Analyses with Clustered Data

Outline

- ~~Sources of clustering: place, time, people, sampling~~
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Growth curves with mixed models
- GEE vs mixed
- Nonlinear outcomes

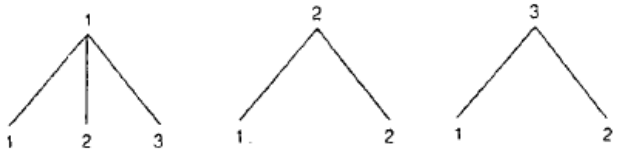
Outline

- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Growth curves with mixed models
- GEE vs mixed
- Nonlinear outcomes

Clusters can be a statistical nuisance or a scientific feature

- Statistical nuisance because the classic formulas to generate confidence intervals, p-values, or other measures of uncertainty assume all observations are independent.
- If the observations are correlated, generally you can extract less information than with independent observations.
- Clusters may also represent interesting substantive questions about the determinants of disease.
- Useful to think about this when choosing an analysis plan.

Questions potentially of interest: measures of people in a place



- Individual level exposure→ individual outcome
  - Do men have higher systolic blood pressure than women?
- Place level exposure → individual outcome
  - Do people who live in places with sufficient greenspace have higher systolic blood pressure than women?
- Cross-level interaction between individual and place level variables
  - Does greenspace have greater benefits for men than women?
- Variance partition/intraclass correlation
  - How much of the variance in SBP is between places vs between people within a place?

Questions potentially of interest: measures of people in a place

- Individual level exposure→ individual outcome
  - Do people with CKD have higher depressive symptoms than those without CKD?
- Cluster level exposure → individual outcome
  - Do patients w/ depression receiving care from psychiatrists have better symptom outcomes than patients receiving care from primary care physicians?
- Cross-level interaction between individual and place level variables
  - Does the effect of psychiatrist vs gp differ for male vs female patients with depression?
- Variance partition/intraclass correlation
  - How much of the variance in patient depressive symptoms is between physician providing care vs between patients receiving care from the same physician?

Why is variance partition of interest?

	Servings of fruit and vegetables combined (n = 13 281)	
	Coefficient	*P*
Race-ethnicity		
Mexican American	0.004	0.9783
Non-Hispanic black	−0.42	0.0003
Other	−0.30	0.1756
Non-Hispanic white (ref)	—	—
Age (centered)	0.03	< 0.0001
Age (centered), squared	−0.0001	0.3297
Female	−0.71	< 0.0001
US-born	−0.85	< 0.0001
Intercept	6.44	< 0.0001
Random effects[2]		
Level 1 variance	9.63	< 0.0001
Level 2 variance	1.99	< 0.0001
Level 3 variance	0.09	0.0309
ICCs[3]		
Level 2 (%)	17.03	
Level 3 (%)	0.74	

- NHANES data: people in census tracts in counties.
- Initial model estimated variance partition controlling for individual variables
- ICC for CT=17%
- 17%=1.99/(9.63+1.99+.09)
- ICC for county much lower

Dubowitz, et al Am J Clin nutr

Why is variance partition of interest?

	Servings of fruit and vegetables combined (n = 13 281)	
	Coefficient	P
Neighborhood SES	0.24	< 0.0001
Race-ethnicity		
Mexican American	0.11	0.4986
Non-Hispanic black	−0.24	0.0509
Other	−0.24	0.2870
Non-Hispanic white (ref)	—	—
Age (centered)	0.02	< 0.0001
Age (centered), squared	−0.0001	0.3277
Female	−0.71	< 0.0001
US-born	−0.83	< 0.0001
Intercept	6.37	< 0.0001
Random effects[2]		
Level 1 variance	9.62	< 0.0001
Level 2 variance	1.97	< 0.0001
Level 3 variance	0.07	0.0490
ICCs[3]		
Level 2 (%)	16.90	
Level 3 (%)	0.58	

- NHANES data: people in census tracts in counties.
- Next model added a neighborhood characteristic
- A 1 SD increase in neighborhood SES predicted 0.24 additional daily servings of F & V.
- Neighborhood SES did not explain the remaining variance in dietary intake across census tracts

Dubowitz, et al Am J Clin nutr

Intraclass correlation suggests causation of higher level variables.

