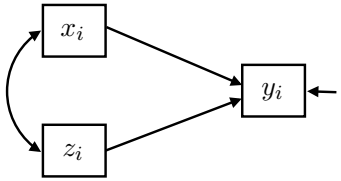
$$p(s_i^x = 1 | z_i, x_i) = \begin{cases} 0.1 & \text{if } z_i > 0 \text{ and } x_i > 1 \\ 0.2 & \text{if } z_i < 0 \text{ and } x_i < 0 \\ 1.0 & \text{otherwise} \end{cases}$$

- Results:

		Inconsistency (%)				SE×10^4		
s^x	Method	β_0	β_z	β_x	σ	β_0	β_z	β_x
z,x	CC	0	0	0	0	13	15	17

Structural equation modeling (SEM)

- Model all variables jointly, with unstructured means, variances, covariances among covariates



- Like multiple imputation (MI) of missing values, unless auxiliary variables used

- Consistent estimation requires MAR

$$p(s_i^x|z_i, x_i, y_i) = p(s_i^x|z_i, y_i)$$

  - Whether income is reported can depend on life satisfaction (and age) **but cannot depend on income**

[Muthén, Kaplan & Hollis, 1987]

Consistency of SEM and CC

- Previous mechanism for selection of x_i

		Inconsistency (%)				SE$\times 10^4$		
s^x	Method	β_0	β_z	β_x	σ	β_0	β_z	β_x
z,x	CC	0	0	0	0	13	15	17
	SEM	-5	-24	6	3	11	13	15

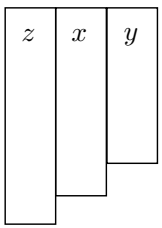
- New mechanism for selection of x_i

$$p(s_i^x = 1|z_i, y_i) = \begin{cases} 0.1 & \text{if } z_i > 0 \text{ and } y_i > 1.8 \\ 0.2 & \text{if } z_i < 0 \text{ and } y_i < 2.3 \\ 1.0 & \text{otherwise} \end{cases}$$

		Inconsistency (%)				SE$\times 10^4$		
s^x	Method	β_0	β_z	β_x	σ	β_0	β_z	β_x
z,y	CC	7	-20	-8	-4	13	15	13
	SEM	0	0	0	0	11	13	13

How useful is SEM/MI when some missingness patterns never occur?

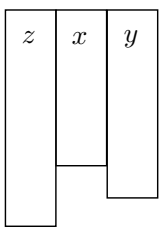
- y never observed when x missing, $p(s^y=1|z, x, y, s^x=0) = 0$



No info on $p(y|z, x)$ when y is missing
SEM equivalent to CC for estimating $p(y|z, x)$

	SEM	CC
s^y	z,x	z,x
s^x	z,x	z,x

- x never observed when y missing, $p(s^x=1|z, x, y, s^y=0) = 0$

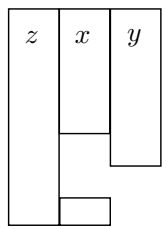


No info on $p(y|z, x)$ when y is missing
SEM **not** equivalent to CC

	SEM	CC
s^y	z	z,x
s^x	z,y	z,x

How useful is SEM/MI when all missingness patterns occur?

- No missingness patterns ruled out (except z always observed)



Some info on $p(y|z, x)$ when y missing, x observed (helps predict missing x where y observed [Little, 1992])

	SEM	CC	$s^y = 1$
s^y	z	z,x	z
s^x	z	z,x	z,y

Conclusions for cross-sectional data: regression models

- Use CC if selection never depends on y, given covariates
  - Often reasonable if outcome occurs **after** covariate information collected
- (In logistic regression, CC can be consistent if selection depends on y [Vach & Illy, 1997])
- Otherwise, consider discarding some individuals – extension of "subsample ignorable likelihood" [Little & Zhang, 2011]
  - Variables that predicts missingness of themselves or other variables cannot be missing

