Multilevel and Longitudinal Modelling

Sophia Rabe-Hesketh
University of California, Berkeley
Institute of Education, London

and

Anders Skrondal
Norwegian Institute of Public Health, Oslo

Institute of Education
Bloomsbury Doctoral Training Center for the Social Sciences
June 2012

OUTLINE

I. Random intercept models (slide 3)

II. Random coefficient models (slide 31)

III. Multilevel logistic regression (slide 65)

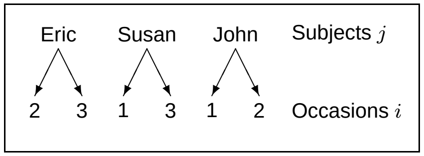
IV. Longitudinal data and alternatives to multilevel modelling (slide 95)

I. Random intercept models

- Clustered data, unobserved heterogeneity and dependence
- Random intercept models
- Intraclass correlation
- Example: GHQ test-retest data
- Estimation, testing and confidence intervals
- Empirical Bayes prediction and shrinkage
- Fixed versus random effects

Clustered data

- An important assumption in linear regression and logistic regression is that units (usually people) are independent (given covariates x)
- An important violation is due to clustered data with responses y_{ij} on units i grouped in clusters j:
  - Students i clustered in schools j
  - Siblings i clustered in families j
  - Repeated observations i clustered in people j (longitudinal, repeated measures, or panel data)



- General terms: level-1 units i clustered in level-2 units j

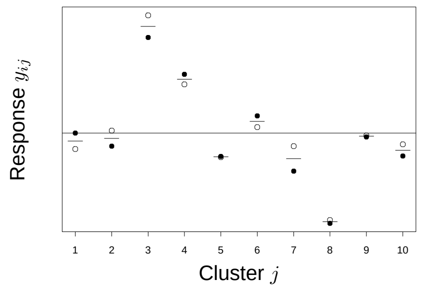
Unobserved heterogeneity

- Could not hope to explain all variability between clusters (e.g. schools) using observed covariates x
  - For instance, the school atmosphere, parents' involvement, teachers' enthusiasm and competence, etc., cannot all be measured
- Therefore there is **unobserved heterogeneity** (= unexplained variability) between clusters
- Means that two observations in same cluster are correlated and more similar than observations in different clusters
  - Students in one school tend to have better test results, even after controlling for covariates, than students in another school

Heterogeneity and dependence

- Example: No covariates, two units $i = 1, 2$ per cluster j with responses y_{ij}:

$$y_{ij} \;=\; \beta + \xi_{ij}, \quad \xi_{ij} \text{ is a residual}$$



- ● is y_{1j}
- ○ is y_{2j}
- – is the mean $\frac{1}{2}(y_{1j} + y_{2j})$

- Residuals ξ_{ij} for same cluster usually have same sign, corresponding to **within-cluster correlations** or dependence

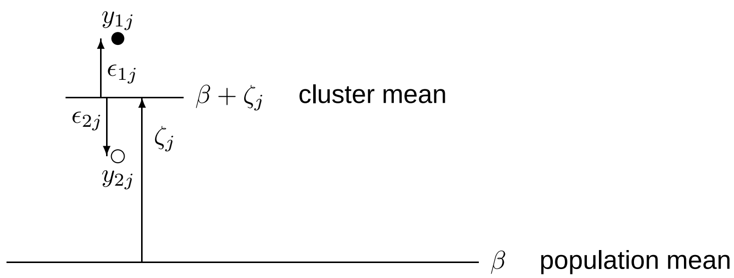
Variance-components model

- Model between-cluster heterogeneity:

$$y_{ij} \;=\; \beta + \underbrace{\zeta_j + \epsilon_{ij}}_{\xi_{ij}}$$

  - Total residual ξ_{ij} split into level-2 residual ζ_j (shared by all members of cluster) and level-1 residual (unit-specific) ϵ_{ij}
  - ζ_j, **random intercept** for cluster j
    - ◇ deviation of true cluster-mean $\beta + \zeta_j$ from overall mean β
    - ◇ independent of $\zeta_{j'}$ for other clusters j'
    - ◇ mean zero and variance ψ (a model parameter)
  - ϵ_{ij}, the **level-1 residual**
    - ◇ deviation of y_{ij} from its true cluster mean $\beta + \zeta_j$
    - ◇ independent of $\epsilon_{i'j'}$ for other i' or j' and of ζ_j and $\zeta_{j'}$
    - ◇ mean zero and variance θ (a model parameter)

Illustration of variance components model



Variance components

- Total residual or error:

$$\xi_{ij} \;=\; \zeta_j + \epsilon_{ij}$$

  - Can view ζ_j and ϵ_{ij} as **error components**

- Total residual variance:

$$\text{var}(\xi_{ij}) \;=\; \text{var}(\zeta_j) + \text{var}(\epsilon_{ij}) \;=\; \overbrace{\psi}^{\text{between}} + \overbrace{\theta}^{\text{within}}$$

  - Variances add up because ζ_j and ϵ_{ij} are independent
  - ψ and θ are therefore **variance components**

- Total variance of y_{ij}:

$$\text{var}(y_{ij}) \;=\; \text{var}(\beta + \xi_{ij}) = \text{var}(\xi_{ij}) = \psi + \theta$$

Conditional independence

- Responses conditionally independent **given random intercept**
- Zero covariance and correlation between measurements on two units i and i', given the random intercept ζ_j,

$$\text{Cor}(y_{ij}, y_{i'j}|\zeta_j) \;=\; 0$$

Intraclass correlation

- Covariance between responses on two units i and i' for the same cluster j (**not conditioning** on ζ_j)

$$\text{Cov}(y_{ij}, y_{i'j}) = \text{E}[(y_{ij} - \underbrace{\beta}_{\text{E}(y_{ij})})(y_{i'j} - \underbrace{\beta}_{\text{E}(y_{i'j})})] = \text{E}[(\zeta_j + \epsilon_{ij})(\zeta_j + \epsilon_{i'j})] = \text{E}[\zeta_j^2] = \psi$$

- Corresponding **intraclass correlation** is covariance divided by product of standard deviations

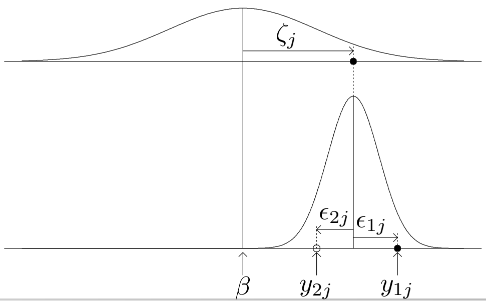
$$\text{Cor}(y_{ij}, y_{i'j}) = \frac{\text{Cov}(y_{ij}, y_{i'j})}{\sqrt{\text{Var}(y_{ij})}\sqrt{\text{Var}(y_{i'j})}} = \frac{\psi}{\sqrt{\psi+\theta}\sqrt{\psi+\theta}} = \frac{\psi}{\psi+\theta} = \rho$$

- Proportion of total variance shared among units in same cluster and therefore due to clusters (similar to coefficient of determination R^2)

$$\rho \;=\; \frac{\text{Var}(\zeta_j)}{\text{Var}(y_{ij})} = \frac{\psi}{\psi+\theta}$$

Distributional assumptions

- Assume that $\zeta_j \sim \text{N}(0, \psi)$
- Assume that $\epsilon_{ij} \sim \text{N}(0, \theta)$
- Hierarchical, two-stage model, reflecting two-stage sampling:
  - $\zeta_j \sim \text{N}(0, \psi) \Longrightarrow$ determines $\beta + \zeta_j$
  - $\epsilon_{ij} \sim \text{N}(0, \theta) \Longrightarrow$ determines $y_{ij} \;=\; \beta + \zeta_j + \epsilon_{ij}$



Parameter estimation (β, ψ, θ)

- Maximum likelihood estimation (ML)
  - If variances were known, would use GLS (generalised least squares) $\Rightarrow$ IGLS (Iterative GLS), iterating between estimation of fixed and random part
  - EM (Expectation-Maximization) algorithm: Treat random effects as missing values
- Restricted maximum likelihood estimation (REML)
  - ML gives downward biased estimate of random intercept variance
  - If cluster size is constant, $n_j = n$, REML gives unbiased estimates (if estimates allowed to be negative)
  - REML is ML applied to ‘residuals’
- Software: `MLwiN`, `HLM`, SPSS: `MIXED`, Stata: `xtmixed`, SAS: `MIXED`, R: `lmer` (all give identical estimates)

Hypothesis testing and confidence intervals

- Inference for β
  - **Wald test**: Use estimated standard error $\widehat{\text{SE}}(\widehat{\beta})$ for test statistic (and confidence interval)

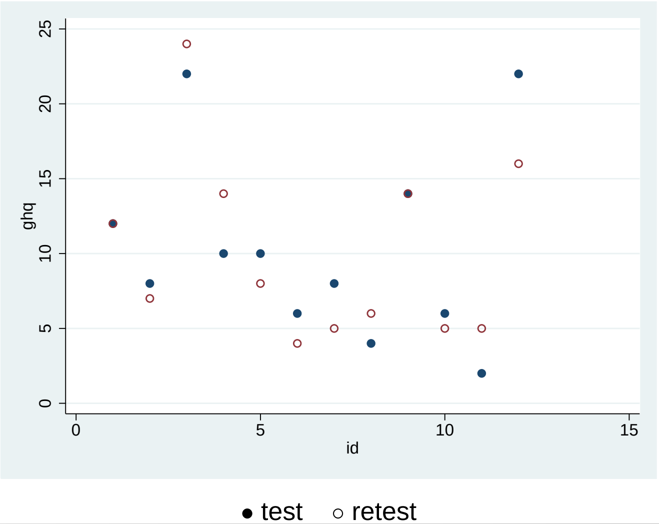
$$H_0 : \beta = \mu_0, \qquad z = \frac{\widehat{\beta} - \mu_0}{\widehat{\text{SE}}(\widehat{\beta})}$$

- Test for zero between-cluster variance $H_0 : \psi = 0$
  - **Likelihood ratio test** (DO NOT USE WALD TEST)
    - ◇ Compare log-likelihood L_1 for random-intercept model with log-likelihood L_0 for ordinary regression model (no ζ_j)
    - ◇ Test statistic $G^2 = 2(L_1 - L_0)$
    - ◇ Asymptotic sampling distribution under H_0 not $\chi^2(1)$ because null hypothesis is on boundary of parameter space since $\psi \geq 0$
    - ◇ Solution: assume $\chi^2(1)$ distribution, but divide p-value by 2

Example: GHQ test-retest data

- General Health Questionnaire (GHQ) to measure psychological distress
- Sum of 12 items, each scored 0,1, or 2
- Completed twice by 12 clinical psychology students, 3 days apart
- Variables:
  - Subject id j
  - Occasion (1:test, 2:retest) i
  - GHQ score y_{ij}

Graph for GHQ data



Maximum likelihood estimates for GHQ data

	Est	(SE)
Fixed part		
β	10.17	(1.68)
Random part		
$\sqrt{\psi}$	5.65	
$\sqrt{\theta}$	1.91	
Log-likelihood	-67.13	

Exercises: GHQ data

- Calculate the estimated intraclass correlation
- Consider the Pearson correlation between test and retest. Is this different than the intraclass correlation? If so, why?

Assigning values to random effects: Empirical Bayes prediction

- ζ_j is a **residual** like ϵ_{ij}
- ζ_j is a random variable, not a model parameter
- As in ordinary regression, sometimes want to **predict** residuals
- Reasons for predicting ζ_j:
  - Residual diagnostics
  - Inference for cluster-mean $\beta + \zeta_j$ or ζ_j
    - Measurement (e.g., GHQ): $\beta + \zeta_j$ is "true score"
    - Institutional performance: ζ_j is "value added"
  - Model interpretation

Assigning values to random effects: Empirical Bayes prediction

- Treat parameter estimates $\widehat{\beta}$, $\widehat{\psi}$ and $\widehat{\theta}$, as known parameter values
- For cluster j, empirical Bayes combines
  1. **Prior distribution** of ζ_j, knowledge about ζ_j before seeing data for the cluster

     $$\text{Prior}(\zeta_j) \qquad \left[\text{normal density } g(0, \widehat{\psi})\right]$$

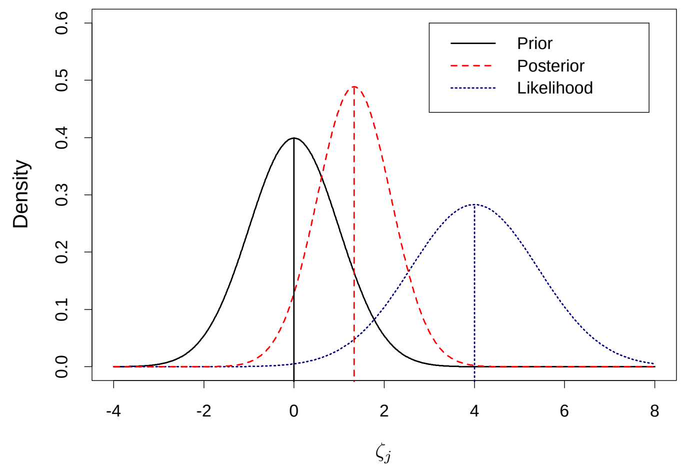
  2. **Likelihood**, knowledge about ζ_j provided by the data $\mathbf{y}_j$ (and $\mathbf{X}_j$)

     $$\text{Likelihood}(\mathbf{y}_j|\zeta_j) \qquad \left[\prod_{i=1}^{n_j} g(\widehat{\beta} + \zeta_j, \widehat{\theta})\right]$$

