Analyses with Clustered Data and Lifecourse Models

Outline

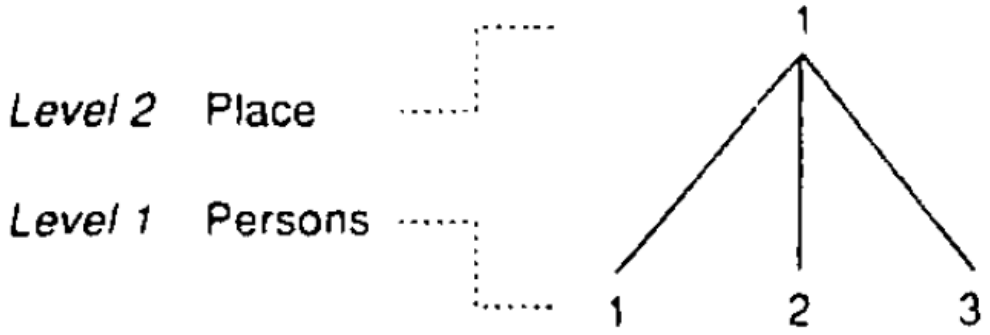
- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Growth curves with mixed models
- Lifecourse timing

Sources of clustering

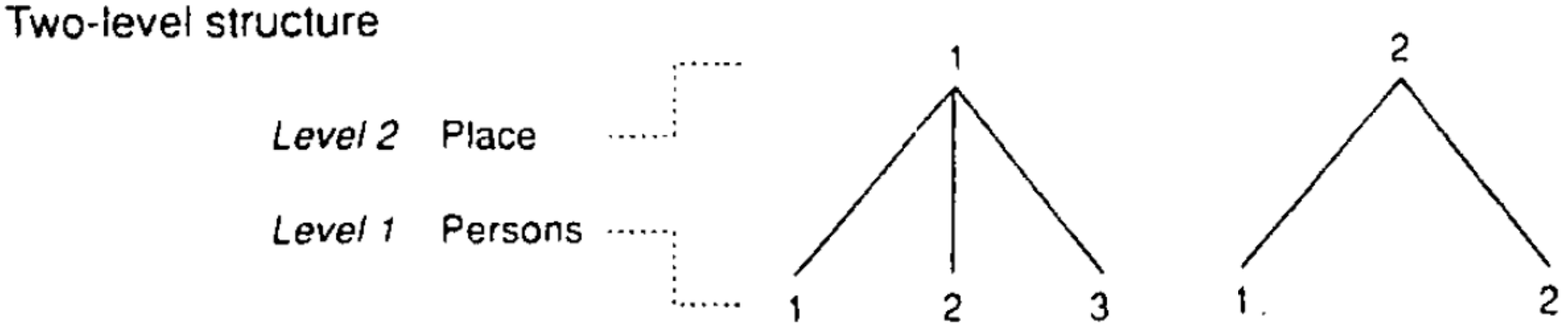
- Sampling places or contexts (neighborhoods, schools, hospitals, clinics, families, siblings)
- Repeatedly measuring the same person
- Clusters may have equal numbers of observations or unequal
- Items within a cluster may be ordered with respect to one another
- Assume that conditional on cluster, observations are independent