- Longitudinal data

II. Multilevel linear models

Longitudinal data and multilevel linear model

- Occasions t (level 1) and individuals i (level 2)

$$y_{it} = \alpha + \beta x_{it} + \gamma z_i + \zeta_i + \epsilon_{it}$$

  - y_{it} is response variable, vector is $\mathbf{y}_i$
  - α is fixed intercept
  - x_{it} is time-varying (level-1) covariate (vector $\mathbf{x}_i$) with coefficient β
  - z_i is time-constant (level-2) covariate with coefficient γ
  - ζ_i is individual-specific (level-2) intercept, a random effect
  - ϵ_{it} is occasion-specific (level-1) error term
- Usual assumptions
  - ζ_i independent across individuals
  - ϵ_{it} independent across occasions and individuals
  - $\zeta_i|x_{it}, z_i \sim N(0, \psi)$ and $\epsilon_{ij}|x_{it}, z_i, \zeta_i \sim N(0, \theta)$
- Maximize Marginal Likelihood (MML)

$$\prod_{i=1}^{N} p(\mathbf{y}_i|z_i, \mathbf{x}_i)$$

Types of MAR for longitudinal data

- Now $\mathbf{s}_i = (s_{i1}, \ldots, s_{iT})'$, with s_{it} indicating that z_i, x_{it}, y_{it} observed
- Missing at random (MAR):

$$p(\mathbf{s}_i|\mathbf{y}_i, z_i, \mathbf{x}_i, \zeta_i) = p(\mathbf{s}_i|\mathbf{y}_i^{\text{obs}}, z_i, \mathbf{x}_i)$$

  - Covariate-dependent selection [Little, 1995]:

$$p(\mathbf{s}_i|\mathbf{y}_i, z_i, \mathbf{x}_i, \zeta_i) = p(\mathbf{s}_i|z_i, \mathbf{x}_i)$$

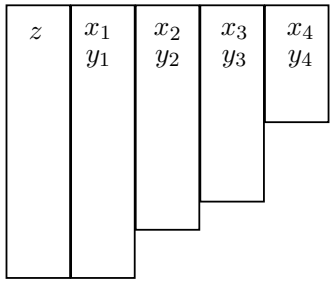
    - ◇ No need to model longitudinal dependence of $\mathbf{y}_i$ given covariates
    - ◇ $\Longrightarrow$ Can use pooled OLS or GEE with robust standard errors
  - Selection dependent on (previous) observed responses:

$$p(\mathbf{s}_i|\mathbf{y}_i, z_i, \mathbf{x}_i, \zeta_i) = p(\mathbf{s}_i|\mathbf{y}_i^{\text{obs}}, z_i, \mathbf{x}_i)$$

    - ◇ Need to model longitudinal dependence correctly
    - ◇ Unless y_{it} affects selection only if observed (implausible), need **monotone missingness**

Advantage of monotone missingness

- Monotone missingness or dropout/attrition: **cannot return**



- If selection of y_{it} and/or x_{it} depends on $y_{i,t-1}$ for $t > 1$
  - Monotone missingness, dependence modeled correctly: consistent estimation
  - Non-monotone missingness: inconsistent estimation
  - Non-monotone missingness, **make monotone**: consistent estimation! [Rabe-Hesketh & Skrondal, 2014]

Types of NMAR for longitudinal data

- Outcome-based selection [Little, 1995]

$$p(\mathbf{s}_i|\mathbf{y}_i, z_i, \mathbf{x}_i, \zeta_i) \;=\; p(\mathbf{s}_i|\mathbf{y}_i^{\text{obs}}, \mathbf{y}_i^{\text{mis}}, z_i, \mathbf{x}_i)$$

  - Only "remedy": model selection jointly with longitudinal data [Diggle & Kenward, 1994], but makes **unverifiable assumptions**
- Random-coefficient-based selection [Little, 1995]

$$p(\mathbf{s}_i|\mathbf{y}_i, z_i, \mathbf{x}_i, \zeta_i) \;=\; p(\mathbf{s}_i|z_i, \mathbf{x}_i, \zeta_i)$$

  - One "remedy": model selection jointly with longitudinal data [Wu & Carroll, 1988], but makes **unverifiable assumptions**
  - Alternative remedy: treat ζ_i as **fixed** [Verbeek & Nijman, 1992]

Fixed-effects approach

- Treat ζ_i as fixed parameter, e.g. by using dummy variables I_{ri} for individuals $r = 1, \ldots, N$ and omitting intercept α

$$y_{it} \;=\; \beta x_{it} + \cancel{\gamma z_i} + \sum_{r=1}^{N} \zeta_r I_{ri} + \epsilon_{it}$$

- Selection based on ζ_i becomes selection based on covariates I_{ri} $\Longrightarrow$ MAR
- γ cannot be estimated because z_i perfectly explained by dummies
- Control for all possible known and unknown individual-level confounders!