- To obtain posterior distribution of random intercept (Bayes Theorem)

$$\text{Posterior}(\zeta_j|\mathbf{y}_j) \propto \text{Prior}(\zeta_j) \times \text{Likelihood}(\mathbf{y}_j|\zeta_j)$$

Empirical Bayes prediction (cont'd)



- Empirical Bayes prediction $\widetilde{\zeta}_j = 1.33$ is **mean of posterior distribution**

Fixed instead of random effects of clusters

- Can view clusters as categories of categorical explanatory variable
- **Fixed effects of cluster**: dummy variable d_{mj} for cluster j (no intercept)

$$y_{ij} = \overbrace{\sum_{m=1}^{J} \alpha_m d_{mj}}^{\alpha_j} + \epsilon_{ij}, \quad d_{mj} = \begin{cases} 1 & \text{if } m = j \\ 0 & \text{if } m \neq j \end{cases} \quad \epsilon_{ij} \sim \mathrm{N}(0, \theta)$$

  - α_j are fixed parameters, representing clusters' population means
  - ϵ_{ij} is a random error term, representing within-cluster variability
- **Random effects of cluster**:

$$y_{ij} = \beta + \zeta_j + \epsilon_{ij}, \qquad \zeta_j \sim \mathrm{N}(0, \psi), \quad \epsilon_{ij} \sim \mathrm{N}(0, \theta)$$

  - β is a fixed parameter, the **population mean**
  - ζ_j and ϵ_{ij} are random error terms

Fixed effects approach for GHQ data (cont'd)

	FE	EST	(SE)	RE	EB	(SE)
Fixed part	α_1	12	(1.35)	$\beta+\zeta_1$	11.9	(1.32)
	α_2	7.5	(1.35)	$\beta+\zeta_2$	7.6	(1.32)
	α_3	23.0	(1.35)	$\beta+\zeta_3$	22.3	(1.32)
	α_4	12.0	(1.35)	$\beta+\zeta_4$	11.9	(1.32)
	α_5	9.0	(1.35)	$\beta+\zeta_5$	9.1	(1.32)
	α_6	5.0	(1.35)	$\beta+\zeta_6$	5.3	(1.32)
	α_7	6.5	(1.35)	$\beta+\zeta_7$	6.7	(1.32)
	α_8	5.0	(1.35)	$\beta+\zeta_8$	5.3	(1.32)
	α_9	14.0	(1.35)	$\beta+\zeta_9$	13.8	(1.32)
	α_{10}	5.5	(1.35)	$\beta+\zeta_{10}$	5.8	(1.32)
	α_{11}	3.5	(1.35)	$\beta+\zeta_{11}$	3.9	(1.32)
	α_{12}	19.0	(1.35)	$\beta+\zeta_{12}$	18.5	(1.32)
Random part	θ	3.7				

- 13 parameters (θ and 12 α_j) for fixed-effects model, compared with 3 parameters (θ, β, ψ) for random-effects model
- In random-effects model, use **empirical Bayes** to assign values to cluster means $\beta + \zeta_j$

Maximum likelihood estimation of cluster-specific effects

- Estimated coefficients $\widehat{\alpha}_j$ of dummies are ML estimates of $\beta + \zeta_j$
  - **Maximum likelihood estimates** of ζ_j, maximum of Likelihood$(\mathbf{y}_j|\zeta_j)$ with $\widehat{\beta}$ treated as known
  - Also called OLS (Ordinary Least Squares) estimates
  - Simply the cluster means of the estimated total residuals $\widehat{\xi}_{ij}$

$$\widehat{\xi}_{ij} = y_{ij} - \widehat{\beta} = \widehat{\zeta_j + \epsilon_{ij}}$$

$$\widehat{\zeta}_j^{\mathrm{ML}} = \frac{1}{n_j} \sum_{i=1}^{n_j} \widehat{\xi}_{ij}$$

Shrinkage

- Empirical Bayes prediction of random intercept can be written as

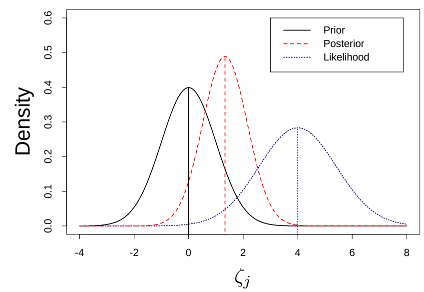
$$\widetilde{\zeta}_j^{\mathrm{EB}} = \widehat{R}_j \widehat{\zeta}_j^{\mathrm{ML}}, \qquad \widehat{R}_j = \frac{\widehat{\psi}}{\widehat{\psi} + \widehat{\theta}/n_j}$$

  - $\widehat{R}_j$ is estimated 'reliability' of ML estimator (true score variance divided by total variance of $\widehat{\zeta}_j^{\mathrm{ML}}$)
  - $\widehat{R}_j$ is **shrinkage factor**, shrinking prediction towards 0 (mean of prior) since $0 \leq \widehat{R}_j \leq 1$
  - More shrinkage (i.e. greater influence of prior) if
    - Small random intercept variance $\widehat{\psi}$ (informative prior)
    - Large level-1 residual variance $\widehat{\theta}$ (non-informative data)
    - Small cluster size n_j (non-informative data)

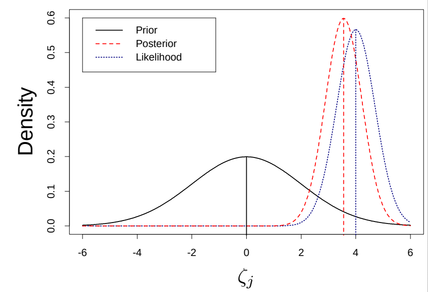
Illustration: Shrinkage

- Cluster with $n_j = 2$ units
- Predicted total residuals $\widehat{\xi}_{1j} = 3$ and $\widehat{\xi}_{2j} = 5$

Much shrinkage:
$\widehat{\psi} = 1, \widehat{\theta} = 4 \Rightarrow \widehat{R}_j = 0.33$



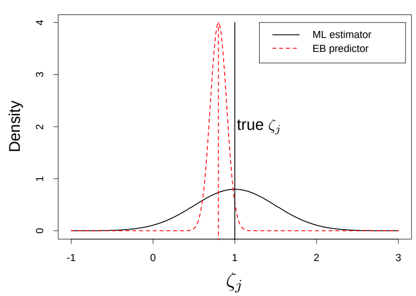
Little shrinkage:
$\widehat{\psi} = 4, \widehat{\theta} = 1 \Rightarrow \widehat{R}_j = 0.89$



Bias/precision trade-off

- EB prediction conditionally biased towards zero unlike ML: For a given cluster, mean EB prediction (over repeated samples of units) closer to zero than true random intercept
- EB has smaller prediction error variance (mean squared error) than ML, i.e. more accurate, especially for small clusters

Example:
Sampling
Distributions



- Also called 'Best Linear Unbiased Predictor' (BLUP)

"Borrowing strength" or partial pooling

- EB for cluster j 'borrows strength' from other clusters
- Estimate of true cluster mean $\beta + \zeta_j$ is:
  - ML:

$$\widehat{\beta} + \widehat{\zeta}_j^{\mathrm{ML}} = \widehat{\beta} + \frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ij} - \widehat{\beta}) = \widehat{\beta} + (\bar{y}_{\cdot j} - \widehat{\beta}) = \bar{y}_{\cdot j}$$

  $\Longrightarrow$ sample mean of cluster j

  - EB:

$$\widehat{\beta} + \widetilde{\zeta}_j^{\mathrm{EB}} = \widehat{\beta} + \widehat{R}_j \widehat{\zeta}_j^{\mathrm{ML}} = \widehat{\beta} + \widehat{R}_j(\bar{y}_{\cdot j} - \widehat{\beta}) = (1 - \widehat{R}_j)\widehat{\beta} + \widehat{R}_j \bar{y}_{\cdot j}$$

  $\Longrightarrow$ weighted mean of:
  sample mean of cluster j and $\widehat{\beta}$, estimate based on all clusters

Fixed versus random effects

Issue	Fixed effects	Random effects
Inference for population of clusters	No $-$	Yes $+$
Number of clusters required	Any number $+$	At least 10 or 20 $-$
Assumptions	None for distribution of intercepts $+$	Intercepts normal, constant variance, etc. $-$
Inference for individual clusters	Yes $+$	Yes, empirical Bayes $+$
Cluster sizes required	Any sizes if many ≥ 2, but overfitting if small $\pm$	Any sizes if many ≥ 2 $+$
Parsimony	A parameter α_j for each cluster $-$	One variance parameter ψ for all cluster $+$

- **Note:** Further issues if there are covariates and for generalized linear mixed models

Exercise: Fixed versus random

- In each situation below, should fixed or random effects be used?

1. Math achievement, 3 schools, 30 to 40 students per school
2. Reading test, 43 countries, about 2000 students per country
3. Longitudinal data on 20 subjects, 3 observations per subject
4. Blood pressure, 10 treatment groups, 20 patients per group
5. Depression, 15 therapists, 3-15 patients per therapist

II. Random coefficient models

- Random intercept model with covariates
- Example: Georgian birthweights
- Between effects, within effects and endogeneity
- Random coefficients

Random intercept model with covariate

- Add covariate to variance components model:

$$y_{ij} = \underbrace{\beta_1 + \beta_2 x_{ij}}_{\text{fixed part}} + \underbrace{\zeta_j + \epsilon_{ij}}_{\text{random part}}$$

- Intercept varies between clusters:

$$y_{ij} = \underbrace{\beta_1 + \zeta_j}_{\substack{\text{intercept} \\ \text{for cluster } j}} + \beta_2 x_{ij} + \epsilon_{ij}$$

Assumptions for random intercept model with covariate

- Assumptions for ϵ_{ij} and ζ_j:
  - $E(\epsilon_{ij}|\zeta_j, \mathbf{X}_j) = 0$
    - $\Rightarrow$ Cov$(\epsilon_{ij}, \mathbf{X}_j) = 0$ [level-1 exogeneity]
    - $\Rightarrow$ variance decomposition
  - ϵ_{ij} independent over units i and clusters j
    $\Rightarrow$ conditional independence of responses given random intercept
  - $E(\zeta_j|\mathbf{X}_j) = 0$
    $\Rightarrow$ Cov$(\zeta_j, \mathbf{X}_j) = 0$ [level-2 exogeneity]
  - ζ_j independent for different j
    $\Rightarrow$ independent clusters in likelihood
- Distributional assumptions (for maximum likelihood):
  - ϵ_{ij} normal with zero mean and variance θ
  - ζ_j normal with zero mean and variance ψ

Regression lines

- Population averaged or **marginal** regression line (mean over population of clusters and populations of units within clusters)

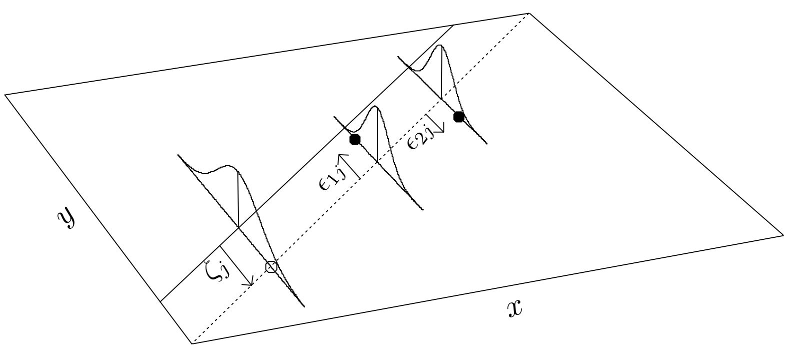
$$\mathrm{E}(y_{ij}|x_{ij}) \;=\; \beta_1 + \beta_2 x_{ij}$$

- Cluster-specific or **conditional** regression line (mean over population of units within cluster j)

$$\begin{aligned}\mathrm{E}(y_{ij}|x_{ij}, \zeta_j) &= \beta_1 + \beta_2 x_{ij} + \zeta_j \\ &= (\beta_1 + \zeta_j) + \beta_2 x_{ij}\end{aligned}$$

  - ψ is variance between cluster-specific intercepts $\beta_1 + \zeta_j$
  - θ is variance of y_{ij} around cluster-specific regression lines

Illustration of random intercept model with covariate



Example: Georgia birthweights

- 878 mothers of five children in Georgia, USA:
  - Child's birth weight in grams y_{ij}
  - Mother's age at the time of the child's birth x_{ij}
- Random intercept model:

$$y_{ij} \;=\; \beta_1 + \beta_2 x_{ij} + \zeta_j + \epsilon_{ij}$$

  - With the usual assumptions stated on slide 33

Estimates for Georgia birthweights (cont'd)

	with age		without age	
	Est	(SE)	Est	(SE)
Fixed part				
β_1	2785.2	(45.2)	3156.3	(14.1)
β_2 [age]	17.1	(2.0)		
Random part				
$\sqrt{\psi}$	354.6		368.4	
$\sqrt{\theta}$	434.2		435.5	
Log-likelihood	$-$33535.7		$-$33572.3	