- Clustering at a place or contextual level implies that there is something about the context that influences the outcome
- If all the diabetics at clinic A have Hb_{A1c}<6.5 and all the diabetics at clinic B have Hb_{A1c} >8, it makes you think that clinic A is doing something right.
- We'll go into the stats more, but conceptually this is a mixed model that partitions variance based on understanding of the sources of clustering
- Common to start with a "null" model: just describe percent of variance in the outcome clustered at each level
- Then a model accounting only for individual (level 1) characteristics, and estimate intraclass correlation coefficients
- Then add contextual variables to try to account for variance at level 2.

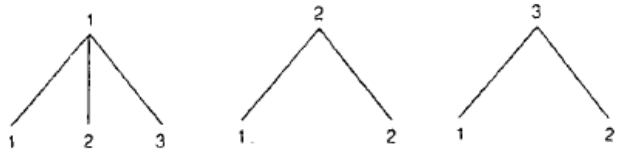
Intraclass correlation suggests causation of higher level variables.

- Multilevel/random effects models can be used to rule out compositional explanations for place level differences, but:
  - Cannot rule out selection differences (sick people go to good hospitals)
  - Controversy regarding what is “downstream” of contextual exposures.
- Draw the DAG!

We can switch from clustering on a place to clustering on a person.

Questions potentially of interest: repeated measures of a person

- Time specific exposure→ time specific outcome
  - Is SBP higher on days when CESD is higher?



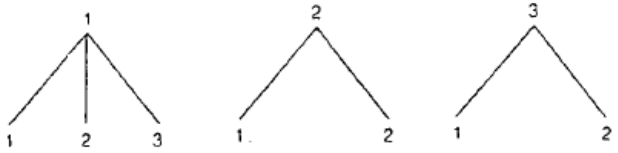
- Person level exposure → time-specific outcome
  - Do men have higher SBP than women?
- Cross-level interaction between individual and place level variables
  - Does CESD have a greater effect on SBP for men than women? have greater benefits for men than women?
- Variance partition/intraclass correlation
  - How much of the variance in MMSE is between person vs between successive observations on the same person?

Questions potentially of interest: repeated measures of a person

- Person level exposure → time-specific outcome
  - Very important category of questions represented by cross-level interactions
  - Does SBP change more quickly for men than for women: measure sex at person level and time at the occasion level.
  - Addresses whether sex of the person modifies the rate of change over time.
  - Major challenge is choosing the time dimension: age vs time from enrollment vs time from some other milestone (e.g., death, diagnosis)

Things we hardly ever ask in epidemiology

- Individual level exposure→ place level outcome



  - Does presence of men affect prevalence of greenspace in a neighborhood?
- Place level exposure → place level outcome
  - Do places with greenspace tend to have better air quality?
  - Beware **ecological fallacy**: drawing inferences about individual level variables based on ecological associations, e.g., places with high prevalence of bathtubs also have prevalence of heart diseasse
- Side note: distinguish compositional from structural characteristics of places
  - Compositional: average income of people in the place
  - Structural: presence of a park in the place

Outline

- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Growth curves with mixed models
- GEE vs mixed
- Nonlinear outcomes

Consequences of ignoring clustering

- Missed scientific questions
  - Individualistic fallacy: assuming that individual-level outcomes can be explained exclusively in terms of individual-level characteristics
- Incorrect standard errors
  - Most variance estimates assume independent observations
- Inefficient estimators
- Sometimes helps handle missing data/loss to follow-up

Incorrect SEs and Loss of efficiency

- Ignoring clustered data generally leads to SEs for individual level exposures that are too small
  - Variance of estimate of mean value of Y=variance of Y/N
  - But in clustered data effective N<actual N
- Same problem if you draw a clustered sample, people within a cluster are not independent and you do not learn as much new information from each person.
  - You must correct your standard errors for this phenomenon.
  - You must anticipate this correction in study design and inflate the sample you need.
  - The ratio of the clustered sample size needed to the simple random sample size is the design effect.
- How does clustering affect the sample size you must draw, compared to a simple random sample?
- Conversely, how much does *ignoring* clustering affect your SEs?