Example of fixed-effects: "Difference-in-Differences"

- (Quasi-)Experiment with intervention $z_i = 1$ and control $z_i = 0$
- Pre-test response y_{i1} and post-test response y_{i2}
- Dummy variable T_{2t} for post-test occasion $T_{22} = 1$, $T_{21} = 0$

$$y_{it} \;=\; \alpha T_{2t} + \delta T_{2t} z_i + \sum_{r=1}^{N} \zeta_r I_{ri} + \epsilon_{it}$$

- Change-scores, or "differences" (eliminate ζ_i)

$$y_{i2} - y_{i1} \;=\; \alpha + \delta z_i + \epsilon_{i2} - \epsilon_{i1}$$

- δ is difference in mean change between intervention and control

Conclusions for multilevel linear model

- MAR useful if covariate-dependent selection
- MAR with observed-response-dependent selection requires
  - Correct model for longitudinal dependence
  - Monotone missing data – throw away responses!
  - Selection to depend on previous responses
- NMAR with random-coefficient-dependent selection
  - Handle using fixed-effects approach, but lose γ
  - (Also possible with random slopes)
- No easy solution for NMAR with outcome-dependent selection

- Longitudinal data

(III.) Multilevel logistic models

[Skrondal & Rabe-Hesketh, 2014]

Multilevel logistic model

- Previous multilevel linear model

$$y_{it} = \alpha + \beta x_{it} + \gamma z_i + \zeta_i + \epsilon_{it}$$

- Multilevel logistic model

$$\text{logit}[p(y_{it}=1|\mathbf{w}_i,\zeta_i)] = \alpha + \beta x_{it} + \gamma z_i + \zeta_i \equiv \nu_{ij}$$

  - Use notation $\mathbf{w}_i$ for all covariates ($\mathbf{x}_i$ and z_i)
- Conditional probability of $\mathbf{y}_i$

$$p(\mathbf{y}_i|\mathbf{w}_i,\zeta_i) = \prod_{t=1}^{T} \frac{\exp(\nu_{ij})^{y_{it}}}{1+\exp(\nu_{ij})}$$

- Treating ζ_i as random leads to same problems as in linear case
- Treating ζ_i as fixed allows for more types of selection
- Cannot use dummy-variable method, but use **conditional maximum likelihood** (CML)

Conditional maximum likelihood (CML)

- ζ_i treated as *fixed parameter* in estimation
- Cond. likelihood contribution of individual i, given $\sum_{t=1}^{T} y_{it} = \tau_i$:

$$\begin{aligned} l_i^{\text{CML}} &\equiv p(\mathbf{y}_i|\textstyle\sum y_{it}=\tau_i, \mathbf{w}_i, \zeta_i) \\ &= p(\mathbf{y}_i|\mathbf{w}_i,\zeta_i) \Big/ p(\textstyle\sum y_{it}=\tau_i|\mathbf{w}_i,\zeta_i) \\ &= \frac{\prod_{t=1}^{T}\exp(\beta x_{it})^{y_{it}}}{\sum_{\mathbf{d}_i\in\mathcal{B}_i}\prod_{t=1}^{T}\exp(\beta x_{it})^{d_{it}}} \end{aligned}$$

where $\mathcal{B}_i = \{\mathbf{d}_i = (d_{i1},\ldots,d_{iT})' \mid d_{it}=0 \text{ or } 1, \text{ and } \sum_t d_{it} = \tau_i\}$
(set of all permutations of 0's and 1's whose sum is τ_i)

- Between-subject component $\alpha + \gamma z_i + \zeta_i$ "conditioned away"

Performance of CML for missing data

- Set of occasions with observed outcomes for individual i is $\mathcal{I}_i$
- $\mathcal{B}_i = \{\mathbf{d}_i |\ d_{it} = 0 \text{ or } 1,\ t \in \mathcal{I}_i, \text{ and } \sum_{t\in\mathcal{I}_i} d_{it} = \tau_i\}$
- Conditional likelihood contribution, given $\mathbf{s}_i$:

$$p(\mathbf{y}_i^{\text{obs}} | \mathbf{s}_i, \sum_{t\in\mathcal{I}_i} y_{it} = \tau_i, \mathbf{w}_i, \zeta_i) \ =$$

$$\frac{\left[\prod_{t\in\mathcal{I}_i} \exp(\beta x_{it})^{y_{it}}\right] \int_{\mathbf{y}_i^{\text{mis}}} p(\mathbf{s}_i|\mathbf{y}_i^{\text{obs}}, \mathbf{y}_i^{\text{mis}}, \mathbf{w}_i, \zeta_i) p(\mathbf{y}_i^{\text{mis}}|\mathbf{w}_i, \zeta_i)\, \mathrm{d}\mathbf{y}_i^{\text{mis}}}{\sum_{\mathbf{d}_i\in\mathcal{B}_i}\left[\prod_{t\in\mathcal{I}_i} \exp(\beta x_{it})^{d_{it}}\right] \int_{\mathbf{y}_i^{\text{mis}}} p(\mathbf{s}_i|\mathbf{d}_i, \mathbf{y}_i^{\text{mis}}, \mathbf{w}_i, \zeta_i) p(\mathbf{y}_i^{\text{mis}}|\mathbf{w}_i, \zeta_i)\, \mathrm{d}\mathbf{y}_i^{\text{mis}}}$$