Between and within-cluster covariates

- Covariates may vary
  - Between clusters, e.g., mother's own birthweight
  - Within clusters, e.g., children's parity (birth order) 1,2,3,4,5
  - Both between and within clusters, e.g., mother's age at birth
    - **Between-cluster variability**: Standard deviation of cluster mean age around overall mean is 3.7
    - **Within-cluster variability**: Standard deviation of age around cluster means is 2.8
    - **Overall variability**: Conventional standard deviation (ignoring clustering) is 4.6

Between and within-cluster effects of covariates

- Previous model:

$$y_{ij} \;=\; \beta_1 + \beta_2 x_{ij} + \zeta_j + \epsilon_{ij}$$

- Coefficient β_2 represents difference in mean birth weight for children whose mothers differ in age by one year
- Two types of comparisons or effects:
  - **Within-mother effect**:
    Same mother, children born at different times (ages)
  - **Between-mother effect**:
    Different mothers giving birth at different ages
- Model assumes that both effects are the same

Between and within cluster effects

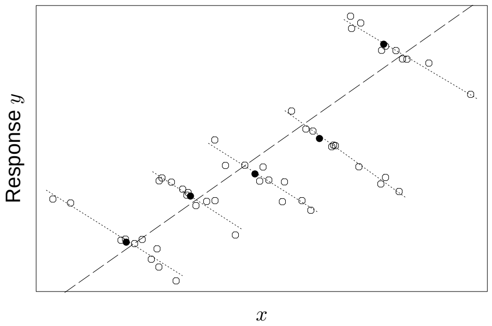
- **Between effect**: Take cluster average of random intercept model

$$\begin{aligned} \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} &= \frac{1}{n_j}\sum_{i=1}^{n_j}[\beta_1 + \beta_2 x_{ij} + \zeta_j + \epsilon_{ij}] \\ \bar{y}_{\cdot j} &= \beta_1 + \beta_2 \bar{x}_{\cdot j} + \underbrace{\zeta_j + \bar{\epsilon}_{\cdot j}}_{e_j} \end{aligned}$$

- **Within effect**: Subtract cluster average random intercept model from random intercept model

$$\begin{aligned} y_{ij} &= [\beta_1 + \beta_2 x_{ij} + \zeta_j + \epsilon_{ij}] \\ -\bar{y}_{\cdot j} &= -[\beta_1 + \beta_2 \bar{x}_{\cdot j} + \zeta_j + \bar{\epsilon}_{\cdot j}] \\ \hline y_{ij} - \bar{y}_{\cdot j} &= \beta_2 (x_{ij} - \bar{x}_{\cdot j}) + \underbrace{\epsilon_{ij} - \bar{\epsilon}_{\cdot j}}_{e_{ij}} \end{aligned}$$

Illustration of different between and within-effects



- Hollow circles: individual units (x_{ij}, y_{ij})
- Dotted lines: within-cluster regression, slope is within-cluster effect
- Solid circles: cluster means $(\bar{x}_{\cdot j}, \bar{y}_{\cdot j})$
- Dashed line: between-cluster regression, slope is between-cluster effect
- Simpson's paradox, cluster-level confounding, **ecological fallacy**

Between and within-cluster estimates for Georgia birthweights

	Between		Within	
	Est	(SE)	Est	(SE)
Fixed part				
β_1	2499.1	(80.7)	2900.1	(51.1)
β_2 [age]	30.4	(3.7)	11.8	(2.3)

- Estimated between-effect much larger than within-effect
- Advantage of clustered data:
  Can distinguish between different kinds of effects!

Exercise: Between and within-effects

- Explain why you think there is a difference between the within and between-effects of mother's age on birth weight

Cluster-level confounding and endogeneity

- Random intercept model (equal between and within-cluster effects)

$$\begin{aligned} y_{ij} &= \beta_1 + \beta_2 x_{ij} + \zeta_j + \epsilon_{ij} \\ &= \beta_1 + \beta_2(x_{ij} - \bar{x}_{\cdot j}) + \beta_2 \bar{x}_{\cdot j} + \zeta_j + \epsilon_{ij} \end{aligned}$$

- Random intercept model assumes exogenous covariate (important if β_2 interpreted as causal effect of x_{ij} on y_{ij})
  - x_{ij} uncorrelated with ζ_j (no cluster-level confounding)
    - $\bar{x}_{\cdot j}$ uncorrelated with ζ_j
    - Assumption not made in within-cluster regression
      $y_{ij} - \bar{y}_{\cdot j} = \beta_2(x_{ij} - \bar{x}_{\cdot j}) + \epsilon_{ij} - \bar{\epsilon}_{\cdot j}$
  - x_{ij} uncorrelated with ϵ_{ij} (no unit-level confounding)
    - $(x_{ij} - \bar{x}_{\cdot j})$ uncorrelated with ϵ_{ij}
- Within-cluster estimate not subject to cluster-level confounding – closer to causal effect?

Allowing and testing for endogeneity

- Concern about bias due to correlation between ζ_j and x_{ij} (especially among economists who call this **endogeneity**)
  - Use within-effect estimator or modify random intercept model:

$$y_{ij} = \beta_1 + \beta_{2w}(x_{ij} - \bar{x}_{\cdot j}) + \beta_{2b}\bar{x}_{\cdot j} + \zeta_j + \epsilon_{ij}$$

    - β_{2w} is within-effect and β_{2b} is between-effect
  - If Cor$(x_{ij}, \zeta_j) \neq 0$
    - $\widehat{\beta}_{2b}$ inconsistent since Cor$(\bar{x}_{\cdot j}, \zeta_j) \neq 0$
    - $\widehat{\beta}_{2w}$ consistent since Cor$((x_{ij} - \bar{x}_{\cdot j}), \zeta_j) = 0$ and Cor$((x_{ij} - \bar{x}_{\cdot j}), \bar{x}_{\cdot j}) = 0$
  - Test of $H_0: \beta_{2w} = \beta_{2b}$, highly significant, $p < 0.001$
  - This test is equivalent to famous Hausman test in econometrics

Fixed instead of random effects of clusters

- Regression with dummy variables d_{mj} for each cluster (and no intercept) – ANCOVA model

$$y_{ij} = \overbrace{\sum_{m=1}^{J} \alpha_m d_{mj}}^{\alpha_j} + \beta_2 x_{ij} + \epsilon_{ij}, \quad d_{mj} = \begin{cases} 1 & \text{if } m = j \\ 0 & \text{if } m \neq j \end{cases}$$

  - Any between-cluster covariate z_j or $\bar{x}_{\cdot j}$ completely collinear with set of dummy variables, i.e., can be written as linear combination of dummy variables:

$$z_j = \sum_{m=1}^{J} z_m d_{mj} \qquad x_{\cdot j} = \sum_{m=1}^{J} x_{\cdot m} d_{mj}$$

    - Cannot include between-cluster covariates
    - **Estimate of β_2 is within-effect**; between-effect absorbed in α_1 to α_J

Fixed versus random effects revisited

Issue	**Fixed effects**		**Random effects**	
Inference for population of clusters	No	–	Yes	+
Number of clusters required	Any number	+	At least 10 or 20	–
Assumptions	None for distribution of intercepts	+	Intercepts normal, constant variance, etc.	–
Effects of cluster-level covariates	No	–	Yes	+
Inference for individual clusters	Yes	+	Yes, Empirical Bayes	+
Cluster sizes required	Any sizes if many ≥ 2, but overfitting if small	$\pm$	Any sizes if many ≥ 2	+
Parsimony	A parameter α_j for each cluster	–	One variance parameter ψ for all cluster	+
Within-cluster effects of covariates	Yes	+	Only with extra work	–

Random coefficient models

- Not only the overall level of the response (intercept) can vary between clusters, but also the slopes of within-cluster covariates
- Simple example:

$$\begin{aligned} y_{ij} &= \overbrace{\beta_1 + \zeta_{1j}}^{\text{intercept}} + \overbrace{(\beta_2 + \zeta_{2j})}^{\text{slope}} x_{ij} + \epsilon_{ij} \\ &= \underbrace{\beta_1 + \beta_2 x_{ij}}_{\text{fixed part}} + \underbrace{\zeta_{1j} + \zeta_{2j} x_{ij} + \epsilon_{ij}}_{\text{random part}} \end{aligned}$$

  - ζ_{1j} is random intercept: Deviation of cluster-specific intercept from mean intercept
  - ζ_{2j} is random slope: Deviation of cluster-specific slope from mean slope

Assumptions for random coefficient models

- Exogeneity assumptions analogous to random intercept model
- Distributional assumptions (for maximum likelihood):
  - ϵ_{ij} normal with zero mean and variance θ
  - (ζ_{1j}, ζ_{2j}) bivariate normal with zero means and unstructured covariance matrix (variances ψ_{11} and ψ_{22} and covariance ψ_{21})

Regression lines

- Population averaged or **marginal** regression line (mean over population of clusters and populations of units within clusters)

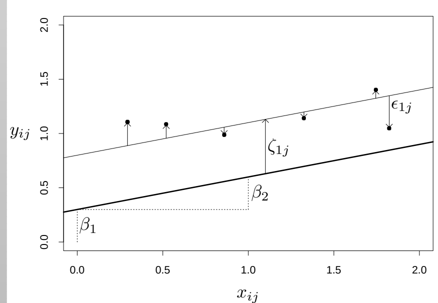
$$\mathrm{E}(y_{ij}|x_{ij}) = \beta_1 + \beta_2 x_{ij}$$

- Cluster-specific or **conditional** regression line (mean over population of units within cluster j)

$$\begin{aligned}\mathrm{E}(y_{ij}|x_{ij}, \zeta_{1j}, \zeta_{2j}) &= \beta_1 + \beta_2 x_{ij} + \zeta_{1j} + \zeta_{2j} x_{ij} \\ &= (\beta_1 + \zeta_{1j}) + (\beta_2 + \zeta_{2j}) x_{ij}\end{aligned}$$

Illustration of random coefficient model

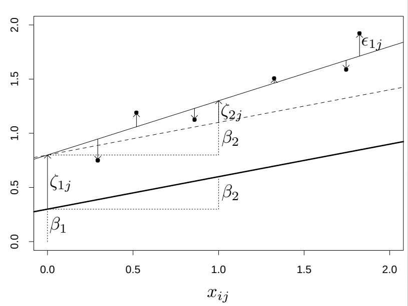
Random intercept model



$$y_{ij} = (\beta_1 + \zeta_{1j}) + \beta_2 x_{ij} + \epsilon_{ij}$$

ζ_{1j}: vertical shift

Random coefficient model



$$y_{ij} = (\beta_1 + \zeta_{1j}) + (\beta_2 + \zeta_{2j}) x_{ij} + \epsilon_{ij}$$

ζ_{1j}: vertical shift at $x_{ij} = 0$

Parameters of random part

- Four unique parameters for random part:
  - Unstructured covariance matrix of intercepts ζ_{1j} and slopes ζ_{2j}:

$$\begin{bmatrix} \mathrm{Var}(\zeta_{1j}) & \mathrm{Cov}(\zeta_{1j}, \zeta_{2j}) \\ \mathrm{Cov}(\zeta_{2j}, \zeta_{1j}) & \mathrm{Var}(\zeta_{2j}) \end{bmatrix} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}, \qquad \psi_{21} = \psi_{12}$$

  - Variance of level-1 residuals ϵ_{ij}: θ
- Easier to interpret standard deviations $\sqrt{\psi_{11}}$, $\sqrt{\psi_{22}}$, $\sqrt{\theta}$ and correlation ρ_{21}

$$\rho_{21} = \frac{\psi_{21}}{\sqrt{\psi_{11}\psi_{22}}}$$

Two-stage formulation

- Raudenbush and Bryk (R&B) define multilevel model in stages:
  - **Level-1 model** with cluster-specific coefficients and unit-specific covariates:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + r_{ij}$$

  - **Level-2 models** for cluster-specific coefficients with cluster-specific covariates:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}w_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}w_j + u_{1j}\end{aligned}$$

    - ◇ 'Intercepts and slopes as outcomes'

Reduced form model

- Substitute level-2 models into level-1 model:

$$\begin{aligned}y_{ij} &= \underbrace{\gamma_{00} + \gamma_{01}w_j + u_{0j}}_{\beta_{0j}} + \underbrace{(\gamma_{10} + \gamma_{11}w_j + u_{1j})}_{\beta_{1j}}x_{ij} + \epsilon_{ij}\\ &= \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + \gamma_{11}w_jx_{ij} + u_{0j} + u_{1j}x_{ij} + \epsilon_{ij}\\ &\equiv \beta_1 + \beta_2x_{ij} + \beta_3w_j + \beta_4w_jx_{ij} + \zeta_{1j} + \zeta_{2j}x_{ij} + \epsilon_{ij}\end{aligned}$$

  - γ_{11} (or β_4) represents a **cross-level interaction** between w_j (level 2) and x_{ij} (level 1)

Example: Inner London Schools

- Inner London School data (65 schools)
  - Graduate Certificate of Secondary Education (GCSE) score (age 16) y_{ij}
  - London Reading Test (LRT) score before entering school (age 11) x_{ij}
  - GCSE and LRT standardized to mean=0, sd=10 (in larger sample)
- Model:

$$y_{ij} = \underbrace{(\beta_1 + \zeta_{1j})}_{\text{Intercept for school } j} + \underbrace{(\beta_2 + \zeta_{2j})}_{\text{Slope for school } j} x_{ij} + \epsilon_{ij}$$