Variance "cost" of clustering a function of rho/ICC and # of obsvns/cluster

- Rho = a measure of homogeneity that behaves like a correlation coefficient. It is the correlation between all possible pairs of elements within a cluster on a particular characteristic (Sudman, 1976)
- ICC=rho= $\sigma^2_{\text{between-cluster}}/(\sigma^2_{\text{between-cluster}} + \sigma^2_{\text{within-cluster}})$
- Rho of 1 indicates extreme homogeneity (e.g., household income of all of 6 children in one household).
- Rho of -1 indicates extreme heterogeneity (e.g., gender w/in a 2 person heterosexual household cluster).
- Small Rho of 0.05 or less common for many characteristics in neighborhoods and schools
- If you cluster randomize, everyone in the cluster has the same value of the random assignment variable, so rho for that variable is 1.

Calculation of rho/ICC

	Servings of fruit and vegetables combined (n = 13 281)	
	Coefficient	*P*
Race-ethnicity		
Mexican American	0.004	0.9783
Non-Hispanic black	−0.42	0.0003
Other	−0.30	0.1756
Non-Hispanic white (ref)	—	—
Age (centered)	0.03	< 0.0001
Age (centered), squared	−0.0001	0.3297
Female	−0.71	< 0.0001
US-born	−0.85	< 0.0001
Intercept	6.44	< 0.0001
Random effects[2]		
Level 1 variance	9.63	< 0.0001
Level 2 variance	1.99	< 0.0001
Level 3 variance	0.09	0.0309
ICCs[3]		
Level 2 (%)	17.03	
Level 3 (%)	0.74	

- Denominator=total variance
- Numerator=variance of mean at the cluster level

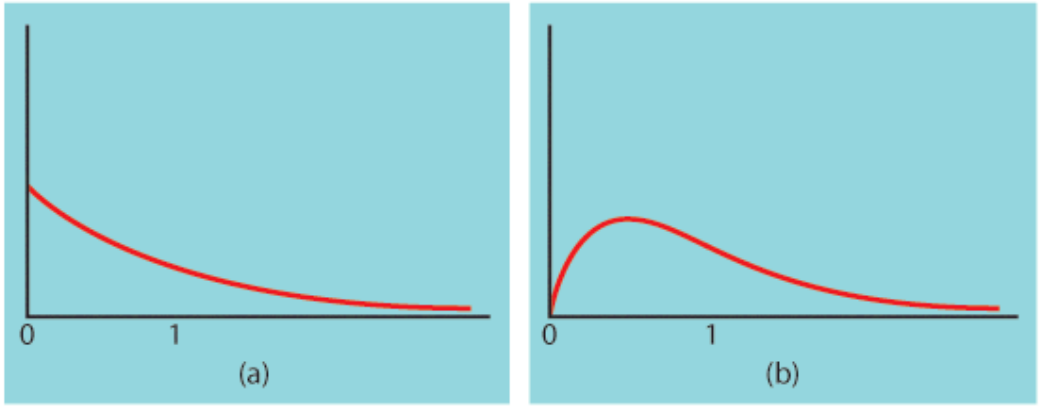
- Can ask about unconditional (didn't control for anything yet) or conditional ICC

Dubowitz, et al Am J Clin nutr

The central limit theorem

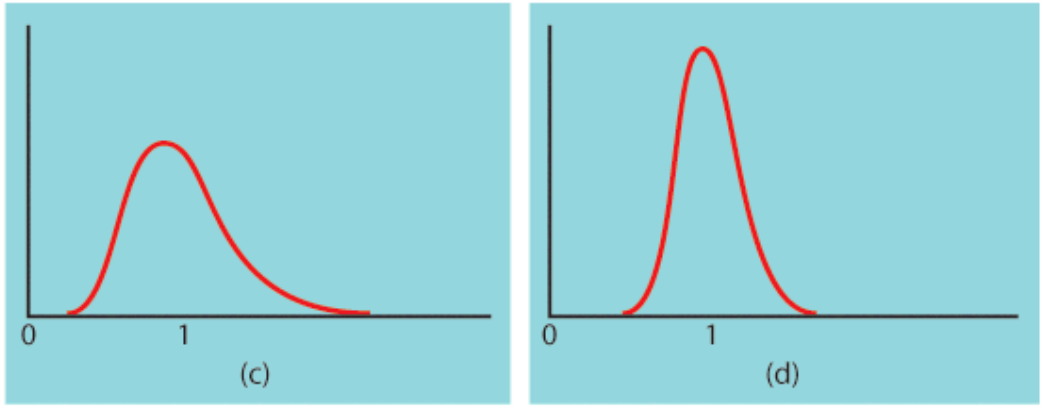
When randomly sampling from any population with mean μ and standard deviation σ, **when *n* is large enough**, the sampling distribution of *x* bar is approximately normal: $\sim N(\mu, \sigma/\sqrt{n})$.