- Consistent estimation generally requires correct joint modeling of $\mathbf{y}_i^{\text{obs}}$ and $\mathbf{s}_i$

CML for covariate, random-effect, and missing-outcome dependent selection

- Let $p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{w}_i, \zeta_i) = p(\mathbf{s}_i|\mathbf{y}_i^{\text{mis}}, \mathbf{w}_i, \zeta_i)$
- Conditional likelihood contribution, given $\mathbf{s}_i$:

$$p(\mathbf{y}_i^{\text{obs}}|\mathbf{s}_i, \sum_{t\in\mathcal{I}_i} y_{it} = \tau_i, \mathbf{w}_i, \zeta_i) \ =$$

$$\frac{\prod_{t\in\mathcal{I}_i} \exp(\beta x_{it})^{y_{it}}}{\sum_{\mathbf{d}_i\in\mathcal{B}_i} \prod_{t\in\mathcal{I}_i} \exp(\beta x_{it})^{d_{it}}}$$

- Consistent estimation of β from standard CML
- Selection can depend on $\mathbf{y}_i^{\text{mis}}$ and ζ_i (unlike MML)

CML for current-outcome-dependent selection

- Current-outcome-dependent selection:

$$\begin{aligned} p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{w}_i, \zeta_i) &= \prod_{t=1}^{T} p(s_{it}|y_{it}) \\ &= \left[\prod_{t\in\mathcal{I}_i} p(s_{it}=1|y_{it}^{\text{obs}})\right] \prod_{t\in\overline{\mathcal{I}}_i} p(s_{it}=0|y_{it}^{\text{mis}}) \\ &\equiv \left[\prod_{t\in\mathcal{I}_i} \pi_t(y_{it}^{\text{obs}})\right] \prod_{t\in\overline{\mathcal{I}}_i} \overline{\pi}_t(y_{it}^{\text{mis}}) \end{aligned}$$

CML for current-outcome-dependent selection (cont'd)

- $p(\mathbf{s}_i|\mathbf{y}_i, \mathbf{w}_i, \zeta_i) = \left[\prod_{t\in\mathcal{I}_i} \pi_t(y_{it}^{\text{obs}})\right] \prod_{t\in\overline{\mathcal{I}}_i} \overline{\pi}_t(y_{it}^{\text{mis}})$
- Conditional likelihood contribution, given $\mathbf{s}_i$:

$$p(\mathbf{y}_i^{\text{obs}}|\mathbf{s}_i, \sum_{t\in\mathcal{I}_i} y_{it} = \tau_i, \mathbf{w}_i, \zeta_i) \ =$$

$$\frac{\prod_{t\in\mathcal{I}_i} \exp([\ln(\pi_t(1)/\pi_t(0))] + \beta x_{it})^{y_{it}}}{\sum_{\mathbf{d}_i\in\mathcal{B}_i} \prod_{t\in\mathcal{I}_i} \exp([\ln(\pi_t(1)/\pi_t(0))] + \beta x_{it})^{d_{it}}}$$

- **Absorb** $\ln(\pi_t(1)/\pi_t(0))$ in α_t to obtain consistent $\widehat{\beta}$

$$\nu_{it} = \underbrace{\alpha_t}_{\ln(\pi_t(1)/\pi_t(0))} + \beta x_{it}, \quad \alpha_1 = 0$$

  - Difference-in-Differences: $\nu_{it} = \alpha_t + \delta T_{2t} z_i + \zeta_i$

CML for lagged-outcome dependent selection

- Inconsistent estimation of β from standard CML with α_t
- Selection depends on lag(1) outcomes:
  - Consistent estimation of β if:
    - ◇ Use complete data
    - ◇ Allow α_t to differ between missingness patterns: "stratify on patterns"
- Selection depends on both lagged and current outcomes
  - Consistent estimation of β if:
    - ◇ Use complete data with $\tau_i = 1$ or $\tau_i = T - 1$
    - ◇ Allow α_t to differ according to τ_i: "stratify on totals"