Represent the hierarchy of the data

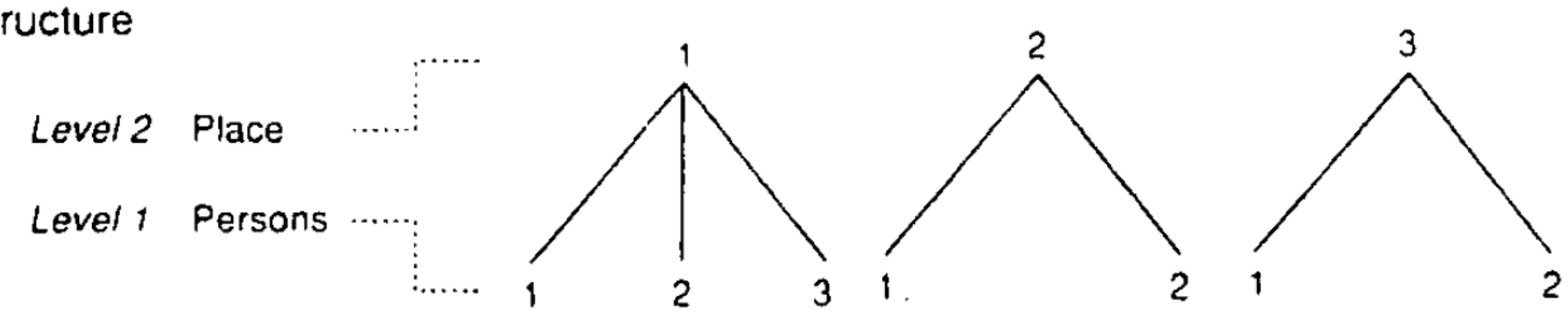
Two-level structure



Represent the hierarchy of the data



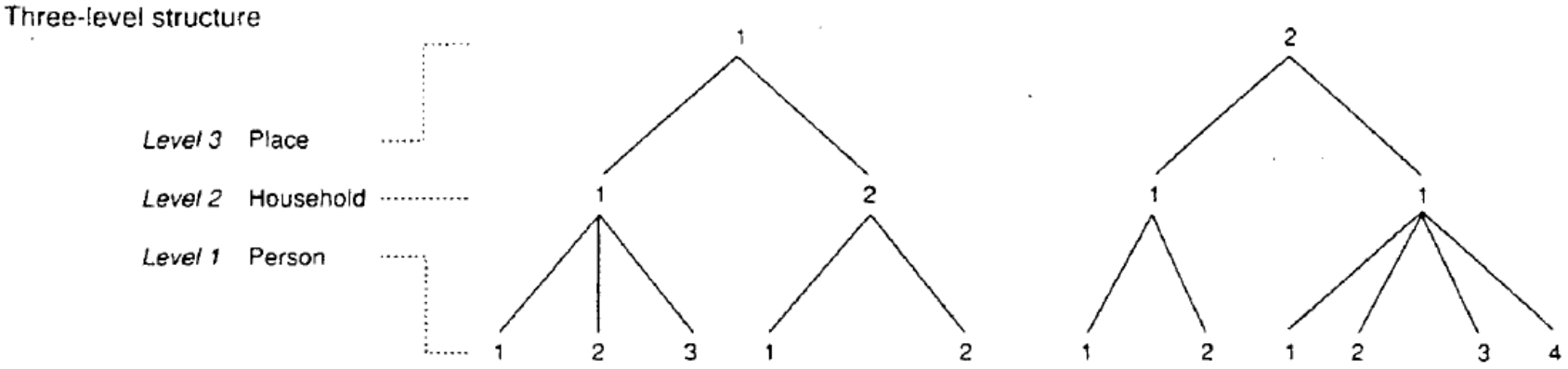
Represent the hierarchy of the data



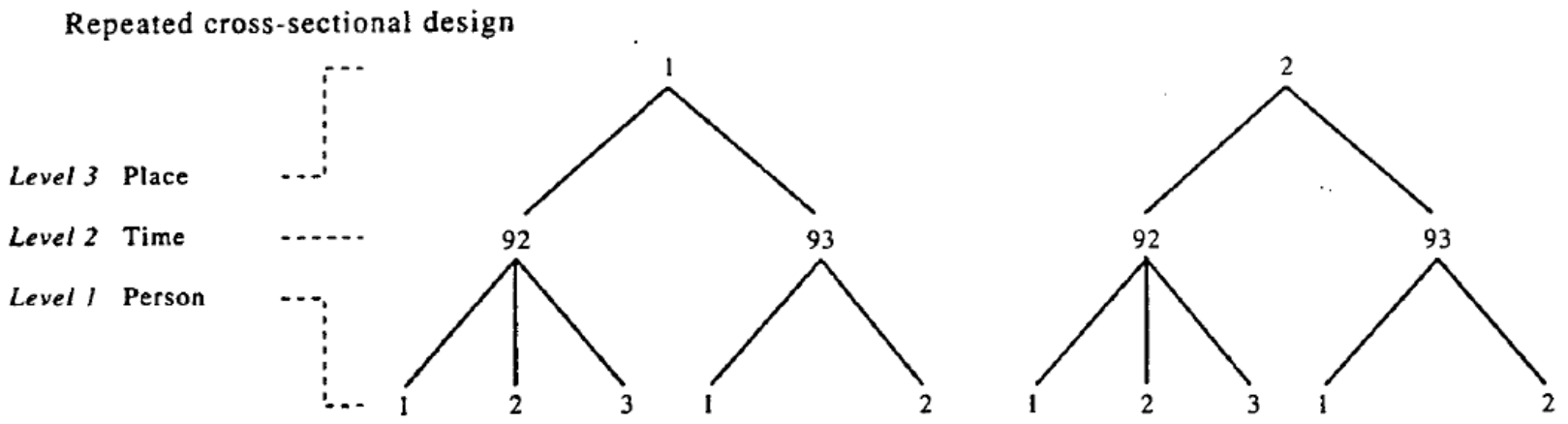
Sources of clustering

- Three level hierarchy:
  - Diabetic people nested within clinics
  - Each patient measured repeatedly
- Cross-classified hierarchy:
  - Diabetics within a clinic
  - Same diabetics live in neighborhoods
- Multivariate response
  - Two outcome measures of the same latent variable
  - Two measures, you think some of the effects of independent variables are the same.