Population with strongly skewed distribution



Sampling distribution of $\bar{x}$ for n = 2 observations

Sampling distribution of $\bar{x}$ for n = 10 observations



Sampling distribution of $\bar{x}$ for n = 25 observations

$\bar{x}$

Approximate design effect formula for estimation of the *mean*

$DEff = \sigma^2_{cluster}/\sigma^2_{SRS} = 1 + rho_X(m_{avg} - 1)$

DEff = the ratio of the variance from a clustered sample *design* to the variance of a *SRS* of the same size

m_{avg} = average number of units drawn from a cluster

σ^2 = V= sample variance of the parameter estimate

- Works if clusters are about the same size.
- Assumes a small proportion of all possible clusters have been sampled.

- Bigger DEff implies you need more data
- DEff increases if rho goes up.
- DEff increases if m goes up.
- What if m=1? (i.e., 1 person in a cluster)
- What if rho=1? (i.e., everyone in cluster identical)

Approximate design effect formula

DEff = the ratio of the variance from a clustered sample design to the variance of a *SRS* of the same size

So, if you estimate the population mean of some variable X, and X has variance σ^2 in the population with a SRS, the variance of your estimated mean in a SRS = σ^2 /n

But the variance of your estimated mean in a clustered sample=Deff* σ^2 /n

Keep in mind: these are approximate formulas.

The relationship between the sample size and the variance of the estimate depends on what you are estimating (e.g. the mean of the population or some other characteristics, perhaps the 25th percentile).

Approximate design effect formula

Keep in mind: these are approximate formulas for the design effect.

The relationship between the sample size and the variance of the estimate depends on what you are estimating (e.g. the mean of the population or some other characteristics, perhaps the 25^{th} percentile).

The design effect also depends on what you are estimating, mean of the population, linear regression coefficient, etc.

For bivariate linear regression, with equal size clusters with exchangeable within cluster correlation

DEff approximately= $1 + rho_X rho_Y (m_{avg} - 1)$

So if either rho_X or rho_Y is 0, DEff is zero.

Examples, DEFF for the mean of X

- DE=1+(m-1)*rho
- m=30, rho=.5
  - DE=1+29*.5=~16
- m=30, rho=.05
  - DE=1+29*.05=~1.15
- m=300, rho=.05
  - DE=1+299*.05=~16
- m=1, rho=.9
  - DE=1+0*.9=1
- m=500, rho=0
  - DE=1+499*0=1

Cluster Size

- As the number of people in the cluster increases, the rho tends to decrease (this is not a law of nature… just generally true)
- However, you do not need to enroll everyone in a cluster, even if you "treat" everyone in the cluster.
- If you randomize cities of 1,000,000 to treatment or non-treatment and then you interviewed all million, the deff would be large even if the ICC for the city was tiny (1+999,999*.01)=1+9,999.99=~10,001.

Cluster Size

- If your clusters are two cities of a million with rho=.01, your effective n would be $2x10^6/1x10^4=200$.
- But you probably won't interview everyone in each city. Perhaps you interview only 100 per city.
- The DE=1+(m-1)*rho=1+99*.01=2.
- Effective n=200/2=100
- Probably more efficient to interview only 100 instead of all million inhabitants and spend the money you saved trying to sample a third city.

Examples?

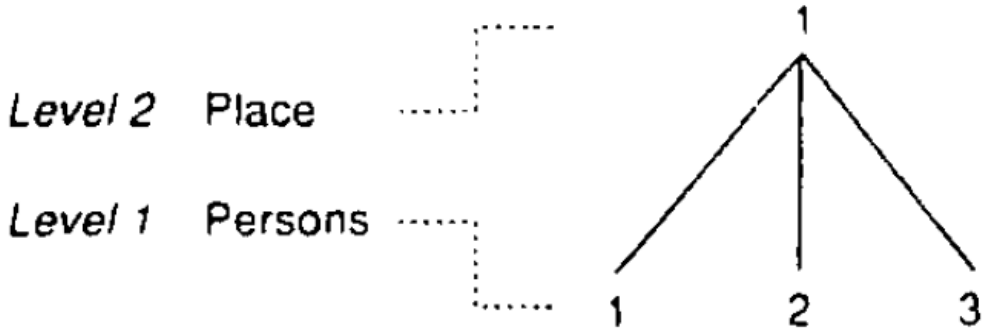
- Rho for ???
- Reflection examples

Outline

- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Growth curves with mixed models
- GEE vs mixed
- Nonlinear outcomes