  - With the usual assumptions

Maximum likelihood estimates for random intercept (RI) and random coefficient (RC) models

	RI Model		RC Model	
Parameter	Est	(SE)	Est	(SE)
Fixed part				
β_1	0.02	(0.40)	-0.12	(0.40)
β_2 [LRT]	0.56	(0.01)	0.56	(0.02)
Random part				
$\sqrt{\psi_{11}}$	3.04		3.01	
$\sqrt{\psi_{22}}$			0.12	
ρ_{21}			0.50	
$\sqrt{\theta}$	7.52		7.44	
Log-likelihood	-14024.80		-14004.61	

Test for zero slope variance

- H_0: $\psi_{22} = 0$ $(\Rightarrow \psi_{21} = 0)$; in other words $\zeta_{2j} = 0$ for all j
- If null hypothesis is true, likelihood ratio (or deviance) statistic G^2 usually has a χ^2 distribution with degrees of freedom equal to difference in number of parameters, here 2 $\Rightarrow$ p-value is < 0.001
- However, for variance component ψ_{22}, null hypothesis is on boundary of parameter space since $\psi_{22} \geq 0$
- Sampling distribution of G^2 under null hypothesis a 1:1 mixture of $\chi^2(2)$ and mass at 0
  $\Rightarrow$ divide p-value of conventional test by 2
  - p-value based on χ^2 distribution with d.f.$= 2$ is $p < 0.001$
  - Dividing by 2 gives same conclusion: random intercept model rejected in favor of random coefficient model

Predicted school-specific regression lines

$$\widehat{\mu}_{ij} \;=\; \widehat{\beta}_1 + \widehat{\beta}_2 x_{ij} + \widetilde{\zeta}_{1j} + \widetilde{\zeta}_{2j} x_{ij}$$

$\widetilde{\zeta}_{1j}$, $\widetilde{\zeta}_{2j}$ are **empirical Bayes** predictions

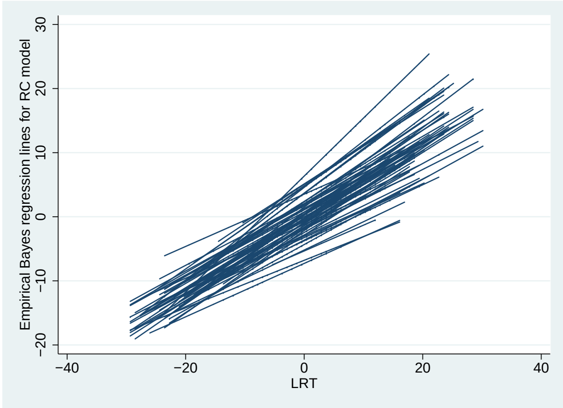
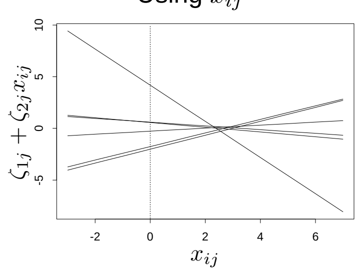


Illustration: Lack of invariance to translation and heteroscedasticity

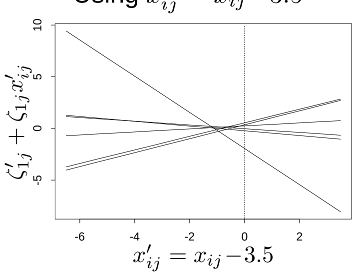
- Graphs of cluster-specific regression lines (with $\beta_1 = \beta_2 = 0$), illustrating effect of translation of x_{ij}:

Using x_{ij}



Large ψ_{11}, negative ψ_{21}

Using $x'_{ij} = x_{ij} - 3.5$



Small ψ_{11}, positive ψ_{21}

- Variance of $\zeta_{1j} + \zeta_{2j}x_{ij}$, and hence of total residual ξ_{ij} decreases with x_{ij} and increases again

Interpreting random part

- $\sqrt{\psi_{11}}$: Standard deviation of **intercepts**
  - Has same units (scale) as y_{ij} and β_1
    $\Rightarrow$ estimate rescaled when y_{ij} rescaled
    $\Rightarrow$ 95% of clusters expected to have intercepts in range $\beta_1 \pm 1.96\sqrt{\psi_{11}}$
  - Is standard deviation of vertical positions of cluster-specific regression lines were $x_{ij} = 0$
    $\Rightarrow$ estimate changes if x_{ij} translated (e.g., mean-centered)
- $\sqrt{\psi_{22}}$: Standard deviation of **slopes**
  - Has same units as β_2 (units of y_{ij} divided by units of x_{ij})
    $\Rightarrow$ cannot compare directly with $\sqrt{\psi_{11}}$
    $\Rightarrow$ estimate rescaled if either x_{ij} or y_{ij} are rescaled
    $\Rightarrow$ 95% of clusters expected to have slopes in range $\beta_2 \pm 1.96\sqrt{\psi_{22}}$

Interpreting random part (cont'd)

- ρ_{21}: Correlation between intercepts and slopes
  - Has no units ($-1 \leq \rho_{21} \leq 1$)
  - Is tendency for clusters with large intercepts to have large slopes $\Rightarrow$ estimate changes if x_{ij} translated
  - **Note:** Never set $\rho_{21} = 0$ (non-equivalent models if x_{ij} translated)
- $\sqrt{\theta}$: Standard deviation of level-1 residual ϵ_{ij}
  - Has same units as y_{ij}, β_1 and $\sqrt{\psi_{11}}$ $\Rightarrow$ estimate rescaled if y_{ij} rescaled
  - Is amount of scatter around cluster-specific regression lines
- **Note:** Since the scaling if y_{ij} and x_{ij} and the translation of x_{ij} matter for interpreting the random part, make meaningful choices
  - e.g., if x_{ij} is annual income in \$, express it as number of thousands above the average, i.e., generate transformed variable $z_{ij} = \frac{x_{ij}-\bar{x}_{..}}{1000}$

Interpreting random part for Inner London Schools

Parameter	Est	(SE)
β_1	−0.12	(0.40)
β_2 [LRT]	0.56	(0.02)
$\sqrt{\psi_{11}}$	3.01	
$\sqrt{\psi_{22}}$	0.12	
ρ_{21}	0.50	
$\sqrt{\theta}$	7.44	

- 95% of intercepts are in the range -6.0 to 5.8 ($-0.12 \pm 1.96 \times 3.01$)
- 95% of slopes are in the range 0.32 to 0.80 ($0.56 \pm 1.96 \times 0.12$)
- When LRT is at its mean, the SD of the school means is 3.01, less than half the within-school SD of 7.44

Warnings about random coefficient models

- Random slope rarely makes sense if there is no random intercept (just like interactions don't make sense without main effects)
- Random slope rarely makes sense without corresponding fixed slope (estimating variance, but constraining mean to zero)
- Models with several random slopes can be hard to estimate
  - With k random slopes (plus 1 random intercept) there are $(k+2)(k+1)/2+1$ parameters in random part e.g., $k=3$ gives 11 parameters in random part
  - Clusters may not provide much information on cluster-specific slopes or their variance if
    - Clusters are small
    - x_{ij} does not vary much within clusters, or varies only in a small number of clusters

Warnings about random coefficient models (cont'd)

- Variance-covariance matrix in random part may (try to) become non 'positive semi-definite' (e.g., negative variances, correlations greater than 1 or less than -1)
  If software does not allow this, get convergence problems
  - It may help to translate and rescale x_{ij}, or to simplify the model
- Overall message: Include random slopes only where strongly suggested by theory

III. Multilevel logistic regression

- Introduction to ordinary logistic regression
- Random intercept logistic regression
- Conditional and marginal relationships

Example: Attitudes to women's roles

- U.S. General Social Survey (GSS), independent samples in 1982 and 1994
- Responses to the question "Do you agree or disagree with this statement?"
  - "Women should take care of running their homes and leave running the country to men"

Year	Agree ($y_i = 1$)	Disagree ($y_i = 0$)	Total
1982 ($x_i = 0$)	122	223	345
1994 ($x_i = 1$)	268	1632	1900
Total	390	1855	2245

- x_i is a dummy variable for year being 1994

Probabilities, odds and odds ratio

- Probability $\equiv$ Proportion of people agreeing in population (Expected number of successes per trial)

$$0 \leq \text{Pr}(y_i = 1) \leq 1$$

  - Probability of agreeing in 1982 estimated as $\widehat{\text{Pr}}(y_i = 1|x_i = 0) = \frac{122}{345} = 0.354$
  - Probability of agreeing in 1994 estimated as $\widehat{\text{Pr}}(y_i = 1|x_i = 1) = \frac{268}{1900} = 0.141$
- Odds $\equiv$ Number of people agreeing per person disagreeing in population (Expected number of successes per failure)

$$0 \leq \text{Odds}(y_i = 1) \leq \infty$$

  - Odds of agreeing in 1982 estimated as $\widehat{\text{Odds}}(y_i = 1|x_i = 0) = \frac{122}{223} = 0.547$
  - Odds of agreeing in 1994 estimated as $\widehat{\text{Odds}}(y_i = 1|x_i = 1) = \frac{268}{1632} = 0.164$
- Odds ratio (OR) $= \dfrac{\text{Odds}(y_i=1|x_i=1)}{\text{Odds}(y_i=1|x_i=0)}$
  - Odds ratio is estimated as $\widehat{\text{OR}} = \frac{0.164}{0.547} = 0.300$

Logistic regression

- Logistic regression

$$\text{Pr}(y_i = 1|x_i) \;=\; \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \;=\; \frac{\text{Odds}(y_i = 1|x_i)}{1 + \text{Odds}(y_i = 1|x_i)}$$

- Log-odds

$$\log\left[\text{Odds}(y_i = 1|x_i)\right] \;\equiv\; \text{logit}[\text{Pr}(y_i = 1|x_i)] \;=\; \beta_1 + \beta_2 x_i$$

- Difference in log-odds for unit change in x_i (from a to $a+1$)

$$\log\left[\text{Odds}(y_i = 1|x_i = a+1)\right] - \log\left[\text{Odds}(y_i = 1|x_i = a)\right]$$
$$= [\beta_1 + \beta_2(a + 1)] - [\beta_1 + \beta_2 a] = \beta_2$$

- Odds ratio for unit change in x_i (from a to $a+1$)

$$\frac{\text{Odds}(y_i = 1|x_i = a+1)}{\text{Odds}(y_i = 1|x_i = a)} \;=\; \exp(\beta_2)$$

Example: *Logistic regression for attitudes to women's roles*

- Variables:
  - Dummy for year being 1994 (x_i)
  - Agreeing with statement (y_i)
- Maximum likelihood estimates:

	Est	(SE)	OR = exp(β)	(95% CI)
β_1	−0.60	(0.11)		
β_2 [1994]	−1.20	(0.13)	0.30	(0.23,0.39)

- 95% CI for OR is $\exp(\widehat{\beta}_2 - 1.96\text{SE}_{\widehat{\beta}_2}), \exp(\widehat{\beta}_2 + 1.96\text{SE}_{\widehat{\beta}_2})$
- Standard error for odds ratio not useful

Logistic regression as generalized linear model

- Linear predictor:

$$\nu_i \equiv \beta_1 + \beta_2 x_i$$

- Conditional expectation of y_i:

$$\mu_i \equiv \text{E}(y_i|x_i) = \text{E}(y_i|\nu_i)$$

  - For continuous responses, this is the population mean
  - For dichotomous responses (0,1), this is the probability $\Pr(y_{ij} = 1|\nu_i)$

Logistic regression as generalized linear model (cont'd)

- Link function $g()$ linking conditional expectation to linear predictor:

$$g(\mu_i) = \nu_i$$

  - Linear regression: $\mu_i = \nu_i$ (identity link)
  - Logistic regression: $\text{logit}(\mu_i) \equiv \log\left[\frac{\mu_i}{1-\mu_i}\right] = \nu_i$ (logit link)
  - Probit regression: $\Phi^{-1}(\mu_i) = \nu_i$ (probit link)
- Distribution of y_i given μ_i from exponential family:
  - Linear regression: Normal with mean μ_i and constant variance θ
  - Logit and probit: Bernoulli with probability μ_i (or binomial $B(1, \mu_i)$) – variance is $\mu_i(1 - \mu_i)$

Latent response y_i^*

- A continuous latent (unobserved) response y_i^* is often assumed to underlie the observed dichotomous response y_i
- Observed response $y_i = 1$ if latent response y_i^* exceeds threshold 0 and $y_i = 0$ otherwise
  - When asked to 'agree' or 'disagree' with a statement, respondent really agrees or disagrees to a certain extent (continuous scale), but is forced to choose one of the two responses
  - y_i^* can be viewed as the **propensity** to have the '1' response or the **utility difference** between alternatives '1' and '0'
    - ◇ e.g., the propensity (or inclination) to have a child vaccinated has to exceed some limit for the parent to have the child vaccinated
  - Death results when some continuous frailty exceeds a limit, or when exposure to some hazardous materials exceeds a limit

Latent response y_i^* (cont'd)

- Idea of latent response introduced by Pearson in 1901
- Yule remarked in 1912:

  > ...all those who have died of smallpox are equally dead: no one is more dead or less dead than another, and the dead are quite distinct from the survivors

- Pearson and Heron responded in 1913:

  > ...if Mr Yule's views are accepted, irreperable damage will be done to the growth of modern statistical theory

- Latent response models useful even if we do not believe in y_i^*

Latent response formulation

- Latent response model is a linear regression model

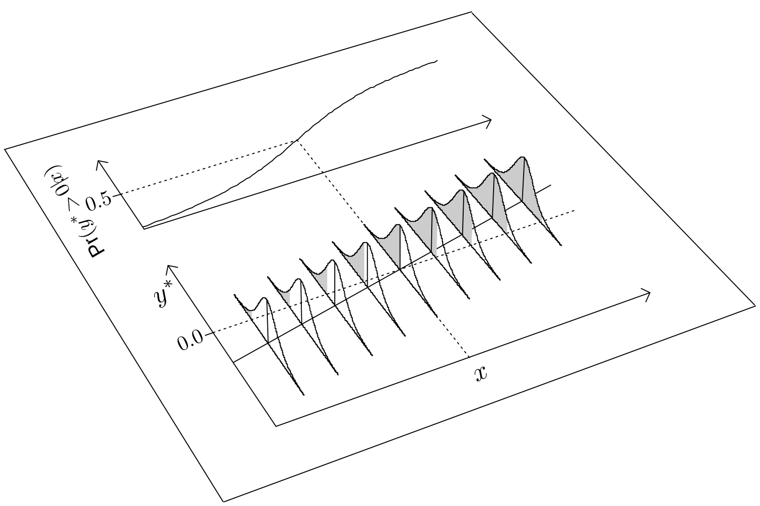
$$y_i^* \;=\; \beta_1 + \beta_2 x_i + \epsilon_i$$

- Observed response results as follows (deterministic):

$$y_i \;=\; \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Logistic regression model:
  - ϵ_i has a standard logistic distribution (variance $\pi^2/3$)
- Probit model:
  - ϵ_i has a standard normal distribution (variance 1)

Latent response formulation of logistic regression



Equivalence of generalized linear model and latent response formulation

- Can calculate the probability that $y_i = 1$ using latent resonse formulation:

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(y_i^* > 0|x_i) = \Pr(\beta_1+\beta_2 x_i+\epsilon_i > 0|x_i) \\ &= \Pr(-\epsilon_i \le \beta_1+\beta_2 x_i|x_i) \\ &= \Pr(\epsilon_i \le \beta_1+\beta_2 x_i|x_i), \quad \text{the CDF of } \epsilon_i \end{aligned}$$

  - Logistic CDF of ϵ_i results in logistic regression:

$$\Pr(y_i = 1|x_i) = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}$$

  - Standard normal CDF $\Phi(\cdot)$ of ϵ_i results in probit regression:

$$\Pr(y_i = 1|x_i) = \Phi(\beta_1+\beta_2 x_i)$$

Example: Toenail infection

- 337 patients with toenail infection randomized to receive terbinafine or itraconazole
- Assessments scheduled at 7 visits; weeks 0, 4, 8, 12, 24, 36, and 48
- Variables:
  - Onycholysis (separation of nail plate from nail bed) y_{ij} (0:none or mild, 1:moderate or severe)
  - Treatment group (0:itraconazole, 1:terbinafine) x_{2j}
  - Exact timing of visit in months x_{3ij}
  - Visit number (1,2,. . .,7)

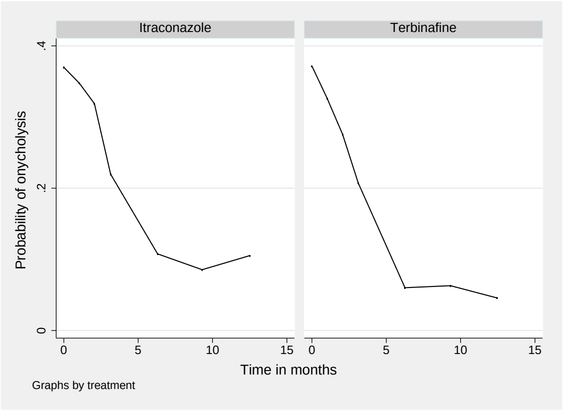
Exploring missingness patterns

Freq.	Percent	Cum.	Pattern
224	76.19	76.19	1111111
21	7.14	83.33	11111.1
10	3.40	86.73	1111.11
6	2.04	88.78	111....
5	1.70	90.48	1......
5	1.70	92.18	11111..
4	1.36	93.54	1111...
3	1.02	94.56	11.....
3	1.02	95.58	111.111
13	4.42	100.00	(other patterns)
294	100.00		XXXXXXX

- 224 patients have complete data, 21 patients missed visit 6, 10 patients missed visit 5, 6 patients dropped out after visit 3, etc.

Plot of raw estimates of marginal probabilities

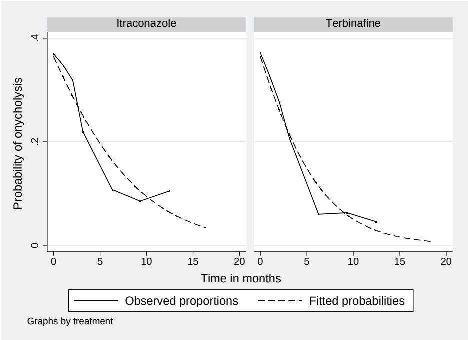
- Proportion with onycholysis at each occasion, versus average time at each visit since randomization



Logistic regression model for marginal probabilities

$$\text{logit}[\Pr(y_{ij}=1|x_{2j},x_{3ij})] \;=\; \beta_1+\beta_2 x_{2j}+\beta_3 x_{3ij}+\beta_4 x_{2j}x_{3ij}$$

- Regression coefficients and odds-ratios have **marginal** or **population averaged** interpretations, comparing prevalences for different population strata
- Plot of predicted probabilities together with raw estimates:



Random intercept logistic regression

- Ordinary logistic regression fits marginal proportions quite well
- However, unobserved heterogeneity between subjects and dependence within subjects are ignored
- Include a random intercept ζ_j:

$$\text{logit}[\Pr(y_{ij}=1|x_{2j},x_{3ij},\zeta_j)] \;=\; \beta_1+\beta_2x_{2j}+\beta_3x_{3ij}+\beta_4x_{2j}x_{3ij}+\zeta_j$$

or

$$y^*_{ij} \;=\; \beta_1+\beta_2x_{2j}+\beta_3x_{3ij}+\beta_4x_{2j}x_{3ij}+\zeta_j+\epsilon_{ij}$$

  - ζ_j enters in same manner as observed covariates
  - Assume $\zeta_j \sim \text{N}(0,\psi)$, independent of x_{2j}, x_{3ij}, and of ϵ_{ij} in latent response formulation (ϵ_{ij} has standard logistic distribution)
- Regression coefficients and odds-ratios have **conditional** or **cluster-specific** interpretations, comparing probabilities holding ζ_j constant

Estimation: Maximum likelihood

- Estimation for categorical responses difficult because marginal (or integrated) likelihood involves integrals that do not have closed form
- Numerical integration
  - Gauss-Hermite (ordinary) quadrature used in `MIXOR`/`MIXNO` (two-level only) and `aML`
  - Adaptive quadrature superior, particularly for large clusters and large variances. Available in SAS: `GLIMMIX` and Stata: `gllamm`, `xtmelogit`, etc., S-PLUS: `glme`, `Mplus`
- Monte Carlo integration
  - Simulated maximum likelihood in `nlogit`, Stata: `mixlogit`
  - Monte Carlo EM - no software?
- Markov chain Monte Carlo (MCMC) with vague priors approximates maximum likelihood and available in `MLwiN` and `WinBUGS`

Estimation: Approximate methods

- Penalized Quasilikelihood (PQL)
  - Two versions: First and second order (PQL-1,PQL-2), the latter being better
    - ◇ PQL-1 in `MLwiN`, `HLM` and SAS: `GLIMMIX`
    - ◇ PQL-2 in `MLwiN`
    - ◇ Even PQL-2 produces biased estimates for small clusters and large level-2 variances
- Laplace: R: `lmer` and Stata: `xtmelogit`
- Sixth order Laplace in `HLM`
- H-likelihood in `Genstat`
- Methods do not provide a likelihood

Maximum likelihood estimates

	Marginal effects		Conditional effects	
	Ordinary logistic		Random intercept logistic	
Parameter	OR	(95% CI)	OR	(95% CI)
Fixed part				
$\exp(\beta_2)$ [`treatment`]	1.00	(0.74, 1.36)	0.85	(0.27, 2.65)
$\exp(\beta_3)$ [`month`]	0.84	(0.81, 0.88)	0.68	(0.62, 0.74)
$\exp(\beta_4)$ [`trt_month`]	0.93	(0.87, 1.01)	0.87	(0.76, 1.00)
Random part				
ψ			16.08	
Log-likelihood	−908.01		−625.39	

Intraclass correlation of latent responses

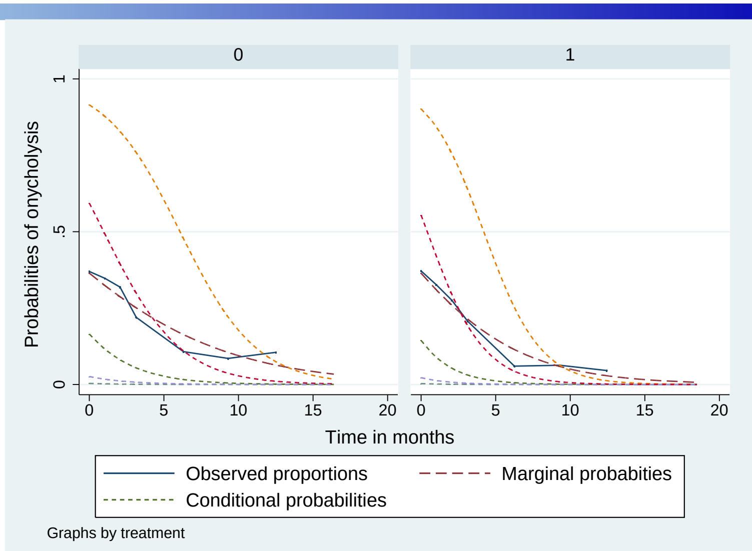
- Correlation between observed responses in the same cluster, given the covariates
$$\text{Cor}(y_{ij}, y_{i'j} | x_{2j}, x_{3ij}, x_{3i'j})$$
is a function of x_{2j}, x_{3ij}, and $x_{3i'j}$
- Therefore, report correlation between **latent responses** in same cluster, given covariates
$$\text{Cor}(y^*_{ij}, y^*_{i'j} | x_{2j}, x_{3ij}, x_{3i'j}) \;=\; \frac{\psi}{\psi + \pi^2/3}$$
  - Estimated intraclass correlation for toenail data:
$$\frac{\widehat{\psi}}{\widehat{\psi} + \pi^2/3} = 0.83$$

Marginal and conditional relationships

- Note that marginal OR closer to 1 than conditional OR
- Marginal probabilities from random intercept model

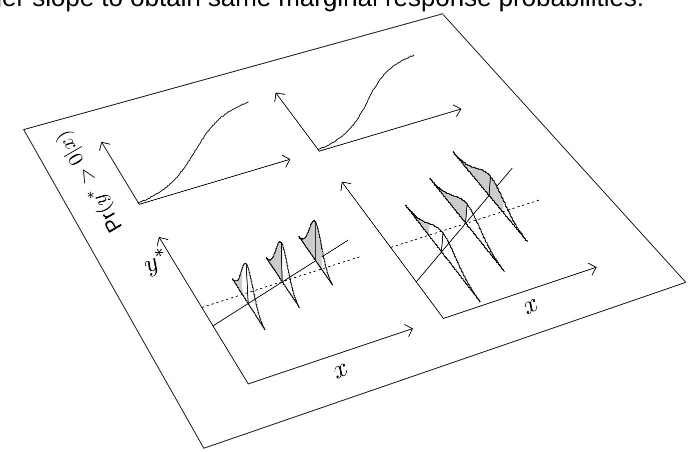
$$\begin{aligned}\text{Pr}(y_{ij} &= 1 | x_{2j}, x_{3ij}) \\ &= \int \text{Pr}(y_{ij} = 1 | x_{2j}, x_{3ij}, \zeta_j) g(\zeta_j; 0, \widehat{\psi})\, \mathrm{d}\zeta_j \\ &= \int \frac{\exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{2j} + \widehat{\beta}_3 x_{3ij} + \widehat{\beta}_4 x_{2j} x_{3ij} + \zeta_j)}{1 + \exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{2j} + \widehat{\beta}_3 x_{3ij} + \widehat{\beta}_4 x_{2j} x_{3ij} + \zeta_j)}\, g(\zeta_j; 0, \widehat{\psi})\, \mathrm{d}\zeta_j\end{aligned}$$

Marginal and conditional relationships (cont'd)



Reason for difference between conditional and marginal effects: Using latent response formulation

- Larger residual standard deviation, $\text{Var}(\zeta_j + \epsilon_{ij}) > \text{Var}(\epsilon_{ij})$, requires larger slope to obtain same marginal response probabilities:



Conditional and marginal effects for probit random intercept model

- Probit random intercept model: $y_{ij}^* = \beta_1 + \beta_2 x_{ij} + \underbrace{\zeta_j + \epsilon_{ij}}_{\xi_{ij}}$