Three level hierarchy

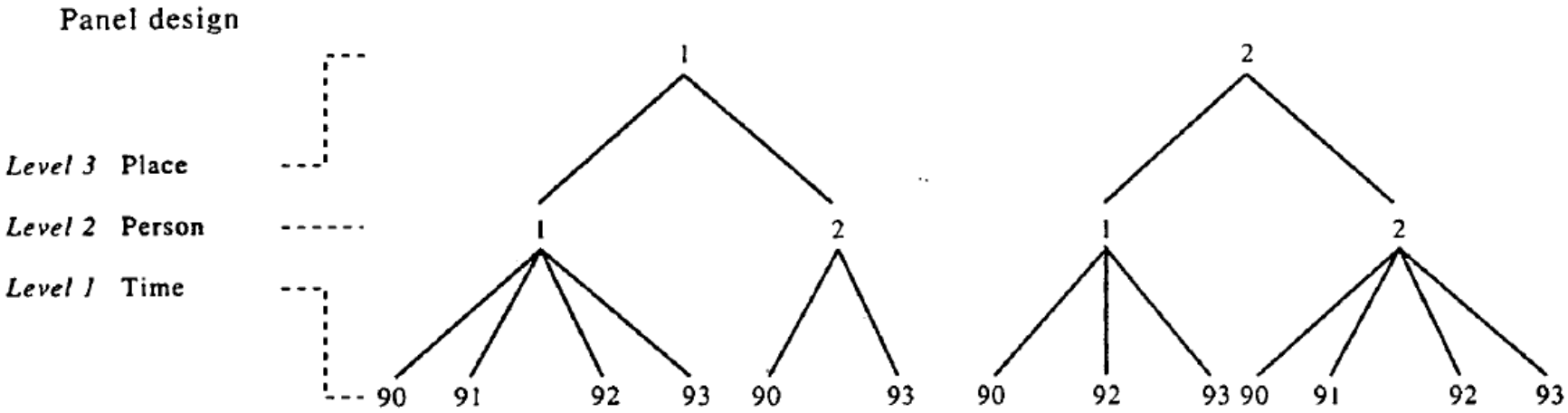


Hierarchy not always obvious: contrast repeated cross sectional design



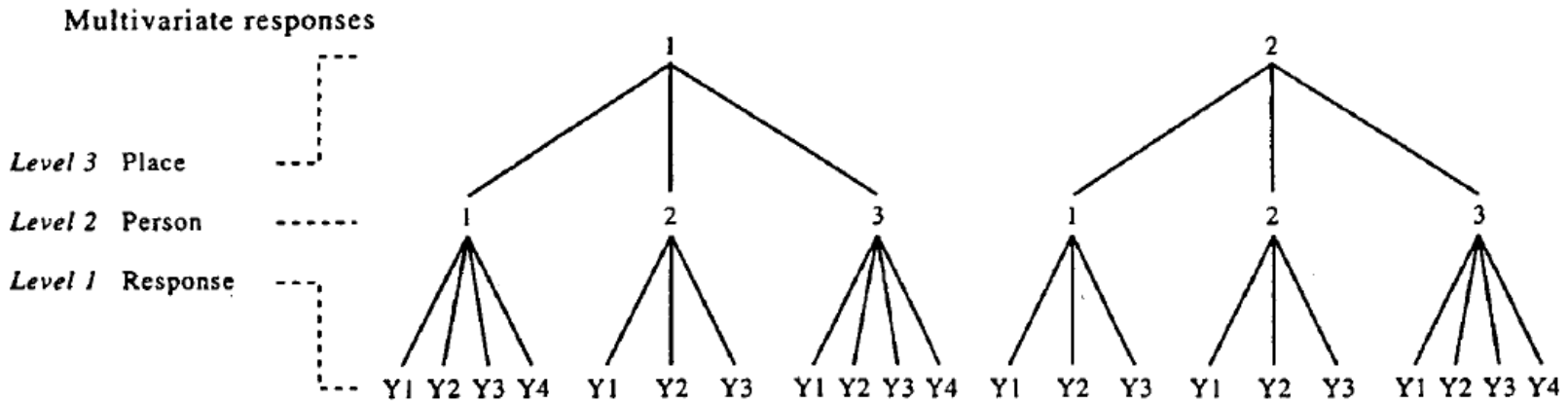