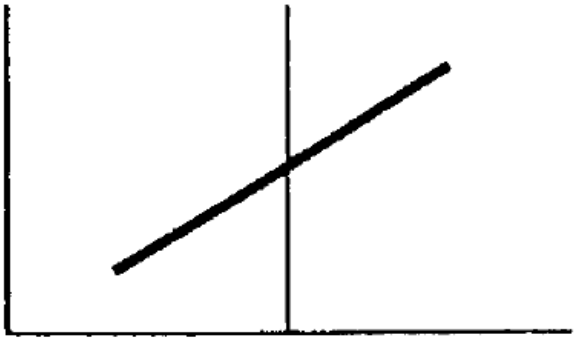
Clustering of people within places

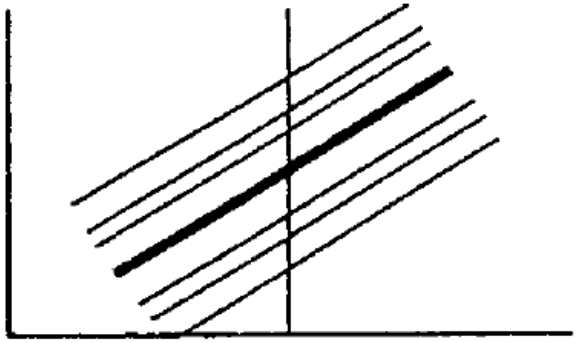
Two-level structure



Y_{ij}= Value of the variable "Y" for person *i* in place *j*
Could also talk about X_{ij} or $X_{j,}$ a feature of place *j* that has the same value for every person *i* in place *j*.

Basic mixed model for clustered data without ordering





We assume places differ by some little bit (which we will call the mu for that place, or μ_{0j}) and everyone in the same place has this same little bit of offset, plus their own personal differences.

When predicting each person's outcome value, we add this place-specific little bit to the intercept, along with differences predicted by any other factor.

Basic mixed model for clustered data without ordering

Two level random intercepts model:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \text{ [Micro model]}$$

$$\beta_{0j} = \beta_0 + \mu_{0j} \text{ [Macro model]}$$

$$\text{[Overall model]} = y_{ij} = \beta_0 + \beta_1 x_{ij} + (\mu_{0j} + \varepsilon_{ij})$$

To fully specify a mixed model, you must also describe the distribution of the random terms

$\mu_{0j} \sim N(0, \sigma^2_{\mu 0})$

$\varepsilon_{ij} \sim N(0, \sigma^2_e)$

The "intercepts" are just place level deviations

Two level random intercepts model:

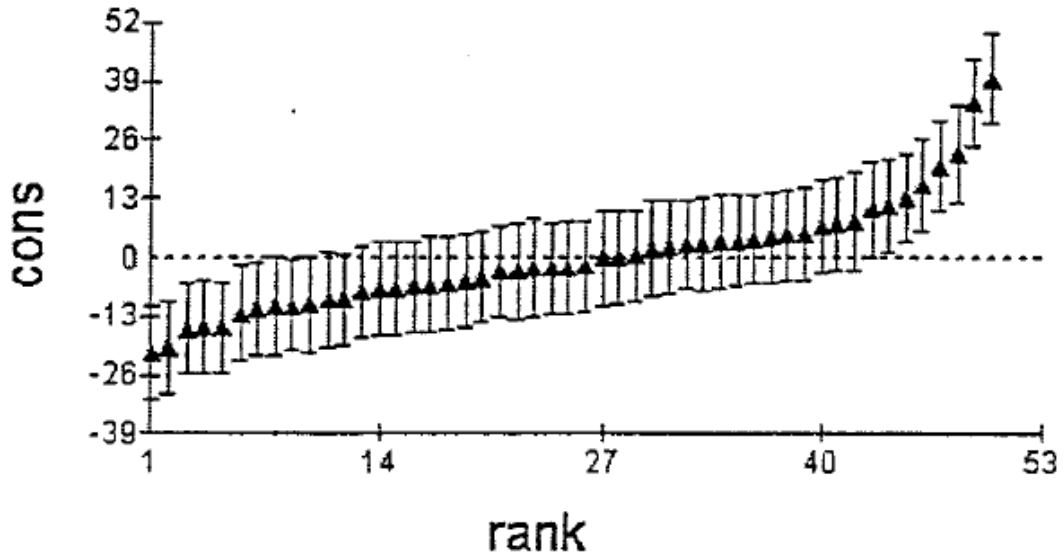
$$[\text{Overall model}] = y_{ij} = \beta_0 + \beta_1 x_{ij} + (\mu_{0j} + \varepsilon_{ij})$$