  $\zeta_j \sim \mathrm{N}(0,\psi)$, $\epsilon_{ij} \sim \mathrm{N}(0,1) \quad \Rightarrow \quad \xi_{ij} = \zeta_j + \epsilon_{ij} \sim \mathrm{N}(0, \psi+1)$
- Conditional probability

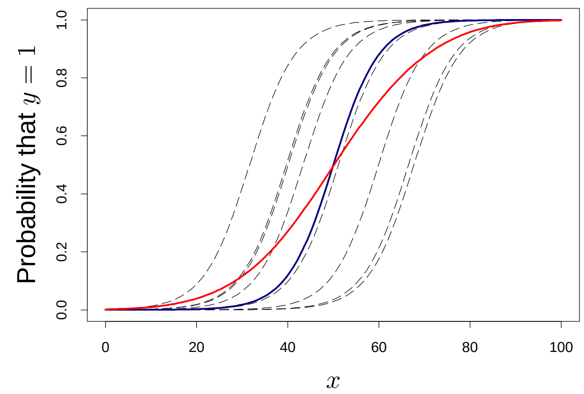
$$\Pr(y_{ij}=1|x_{ij},\zeta_j) = \Phi(\beta_1 + \beta_2 x_{ij} + \zeta_j)$$

- Marginal probability

$$\begin{aligned}
\Pr(y_{ij}=1|x_{ij}) &= \Pr(y_{ij}^*>0|x_{ij}) = \Pr(\beta_1+\beta_2 x_{ij}+\xi_{ij}>0|x_{ij}) \\
&= \Pr(-\xi_{ij} \le \beta_1+\beta_2 x_{ij}|x_{ij}) = \Pr(\xi_{ij} \le \beta_1+\beta_2 x_{ij}|x_{ij}) \\
&= \Pr(\frac{\xi_{ij}}{\sqrt{\psi+1}} \le \frac{\beta_1+\beta_2 x_{ij}}{\sqrt{\psi+1}}|x_{ij}) \\
&= \Phi(\frac{\beta_1+\beta_2 x_{ij}}{\sqrt{\psi+1}})
\end{aligned}$$

- Marginal effect **attenuated** or closer to zero: $|\beta_2/\sqrt{\psi+1}| \le |\beta_2|$

Illustration: Conditional versus marginal relationship



- - - - - cluster-specific (random sample)
——— median
——— marginal or population-averaged

Interpretation of regression parameter for within-cluster covariate

- **Conditional effects** or **subject-specific effects**:
  - Subject-specific odds ratios, e.g. for [month] $a+1$ versus a when [treatment] = 0

$$\exp(\beta_3^{\mathrm{C}}) = \frac{\Pr(y_{ij}=1|x_{ij}=a+1,x_j=0,\zeta_j)}{\Pr(y_{ij}=0|x_{ij}=a+1,x_j=0,\zeta_j)} \Big/ \frac{\Pr(y_{ij}=1|x_{ij}=a,x_j=0,\zeta_j)}{\Pr(y_{ij}=0|x_{ij}=a,x_j=0,\zeta_j)}$$

    - ◇ Comparing odds for particular subject j (conditional on ζ_j)
- **Marginal effects** or **population-averaged effects**:
  - Marginal odds ratios

$$\exp(\beta_3^{\mathrm{M}}) = \frac{\Pr(y_{ij}=1|x_{ij}=a+1,x_j=0)}{\Pr(y_{ij}=0|x_{ij}=a+1,x_j=0)} \Big/ \frac{\Pr(y_{ij}=1|x_{ij}=a,x_j=0)}{\Pr(y_{ij}=0|x_{ij}=a,x_j=0)}$$

    - ◇ Comparing odds for population strata (not conditional on ζ_j)

Interpretation of regression parameter for between-cluster covariate

- **Conditional effects** or **subject-specific effects**:
  - Subject-specific odds ratios, e.g. for [treatment] 1 versus 0 when [month] = 1

$$\exp(\beta_2^{\mathrm{C}}+\beta_4^{\mathrm{C}}) = \frac{\Pr(y_{ij}=1|x_j=1,x_{ij}=1,\zeta_j)}{\Pr(y_{ij}=0|x_j=1,x_{ij}=1,\zeta_j)} \Big/ \frac{\Pr(y_{ij}=1|x_j=0,x_{ij}=1,\zeta_j)}{\Pr(y_{ij}=0|x_j=0,x_{ij}=1,\zeta_j)}$$

    - ◇ Comparing **counterfactual** odds for particular subject j
- **Marginal effects** or **population-averaged effects**:
  - Marginal odds ratios

$$\exp(\beta_2^{\mathrm{M}}+\beta_4^{\mathrm{M}}) = \frac{\Pr(y_{ij}=1|x_j=1,x_{ij}=1)}{\Pr(y_{ij}=0|x_j=1,x_{ij}=1)} \Big/ \frac{\Pr(y_{ij}=1|x_j=0,x_{ij}=1)}{\Pr(y_{ij}=0|x_j=0,x_{ij}=1)}$$

    - ◇ Comparing odds for population strata

Pros and cons of conditional and marginal effects

- Marginal effects
  - Of interest for policy, e.g. public health
  - Not invariant across populations (depend on ψ)
- Conditional effects
  - Of interest for individuals, e.g. patients
  - More useful for investigating causal processes
  - More invariant across populations

Two-stage formulation

- Raudenbush and Bryk -style notation for two-level logistic models
  - Level-1 model

$$\begin{aligned}\varphi_{ij} &\equiv \text{Pr}(y_{ij}=1|\nu_{ij}) = \text{E}(y_{ij}|\nu_{ij})\\ y_{ij}|\varphi_{ij} &\sim \text{Binomial}(1,\varphi_{ij}) \equiv \text{Bernoulli}(\varphi_{ij}) \quad \text{('sampling model')}\\ \text{logit}(\varphi_{ij}) &= \beta_{0j}+\beta_{1j}x_{1ij}+\beta_{2j}x_{2ij} \equiv \nu_{ij} \quad \text{('structural model')}\end{aligned}$$

  - Level-2 models

$$\begin{aligned}\beta_{0j} &= \gamma_{00}+\gamma_{01}w_{1j}+\gamma_{02}w_{2j}+u_{0j}\\ \beta_{1j} &= \gamma_{10}+\gamma_{11}w_{1j}+\gamma_{12}w_{2j}+u_{1j}\\ \beta_{2j} &= \gamma_{20}\end{aligned}$$

where $(u_{0j},u_{1j})' \sim \text{N}(\mathbf{0},\boldsymbol{\tau}), \quad \boldsymbol{\tau} = \begin{bmatrix}\tau_{00} & \tau_{01}\\ \tau_{10} & \tau_{11}\end{bmatrix}$

IV. Longitudinal data and alternatives to multilevel modeling

- Longitudinal data
- Example: Wage and experience
- Linear growth curve models
- Nonlinear growth
- Example: Children's growth
- Fixed effects approach
- Marginal versus multilevel approach
- Autoregressive approaches
- Dropout and missing data
- Three-level models
- Example: Sustaining effects study

Longitudinal studies

- Panel surveys
  - All subjects followed up at the same panel waves $\Longrightarrow$ balanced data
- Cohort studies (as defined in epidemiology)
  - Cohort is any group of individuals, often same age ("birth cohort")
  - Generally, not followed up at the same time $\Longrightarrow$ unbalanced data
  - Intervention studies and clinical trials are special cases
- Other related types of studies (not discussed here)
  - Time-series for a single unit over time
  - Longitudinal information collected retrospectively $\Longrightarrow$ Recall bias
  - Survival, durations, or time-to event data

Longitudinal data

- Variables for subject j at occasion (e.g., panel wave) i
  - Response variable (time-varying) y_{ij}
  - Explanatory variable
    - ◇ Subject-specific (time-constant) x_j, e.g. gender
    - ◇ Occasion-specific x_i, e.g. calendar time
    - ◇ Subject and occasion-specific (time-varying) x_{ij}, e.g. marital status
- Longitudinal data are balanced if occasions for each subject correspond to same time points
  - Can treat responses at different occasions as different variables & use multivariate methods (e.g., Structural equation modeling)
  - Can model means and covariances more freely
- Intermittent missing data and dropout or attrition are common

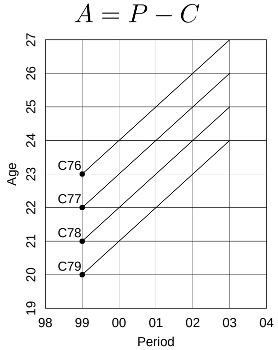
Longitudinal data (cont'd)

Subject j	Occ. i	Born x_{1j}	Age x_{2ij}	Year x_{3i}	Gender x_{4j}	Married x_{5ij}	Happiness y_{ij}
1	1	1960	35	1995	Male	1	1
1	2	1960	40	2000	Male	0	6
1	3	1960	45	2005	Male	1	3
2	1	1980	15	1995	Female	0	5
2	2	1980	20	2000	Female	0	4
2	3	1980	25	2005	Female	1	1
		Cohort	Age	Period			

Three time scales

- Age A: Time since birth
- Period P: Current calendar time (time since birth of Christ)
- Cohort C: Calendar time at time of birth

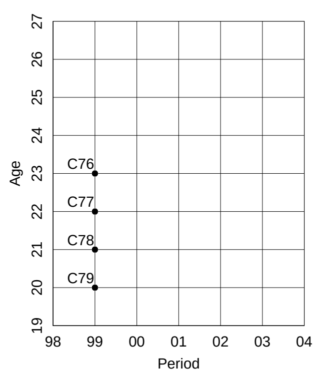
$$A = P - C$$



- Alternative age-like timescale: Time since subject-specific event such as surgery (then cohort becomes time of surgery)

Age-Period-Cohort effects: Cross-sectional study

- One period P
  $\Longrightarrow$ cannot estimate effect of period
- Different ages A_j, $\quad A_j = P - C_j$
  $\Longrightarrow$ age and cohort effects confounded



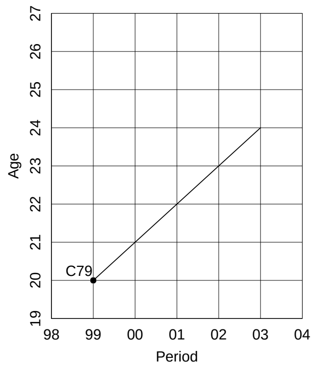
e.g., explanations for older people being more conservative:

(1) later stage in life A_j

(2) born longer ago (into a different 'era') C_j

Age-Period-Cohort effects: Longitudinal study, one cohort

- One cohort C
  $\Longrightarrow$ cannot estimate effect of cohort
- Different periods P_i and ages A_i, $\quad A_i = P_i - C$
  $\Longrightarrow$ period and age effects confounded



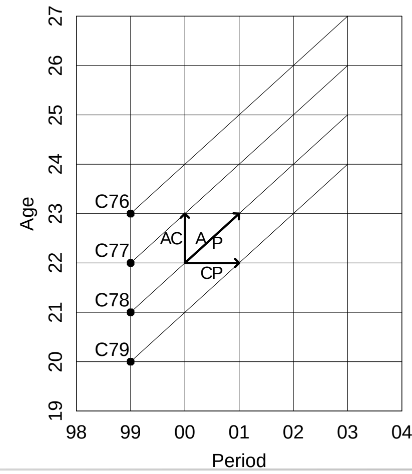
e.g., explanations for salary increases:

(1) more experience A_i

(2) inflation P_i

Age-Period-Cohort effects: Longitudinal study, several cohorts

- Several cohorts C_j, different periods P_i and ages A_{ij}, $\quad A_{ij} = P_i - C_j$
  $\Longrightarrow$ can estimate effects of two time scales, but confounded with third



Pick time scales believed to be most important

$\Rightarrow$ e.g., Conservatism depends on age and cohort (ignore period)

$\Rightarrow$ e.g., Salary depends on age and period (ignore cohort)

Other terms for design:
Accelerated longitudinal
Cohort-sequential

Example: Wage and experience

- US National Longitudinal Survey of Youth 1979 (NLSY79)
  - Representative sample of non-institutionalized, civilian U.S. youth
  - 6,111 men and women, aged 14-21 in Dec 31, 1978
  - Subsample of 545 considered here:
    - ◇ Full-time working males who completed schooling by 1980
    - ◇ Complete data for 1980-1987
  - Variables:
    - ◇ Subject identifier j
    - ◇ Log hourly wage lny_{ij}
    - ◇ Education (number of years) E_j
    - ◇ Labor market experience (in years) L_{ij}
    - ◇ Period (1980-1987) P_i
- How does log hourly wage depend on labor market experience L_{ij} and period P_i, controlling for education E_j?