Time is level 2 because People are sampled within Waves in each Place

With a repeated measures design



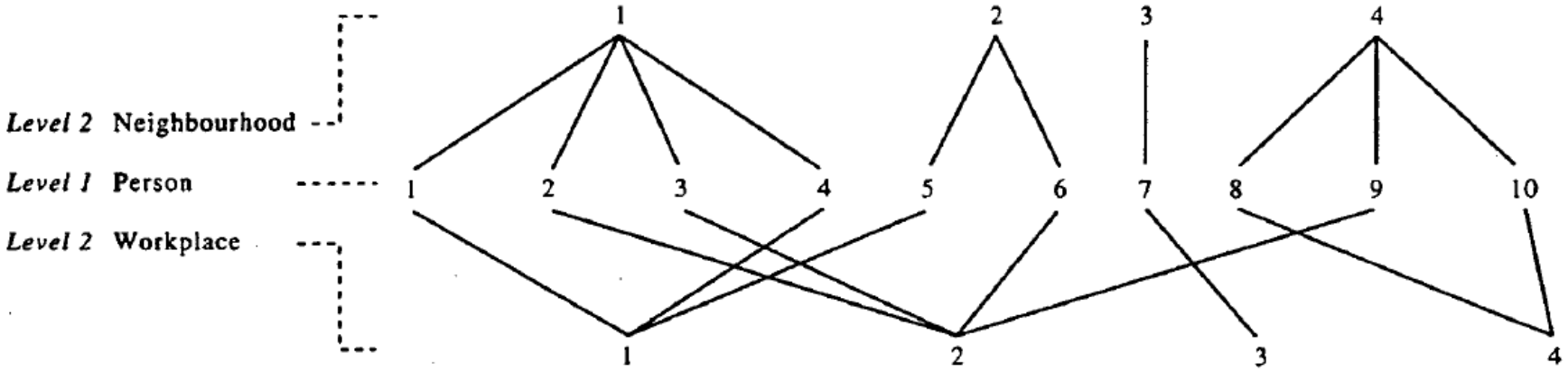
Time is level 1 because Annual observations are sampled within People in each Place