To fully specify a mixed model, you must also describe the distribution of the random terms

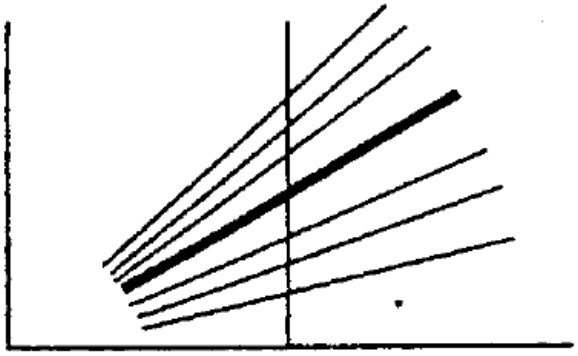
$\mu_{0j} \sim N(0, \sigma^2_{\mu 0})$

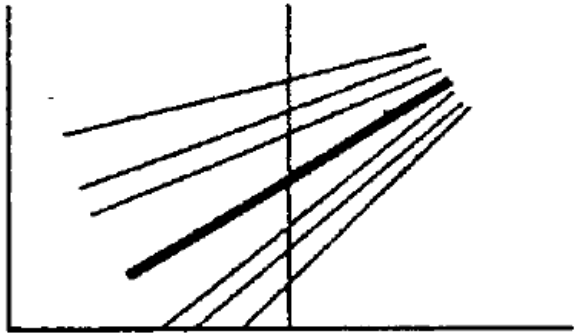
$\varepsilon_{ij} \sim N(0, \sigma^2_e)$

You can plot the μ_{0j}s to get a sense of the distribution



Random slopes model





Random slopes model (implicitly, assume you have a random intercept as well):

Not only do places differ by a little bit (the μ_{0j} random intercept) to the same extent for everyone in the place, but the effect of exposure differs depending on the place (μ_{1j}), so the “slope”, or coefficient is different.

Random slopes model

Two level random slopes model (implicitly, assume you have a random intercept as well):

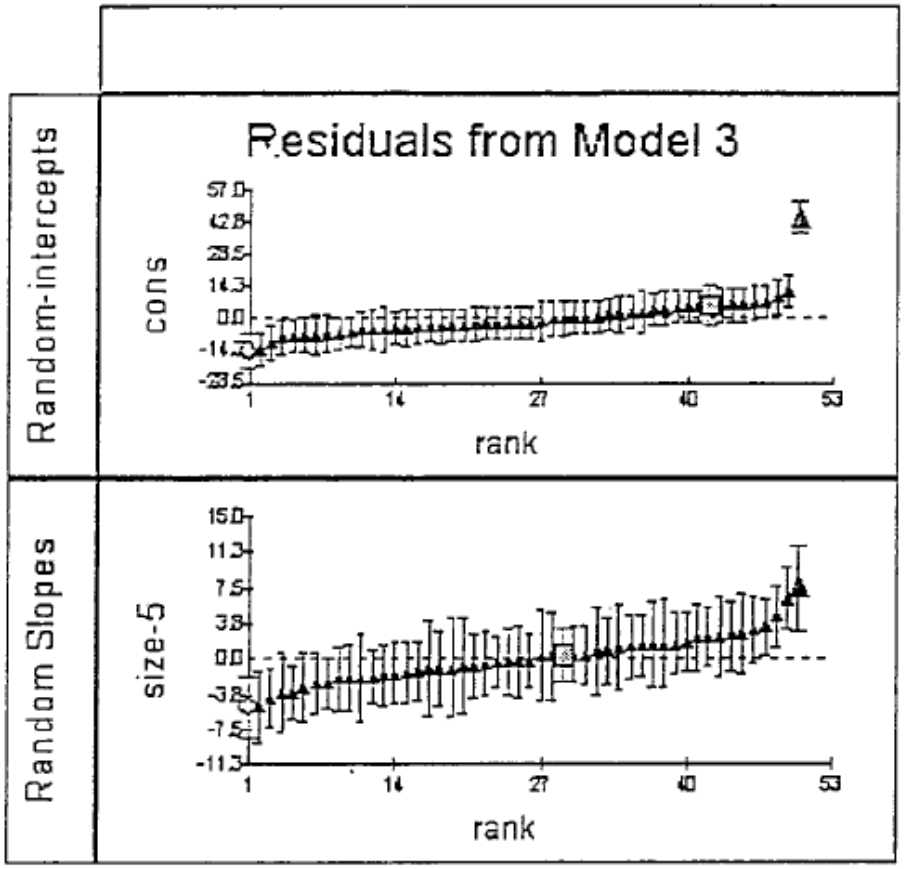
$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \varepsilon_{ij} \text{ [Micro model]}$$