Randomized clinical trial data

- 1216 women randomized to receive 100mg ($z_i = 0$) or 150mg ($z_i = 1$) of hormonal contraceptive by injection every 90 days
- Women recorded bleeding pattern every day (menstrual diaries)
- Outcome for each 90-day interval ($t = 1, 2, 3, 4$): amenorrhea (y_{it}) (at least 80 consecutive days without bleeding or spotting)
- Many women discontinued treatment and lost to follow-up

pattern	freq.	cumul. percent
1111	714	59
111.	84	66
11..	155	78
1...	198	95
....	65	100

y_{it} for ith interval observed if:
- Received injection before interval – may depend on $y_{i,t-1}$
- Visits at end of interval – may depend on y_{it}

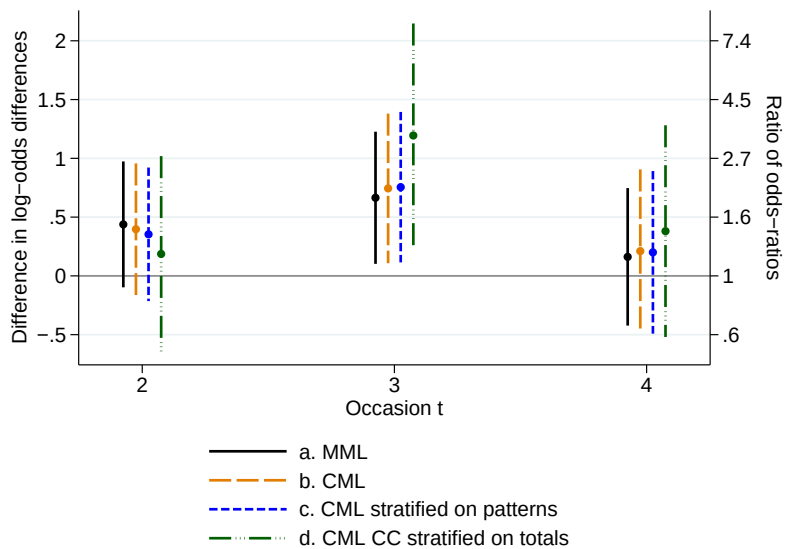
Model and estimates

$$\text{logit}\{p(y_{it} = 1 \mid z_i, \zeta_i)\} = \alpha_t + \beta_1 z_i + \beta_2 z_i T_{2t} + \beta_3 z_i T_{3t} + \beta_4 z_i T_{4t} + \zeta_i$$

	unstratified		stratified on patterns	stratified on totals CC, $\tau \neq 2$
	a: MML	b: CML	c: CML	d: CML
β_2	0.44	0.40	0.35	0.19
	(0.27)	(0.29)	(0.29)	(0.42)
β_3	0.66	0.74	0.75	1.19
	(0.29)	(0.32)	(0.33)	(0.49)
β_4	0.16	0.21	0.20	0.38
	(0.30)	(0.35)	(0.35)	(0.46)
N	1151	487	487	259

Selection at occasion t can depend on:
- a: $\mathbf{y}_i^{\text{obs}}$, $\mathbf{w}_i$
- b: $\mathbf{y}_i^{\text{mis}}$, ζ_i, $\mathbf{w}_i$ (also y_{it} if multiplicative model)
- c: $y_{i,t-1}$
- d: $y_{i,t-1}$, y_{it}

Graph of treatment effect estimates (log-odds scale)



Conclusions for multilevel logistic regression

- MML: Selection can depend on **observed outcomes**, covariates
- CML: Selection can depend on **missing outcomes, random intercepts**, covariates
- Stratified CML: Selection can depend on **current and lagged outcomes (observed or missing)**
- Useful for sensitivity analysis

Overall conclusions

- Fanciest methods not always best
- Think carefully about plausible selection mechanisms
- Check missingness patterns
- Sometimes handle missing data by creating more missing data
- Importance of sensitivity analysis

References to other authors

- Little (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* 87, 1227-1237.
- Little (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90, 1112-1121.
- Little & Zhang (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society, Series C* 60, 591-605.
- Muthén, Kaplan & Hollis (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika* 52, 431-462.
- Rubin (1976). Inference and missing data. *Biometrika* 63, 581-592.
- Vach & Illy (1997). Biased estimation of odds ratios from incomplete covariate data due to violationof the missing at random assumption. *Biometrical Journal*, 39, 13-18.
- Verbeek & Nijman (1992). Testing for selectivity bias in panel data models. *International Economic Review* 33, 681-703.

References to our work

- Skrondal & Rabe-Hesketh (2014). Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika* 101, 175-188.
- Rabe-Hesketh & Skrondal (2014). Handling missing data by creating more missing data. *Under revision*.