Multivariate responses



e.g., Y1=walking speed, Y2=balance measure, Y3=grip strength and Y4=lung function

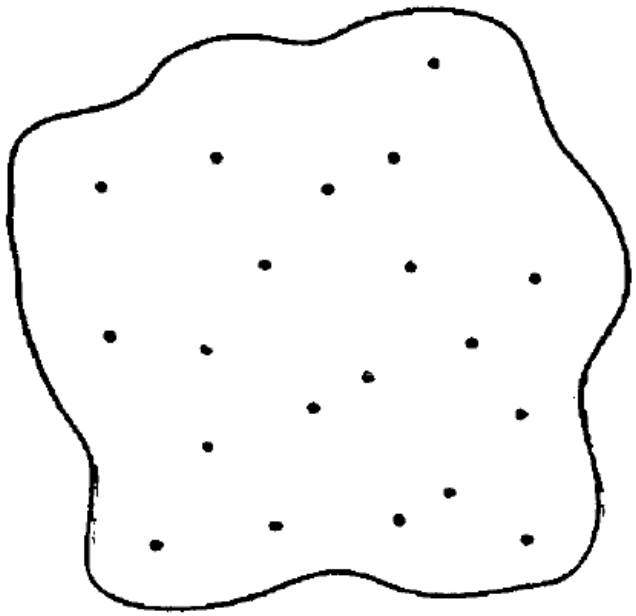
Cross-classified models



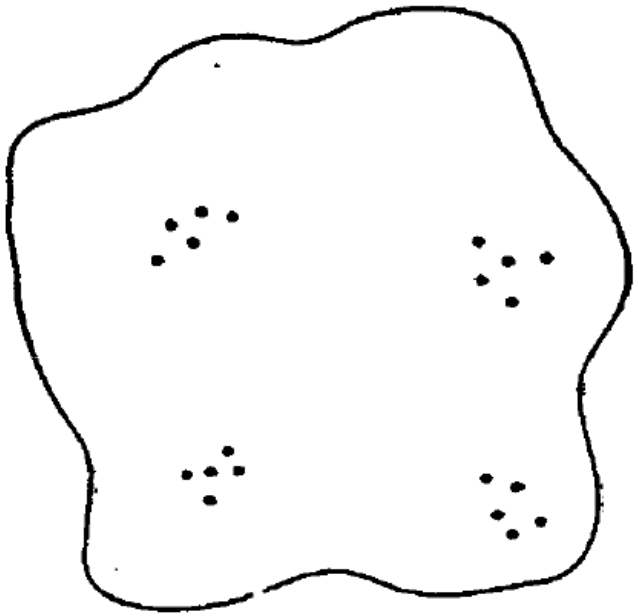
- People are nested within neighborhoods and also within worksites.
- Some people who live in the same neighborhood work at the same place, others do not.

Sampling design creates clustering

a) Simple random sample



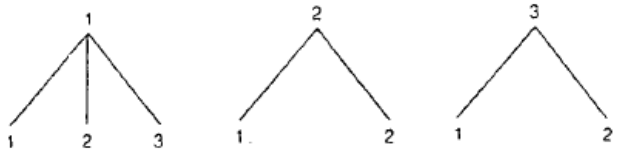
b) Two stage sample



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

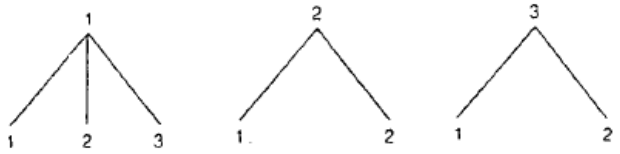
Questions potentially of interest: measures of people in a place



- 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?

Things we hardly ever ask

- 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
- 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

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
- A 1 SD increase in neighborhood SES predicted 0.24 additional daily servings of F & V.