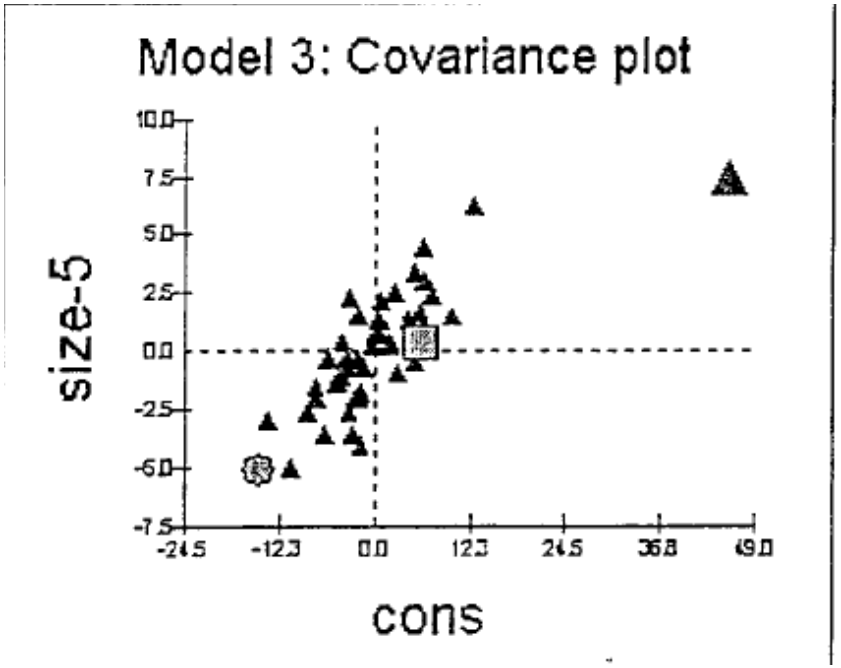
$$\beta_{0j} = \beta_0 + \mu_{0j} \text{ [Macro model for intercept]}$$

$$\beta_{1j} = \beta_1 + \mu_{1j} \text{ [Macro model for slope]}$$

$$y_{ij} = \beta_0 x_{0ij} + \beta_1 x_{1ij} + (\mu_{1j}x_{1ij} + \mu_{0j}x_{0ij} + \varepsilon_{ij}x_{0ij})$$

Potential covariance of random intercept and slope





For a random slope model, you must describe the variance and covariance of the random intercept and slope

Substantive questions from the intercept-slope covariance

For a random slope model, you must describe the variance and covariance of the random intercept and slope

In places that average poor health outcomes for whites, are racial disparities especially large?

In hospitals that have poor functional outcomes for mild stroke patients, does stroke severity have a smaller or larger impact on functional outcomes? In other words, do hospitals "specialize" in optimizing outcomes for severe strokes (negative covariance) or are hospitals that are bad for mild strokes terrible for severe strokes (positive covariance)

Outline

- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Growth curves with mixed models
- GEE vs mixed
- Nonlinear outcomes