Time scales in NLSY79

- Note that there are at least 5 time-scales:
$$A_{ij} \;=\; 6 + E_j + L_{ij} \;=\; P_i - C_j$$
  - A_{ij} determined by (and thus confounded with) E_j and L_{ij}
  - C_j determined by (and thus confounded with) P_i, L_{ij} and E_j
- Random intercept model: ln$y_{ij} = \beta_1 + \beta_2 L_{ij} + \beta_3 P_i + \beta_4 E_j + \zeta_j + \epsilon_{ij}$

	Est	(SE)	exp(Est)
Fixed Part:			
β_1	−52.99	(23.23)	
β_2 [L_{ij}]	0.04	(0.01)	1.04
β_3 [P_i]	0.03	(0.01)	1.03
β_4 [E_j]	0.10	(0.01)	1.11
Random Part:			
$\sqrt{\psi}$	0.34		
$\sqrt{\theta}$	0.35		

Linear growth curve models

- Appropriate for balanced or unbalanced data
- In R&B two-stage formulation, linear growth curve model (level 1):

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + r_{ij}$$

  - Each subject grows linearly, starting at level β_{0j} (when $t_{ij} = 0$) and growing at a rate of β_{1j} per unit of time (e.g., year)
  - Define level-2 models to explain variability in initial status β_{0j} and growth rate β_{1j} using subject-specific covariate x_j

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} x_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} x_j + u_{1j} \end{aligned}$$

- Reduced form formulation:

$$y_{ij} = \beta_1 + \beta_2 x_j + \beta_3 t_{ij} + \beta_4 x_j t_{ij} + \zeta_{1j} + \zeta_{2j} t_{ij} + \epsilon_{ij}$$

Nonlinear growth: Polynomial model

- Change fixed part of model to allow for nonlinear growth
- Saturated model for balanced data: use dummy variables for each time point except first
- Polynomial model

$$y_{ij} = \beta_{1j} + \beta_{2j} t_{ij} + \beta_{3j} t_{ij}^2 + \cdots + \beta_{p+1,j} t_{ij}^p + \cdots$$

  - Smooth function, but can get weird artifacts
  - Order of polynomial p determines flexibility – number of extrema is $p - 1$
  - Not all coefficients must be random
  - For balanced data, cannot have more coefficients than time points

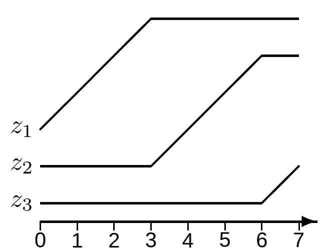
Nonlinear growth: Piecewise linear model

- Model, with linear spline basis functions z_{kij}

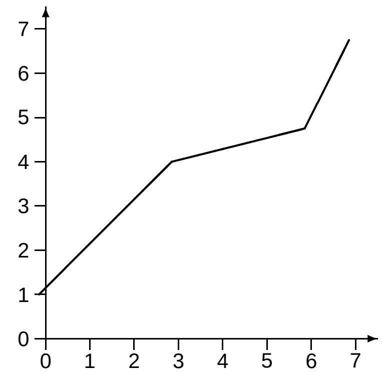
$$y_{ij} = \beta_1 + \beta_2 z_{1ij} + \cdots + \beta_{K+1} z_{Kij} + \cdots$$

- Example: $t_{ij} = i$, $i = 0, \ldots, 7$, and spline knots at $\tau_1 = 3$, $\tau_2 = 6$

t_{ij}	Interval	z_{1ij}	z_{2ij}	z_{3ij}
0	1	0	0	0
1	1	1	0	0
2	1	2	0	0
3	1	3	0	0
4	2	3	1	0
5	2	3	2	0
6	2	3	3	0
7	3	3	3	1



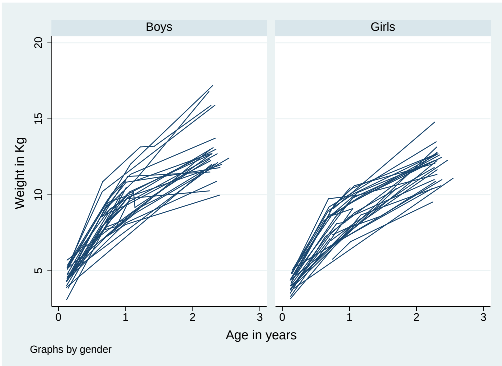
Nonlinear growth: Piecewise linear model (cont'd)



$1 + z_{1ij} + 0.25 z_{2ij} + 2 z_{3ij}$

Example: Children's growth

- Asian children in Britain weighed from age 6 weeks to 27 months:
  - Weight in Kg
  - Age in years
  - Gender (1:boy, 2:girl)
- Plot of observed trajectories



Maximum likelihood estimates (fixed part)

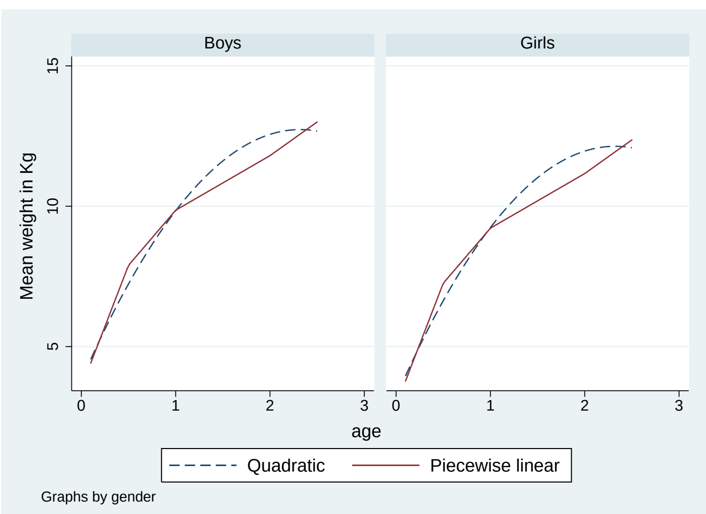
- Polynomial (quadratic)

	Est	(SE)
β_1	3.75	(0.17)
β_2 [girl]	−0.54	(0.21)
β_3 [age]	7.81	(0.25)
β_4 [agesq]	−1.66	(0.09)

- Piecewise linear (4 pieces), knots at 0.5, 1, 2

	Est	(SE)
β_1	3.34	(0.18)
β_2 [girl]	−0.64	(0.20)
β_3 [age1]	8.71	(0.45)
β_4 [age2]	3.93	(0.40)
β_5 [age3]	1.95	(0.70)
β_6 [age4]	2.40	(0.38)

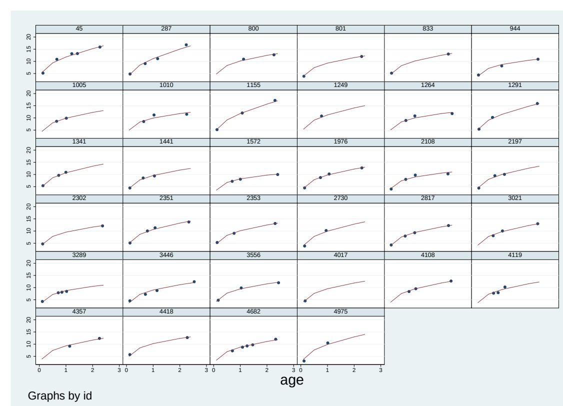
Estimated population averaged trajectories



Estimated subject-specific trajectories

- 'Trellis graph' of estimated cluster-specific trajectories (for boys)

$$\widehat{\mu}_{ij} = \widehat{\beta}_1 + \widehat{\beta}_2 \texttt{girl}_j + \widehat{\beta}_3 \texttt{age1}_{ij} + \widehat{\beta}_4 \texttt{age2}_{ij} + \widehat{\beta}_5 \texttt{age3}_{ij} + \widehat{\beta}_6 \texttt{age4}_{ij} + \widetilde{\zeta}_{1j} + \widetilde{\zeta}_{2j} \texttt{age}_{ij}$$



Fixed-effects models

- Avoid endogeneity or subject-level confounding by using fixed-effects models to estimate within-effects
  - Subjects truly act as their own controls
- For linear and log-linear models
  - Include dummy variables, or use conditional maximum likelihood (in linear case by sweeping out the subject mean)
- For logistic regression models:
  - Cannot include dummy variables for subjects due to incidental parameter problem, leading to inconsistent estimates of within-effects
  - Can use conditional logistic regression (conditional maximum likelihood, conditioning on sum of responses for subjects)

Disadvantages of fixed-effects models

- Cannot include subject-level covariates such as gender
- Inefficient if covariate(s) and/or response variable vary mostly between subjects
- Allows only for subject-specific intercepts (not slopes) for logistic regression
- Not possible for probit or ordinal models
- No direct information on unobserved heterogeneity
- Cannot make predictions for units in new clusters

Reminder: Marginal versus conditional

$$y_{ij} = \beta_1 + \beta_2 t_{ij} + \underbrace{\zeta_{1j} + \zeta_{2j} t_{ij} + \epsilon_{ij}}_{\xi_{ij}}$$

- Can consider conditional, or subject-specific expectation, given random effects ζ_{1j}, ζ_{2j}:

$$\text{E}(y_{ij}|t_{ij}, \zeta_{1j}, \zeta_{2j}) = \beta_1 + \beta_2 t_{ij} + \zeta_{1j} + \zeta_{2j} t_{ij}$$

- Conditional variance is θ and conditional covariances are zero
- Can consider marginal mean, variances and covariances

$$\text{E}(y_{ij}|t_{ij}) = \beta_1 + \beta_2 t_{ij}$$

$$\text{Var}(y_{ij}|t_{ij}) = \text{Var}(\xi_{ij}|t_{ij}) \;=\; \psi_{11} + 2\psi_{21} t_{ij} + \psi_{22} t_{ij}^2 + \theta$$

$$\text{Cov}(y_{ij}, y_{i'j}|t_{ij}, t_{i'j}) = \text{Var}(\xi_{ij}, \xi_{i'j}|t_{ij}, t_{i'j}) \;=\; \psi_{11} + \psi_{21}(t_{ij} + t_{i'j}) + \psi_{22} t_{ij} t_{i'j}$$

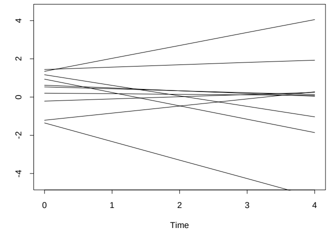
Marginal covariance matrix for linear growth curve model (5 occasions, $t = 0, 1, 2, 3, 4$)

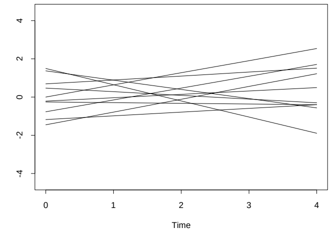
Random part

$$\boldsymbol{\Psi} = \begin{bmatrix} 1.00 & 0.10 \\ 0.10 & 0.25 \end{bmatrix} \qquad \theta = 0.5$$

$$\boldsymbol{\Psi} = \begin{bmatrix} 1.00 & -0.40 \\ -0.40 & 0.25 \end{bmatrix} \qquad \theta = 0.5$$

Subject-specific trajectories (β_1=β_2=0)





Residual variances

$$\begin{bmatrix} 1.50 & 1.95 & 2.90 & 4.35 & 6.30 \end{bmatrix}$$

$$\begin{bmatrix} 1.50 & 0.95 & 0.90 & 1.35 & 2.30 \end{bmatrix}$$

Residual correlations

$$\begin{bmatrix} 1.00 & & & & \\ 0.63 & 1.00 & & & \\ 0.58 & 0.76 & 1.00 & & \\ 0.51 & 0.74 & 0.84 & 1.00 & \\ 0.46 & 0.71 & 0.84 & 0.90 & 1.00 \end{bmatrix}$$

$$\begin{bmatrix} 1.00 & & & & \\ 0.50 & 1.00 & & & \\ 0.17 & 0.32 & 1.00 & & \\ -0.14 & 0.13 & 0.45 & 1.00 & \\ -0.32 & 0.00 & 0.42 & 0.68 & 1.00 \end{bmatrix}$$

Population-averaged or marginal approach to longitudinal data

- Instead of modeling individual trajectories (multilevel approach), model mean response and covariance matrix of (total) residual directly as functions of time ('Marginal model')
- Popular residual covariance structures
  - **Compound symmetric** or **exchangeable**: All variances equal and all covariances (and hence correlations) equal
    - If correlation is positive, random intercept model with variance $\psi + \theta$ and covariance ψ
  - **Autoregressive**: Correlations fall off as time lag increases
    - Popular special case: first order autoregressive, AR(1)

$$\text{Cor}(\xi_{ij}, \xi_{i'j}) = \alpha^{|t_i - t_{i'}|}$$

  - **Unstructured**: Each variance and covariance is freely estimated
    - Seems best, but inefficient (imprecise) if many time points

Illustration with three time-points

- Maximum likelihood estimates of residual variances and correlation matrices (alcohol use data, not discussed here)

Unstructured	AR(1)	Exchangeable	Growth curve model
$\begin{bmatrix} 0.52 & 0.77 & 1.11 \end{bmatrix}$	$\begin{bmatrix} 0.80 & 0.80 & 0.80 \end{bmatrix}$	$\begin{bmatrix} 0.80 & 0.80 & 0.80 \end{bmatrix}$	$\begin{bmatrix} 0.60 & 0.72 & 1.15 \end{bmatrix}$
$\begin{bmatrix} 1.00 & & \\ 0.44 & 1.00 & \\ 0.26 & 0.53 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & & \\ 0.49 & 1.00 & \\ 0.24 & 0.49 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & & \\ 0.40 & 1.00 & \\ 0.40 & 0.40 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & & \\ 0.38 & 1.00 & \\ 0.28 & 0.57 & 1.00 \end{bmatrix}$
-293.0 (6)	-299.3 (2)	-303.2 (2)	-294.3 (4)