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

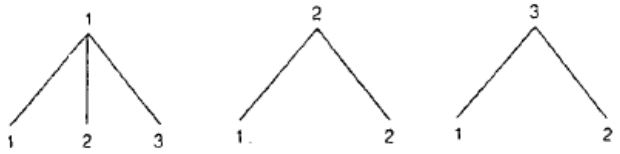
Dubowitz, et al Am J Clin nutr

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 hba1c<6.5 and all the diabetics at clinic B have hba1c>8, it makes you think that clinic A is doing something right.
- Common to start with a model accounting only for individual characteristics, and estimate intraclass correlation coefficients
- Then add contextual variables to try to eliminate
- 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.

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)

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 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 people/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= $s^2_{\text{between-cluster}}/(s^2_{\text{between-cluster}} + s^2_{\text{within-cluster}})$
- 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 conditional or unconditional ICC

Dubowitz, et al Am J Clin nutr

Approximate design effect formula for estimation of the *mean*

$DE = s^2_{cluster}/s^2_{SRS} = 1 + rho_X(m_{avg} - 1)$

DE = 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

s^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.

- Increases if rho goes up.
- 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

DE = 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=DE* 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 25th 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

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

So if either rho_X or rho_Y is 0, DEFF is one.

Note that for level 1 covariates, accounting for clustering can improve SEs, because it reduces extraneous variance in the outcome.

3 Situations

- Substantive question is level 2 exposure (greenspace)
- Substantive question is level 1 exposure, level 1 exposure is not clustered (male)
- Substantive question is level 1 exposure, level 1 exposure *is* clustered (BMI)

```
set obs 200
gen nbhd=_n
gen greenspace=uniform()>.3
gen U_nbd=uniform()>.5
expand 25
sort nbhd
by nbhd: gen id=_n
gen male=uniform()<.4
gen BMI=invnorm(uniform())+1*U_nbd      * this induces clustering and confounding
gen U_ind=uniform()<.5
gen sbp=.3*greenspace+1*U_nbd+.2*male+.2*BMI+.1*U_ind+invnorm(uniform())
```

3 Situations

```
. reg sbp greenspace male BMI

      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  3,  4996) =  294.14
       Model |  1045.02836     3  348.342786            Prob > F      =  0.0000
    Residual |  5916.71436  4996   1.1842903           R-squared     =  0.1501
-------------+------------------------------           Adj R-squared =  0.1496
       Total |  6961.74272  4999  1.39262707           Root MSE      =  1.0883

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  greenspace |   .2223232   .0341066     6.52   0.000     .1554593    .2891871
        male |   .1829174   .0313513     5.83   0.000     .1214551    .2443798
         BMI |   .3929799   .0137963    28.48   0.000     .3659331    .4200267
       _cons |   .4828133   .0322205    14.98   0.000     .4196469    .5459796
```

3 Situations

```
. reg sbp greenspace male BMI

      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(  3,  4996) =  29
       Model |  1045.02836     3  348.342786            Prob > F      =  0.
    Residual |  5916.71436  4996   1.1842903           R-squared     =  0.
-------------+------------------------------           Adj R-squared =  0.
       Total |  6961.74272  4999  1.39262707           Root MSE      =  1.

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Inter
-------------+----------------------------------------------------------------
  greenspace |   .2223232   .0341066     6.52   0.000     .1554593    .289
        male |   .1829174   .0313513     5.83   0.000     .1214551    .244
         BMI |   .3929799   .0137963    28.48   0.000     .3659331    .420
       _cons |   .4828133   .0322205    14.98   0.000     .4196469    .545
```

```
. xtmixed sbp greenspace male BMI || nbhd:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -7223.7291
Iteration 1:   log likelihood = -7223.7291

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      5000
Group variable: nbhd                            Number of groups   =       200

                                                Obs per group: min =        25
                                                               avg =      25.0
                                                               max =        25

                                                Wald chi2(3)       =    329.72
Log likelihood = -7223.7291                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  greenspace |   .2110027   .0833436     2.53   0.011     .0476523    .3743532
        male |   .1959772    .028909     6.78   0.000     .1393165    .2526378
         BMI |    .229141   .0138229    16.58   0.000     .2020487    .2562333
       _cons |   .5603388   .0717547     7.81   0.000     .4197022    .7009755
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
----------------------------+-------------------------------------------------
nbhd: Identity              |
                  sd(_cons) |   .4941141     .02946      .4396196    .5553637
----------------------------+-------------------------------------------------
               sd(Residual) |   .9862128   .0100773       .966658    1.006163
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   583.64 Prob >= chibar2 = 0.0000
```