Generalized estimating equations (GEE)

- Covariance structures for residuals are natural in linear models, giving multivariate regression models that can be estimated by maximum likelihood (ML)
- For binary and other non-continuous outcomes, can pretend that this is still possible
  - Specify structures for means and covariances $\Rightarrow$ Quasilikelihood
  - Use **estimating equations**, like "score equations" for ML
  - Estimation alternates between estimation of
    1. Regression coefficients: Generalized least squares for linearized model
    2. Covariance parameters: Moment estimators based on residuals
  - Not a true statistical model

Advantages of multilevel over marginal approach

- Multilevel model 'explains' covariance structure in terms of variability in intercepts and slopes
  - In marginal model, tempting to specify meaningless structures, such as constant variance over time in growth model (as in standard GEE)
- Multilevel model provides conditional or subject-specific interpretation $\Rightarrow$ stable across populations differing in between-subjects variability
- Multilevel model is proper statistical model for any response type
  - Can conceptualize as data-generating mechanism
  - Can simulate from the model
  - Can derive marginal relationships
  - Can make predictions and perform diagnostics based on predictions
  - Can perform likelihood ratio tests

Advantages of marginal over multilevel approach

- Permits more flexible covariance structures, e.g., negative intraclass correlation
- For non-continuous responses:
  - Marginal approach has marginal or population-averaged interpretation
    - Descriptive and easy to interpret; less likely to get extreme coefficients
  - Marginal approach via GEE gives consistent estimates of regression coefficients even if covariance structure misspecified (assuming correct fixed part)
  - GEE is computationally efficient (e.g., no numerical integration)

Models with autoregressive (AR) responses

- AR(1) model for response, conditioning on previous response $y_{i-1,j}$:

$$y_{ij} \;=\; \beta_1 + \gamma y_{i-1,j} + \beta_2 x_{ij} + \epsilon_{ij}, \qquad |\gamma| < 1$$

- Also called dynamic, lagged response or transition models
- Should be used only if effect γ of lagged response is of substantive interest ('state dependence' for binary responses)
- Advantage:
  - Easy to implement in linear as well as non-linear models
- Disadvantages:
  - Only sensible for equally spaced time-points
  - Discarding data: Lags missing for first occasion, missing responses and subsequent responses discarded
  - Initial conditions problem if true model contains subject-specific effects ζ_j

Models with autoregressive (AR) residuals

- AR(1) model for residual, conditioning on previous residual $\epsilon_{i-1,j}$

$$\epsilon_{ij} \;=\; \alpha\epsilon_{i-1,j} + \delta_{ij}, \qquad \delta_{ij} \sim \mathrm{N}(0,\sigma^2) \quad \mathrm{Cor}(\epsilon_{i-1,j}, \delta_{ij}) = 0$$

- Correlation structure is

$$\mathrm{Cor}(\epsilon_{ij}, \epsilon_{i'j}) \;=\; \alpha^{|t_{ij} - t_{i'j}|}, \quad |\alpha| < 1$$

Dropout and missing data

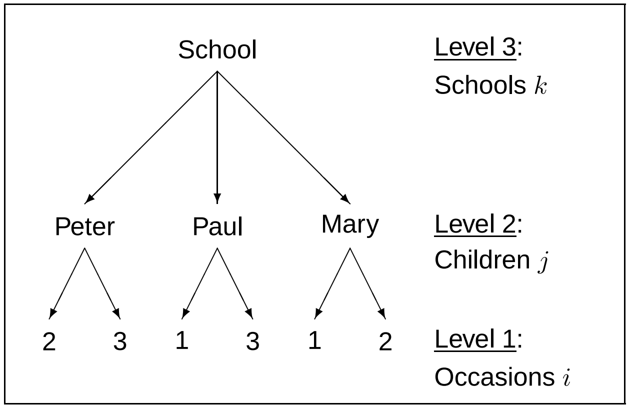
- Dropout or attrition is common where subjects are lost to follow-up from some time onwards (monotone missingness)
- Intermittent missing data also occur (e.g., subjects miss appointments but return)
- Old-fashioned methods & software (e.g., repeated measures ANOVA in SPSS) use listwise deletion, where all subjects with incomplete data are dropped
- Multilevel modeling & other modern methods (& modern software) use all available data
- Depending on reasons for dropout and missing data and on estimation method, both approaches can give inconsistent estimates

Types of missing data

- Missing completely at random (MCAR):
  $\Longrightarrow$ consistent estimates from ‘listwise’ data but inefficient
- Covariate-dependent dropout
  $\Longrightarrow$ consistent estimates if covariates that relate to missingness are included in model
- Missing at random (MAR):
  probability of missingness can depend on covariates and observed responses
  $\Longrightarrow$ consistent estimates if maximum likelihood used and model correctly specified
- Not missing at random (NMAR):
  probability of missingness depends on what that response would have been
  $\Longrightarrow$ Problems with all methods; can attempt to model missingness

Three-level data

- Units i nested in clusters j nested in superclusters k
  e.g. occasions i in children j in schools k

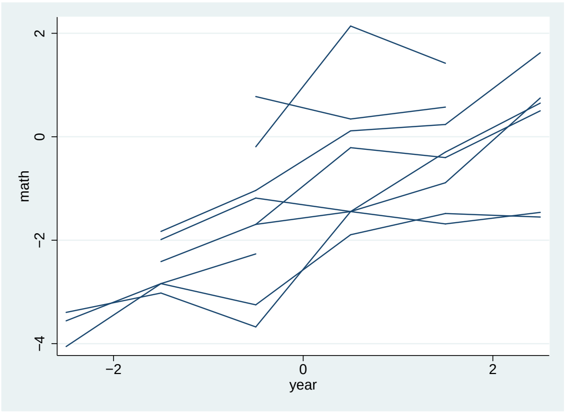


Example: Sustaining effects study

- Longitudinal survey of children in the six primary school years
- Primary sampling units were urban public primary schools
- 60 schools, 1721 students, 6 panel waves
- Variables:
  - Level 1 (occasion)
    - [Math]: Math test score from item response model y_{ijk}
    - [Year]: Year of study minus 3.5 a_{1ijk}
      (values -2.5, -1.5, -0.5, 0.5, 1.5, 2.5)
  - Level 2 (child)
    - [Black]: Dummy variable for being African American x_{1jk}
    - [Hispanic]: Dummy variable for being Hispanic x_{2jk}
  - Level 3 (school)
    - [Lowinc]: Percentage of students from low income families w_{1k}

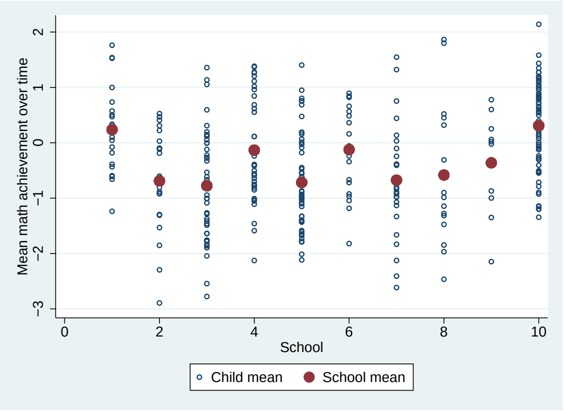
Variability between and within children

- Observed growth trajectories for 9 children from the same school



Variability between and within schools

- Mean math score over time for children from 10 schools



Three-level model

$$
\begin{aligned}
y_{ijk} &= \beta_1 + \beta_2 w_{1k} + \beta_3 x_{1jk} + \beta_4 x_{2jk} \\
&+ \beta_5 a_{1ijk} + \underbrace{\beta_6 w_{1k} a_{1ijk} + \beta_7 x_{1jk} a_{1ijk} + \beta_8 x_{2jk} a_{1ijk}}_{\text{Interactions}} \\
&+ \underbrace{\zeta_{1jk}^{(2)} + \zeta_{2jk}^{(2)} a_{1ijk}}_{\text{Child}} + \underbrace{\zeta_{1k}^{(3)} + \zeta_{2k}^{(3)} a_{1ijk}}_{\text{School}} + \underbrace{\epsilon_{ijk}}_{\text{Occ.}}
\end{aligned}
$$

Maximum likelihood estimates

Fixed part

		Est	(SE)
$\beta_1 \equiv \gamma_{000}$	[Cons]	0.141	(0.127)
$\beta_2 \equiv \gamma_{001}$	[Lowinc]	−0.008	(0.002)
$\beta_3 \equiv \beta_{01}$	[Black]	−0.502	(0.078)
$\beta_4 \equiv \beta_{02}$	[Hispanic]	−0.319	(0.086)
$\beta_5 \equiv \gamma_{100}$	[Year]	0.875	(0.039)
$\beta_6 \equiv \gamma_{101}$	[Lowinc]×[Year]	−0.001	(0.000)
$\beta_7 \equiv \beta_{11}$	[Black]×[Year]	−0.031	(0.022)
$\beta_8 \equiv \beta_{12}$	[Hispanic]×[Year]	0.043	(0.025)

Random part

	Est
$\sqrt{\psi_{11}^{(2)}}$	0.789
$\sqrt{\psi_{22}^{(2)}}$	0.105
$\rho_{21}^{(2)}$	0.561
$\sqrt{\psi_{11}^{(3)}}$	0.279
$\sqrt{\psi_{22}^{(3)}}$	0.089
$\rho_{21}^{(3)}$	0.033
$\sqrt{\theta}$	0.55

Interpretation of estimates

- In the middle of primary school, controlling for school mean income,
  - black and Hispanic children score on average 0.50 points and 0.32 points lower than white children, respectively
  - within ethnic groups, children's mean scores have a standard deviation of 0.79 within schools and 0.28 between schools; the standard deviation of scores around child-specific regression lines is 0.55
- On average, the mean math score increases 0.88 points per year for white children in schools with no low income children and this increase does not differ significantly for blacks or Hispanics
- The average annual increase in mean math scores is somewhat lower in schools with low income children, for a given ethnicity
- After controlling for ethnicity and school mean income, the average annual increase in math scores has a within-school standard deviation of 0.11 and a between-school standard deviation of 0.09

Model using three-stage (R&B) formulation

- Level-1 model:

$$y_{ijk} = \pi_{0jk} + \pi_{1jk} a_{1ijk} + e_{ijk}$$

  - Linear growth model
- Level-2 models:

$$\pi_{pjk} = \beta_{p0k} + \beta_{p1} x_{1jk} + \beta_{p2} x_{2jk} + r_{pjk}, \quad p = 0, 1$$

  - Mean intercept and slope depend on [Black] and [Hispanic]
  - Intercept and slope vary randomly between students within ethnic groups
- Level-3 models:

$$\beta_{p0k} = \gamma_{p00} + \gamma_{p01} w_{1k} + u_{p0k}, \quad p = 0, 1$$

  - Mean intercept and slope depend on [Lowinc]
  - Intercept and slope vary randomly between schools with given [Lowinc]

Deriving the reduced form

- Substitute level-3 models into level-2 models

$$\begin{aligned}\pi_{pjk} &= \underbrace{\gamma_{p00} + \gamma_{p01} w_{1k} + u_{p0k}}_{\beta_{p0k}} + \beta_{p1} x_{1jk} + \beta_{p2} x_{2jk} + r_{pjk} \\ &= \gamma_{p00} + \gamma_{p01} w_{1k} + u_{p0k} + \beta_{p1} x_{1jk} + \beta_{p2} x_{2jk} + r_{pjk}, \quad p = 0, 1\end{aligned}$$

- Substitute level-2 models into level-1 model

$$\begin{aligned}y_{ijk} &= \underbrace{\gamma_{000} + \gamma_{001} w_{1k} + u_{00k} + \beta_{01} x_{1jk} + \beta_{02} x_{2jk} + r_{0jk}}_{\pi_{0jk}} \\ &\quad + \underbrace{(\gamma_{100} + \gamma_{101} w_{1k} + u_{10k} + \beta_{11} X_{1jk} + \beta_{12} X_{2jk} + r_{1jk})}_{\pi_{1jk}} a_{1ijk} + e_{ijk} \\ &= \gamma_{000} + \gamma_{001} w_{1k} + \beta_{01} x_{1jk} + \beta_{02} x_{2jk} \\ &\quad + \gamma_{100} a_{1ijk} + \gamma_{101} w_{1k} a_{1ijk} + \beta_{11} x_{1jk} a_{1ijk} + \beta_{12} x_{2jk} a_{1ijk} \\ &\quad + r_{0jk} + r_{1jk} a_{1ijk} + u_{00k} + u_{10k} a_{1ijk} + e_{ijk}\end{aligned}$$

Further reading

-  Rabe-Hesketh & Skrondal (2012): Applied multilevel modeling (MLM) using Stata
- Snijders & Bosker (2011): Excellent introduction to MLM
- Fitzmaurice, Laird & Ware (2011): Most accessible biostatistical book on longitudinal data analysis (LDA)
- Wooldridge (2010): Most accessible econometric book on LDA
- Goldstein (2010): Generalized linear mixed models (GLMM)
- Raudenbush & Bryk (2002): GLMM
- McCulloch, Searle & Neuhaus (2008): Theoretical treatment of LMMs and GLMMs
-  Skrondal & Rabe-Hesketh (2004): GLMM & Generalized latent variable modeling