3 Situations

```
. reg sbp greenspace male BMI

      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(  3,  4996) =  29
       Model |  1045.02836     3  348.342786            Prob > F      =  0.
    Residual |  5916.71436  4996   1.1842903           R-squared     =  0.
-------------+------------------------------           Adj R-squared =  0.
       Total |  6961.74272  4999  1.39262707           Root MSE      =  1.

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Inter
-------------+----------------------------------------------------------------
  greenspace |   .2223232   .0341066     6.52   0.000     .1554593    .289
        male |   .1829174   .0313513     5.83   0.000     .1214551    .244
         BMI |   .3929799   .0137963    28.48   0.000     .3659331    .420
       _cons |   .4828133   .0322205    14.98   0.000     .4196469    .545
```

```
. xtmixed sbp greenspace male BMI || nbhd:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -7223.7291
Iteration 1:   log likelihood = -7223.7291

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      5000
Group variable: nbhd                            Number of groups   =       200

                                                Obs per group: min =        25
                                                               avg =      25.0
                                                               max =        25

                                                Wald chi2(3)       =    329.72
Log likelihood = -7223.7291                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nspace       |   .2110027   .0833436     2.53   0.011     .0476523    .3743532
        male |   .1959772    .028909     6.78   0.000     .1393165    .2526378
         BMI |    .229141   .0138229    16.58   0.000     .2020487    .2562333
       _cons |   .5603388   .0717547     7.81   0.000     .4197022    .7009755
------------------------------------------------------------------------------

------------------------------------------------------------------------------
om-effects Parameters |  Estimate   Std. Err.      [95% Conf. Interval]
----------------------+-------------------------------------------------------
Identity              |
             sd(_cons) |  .4941141     .02946      .4396196    .5553637
----------------------+-------------------------------------------------------
          sd(Residual) |  .9862128   .0100773       .966658    1.006163
------------------------------------------------------------------------------
t vs. linear regression: chibar2(01) =   583.64 Prob >= chibar2 = 0.0000
```

```
GEE population-averaged model                   Number of obs      =      5000
Group variable:                        nbhd     Number of groups   =       200
Link:                                  identity Obs per group: min =        25
Family:                                Gaussian                avg =      25.0
Correlation:                           exchangeable            max =        25
                                                Wald chi2(3)       =    344.65
Scale parameter:                       1.216765 Prob > chi2        =    0.0000

                          (Std. Err. adjusted for clustering on nbhd)
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  greenspace |   .2110027    .086821     2.43   0.015     .0408366    .3811688
        male |   .1959772   .0283322     6.92   0.000     .1404471    .2515072
         BMI |    .229141   .0133301    17.19   0.000     .2030145    .2552675
       _cons |   .5603388   .0760099     7.37   0.000     .4113622    .7093155
------------------------------------------------------------------------------
```

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

Inefficient estimators

- Alternatives to mixed models or GEE
  - Do men average higher SBP than women?
    - Take the average value of SBP for each person over all repeated measures
    - Treat as a single outcome measure and regress on sex
    - Choose one observation and use that
  - Do residents of neighborhoods with greenspace average lower SBP?
    - Average SBP for all individuals in each neighborhood
    - Treat n=# of neighborhoods
    - Inefficient with unbalanced data

GEE uses most efficient weighting of correlated observations

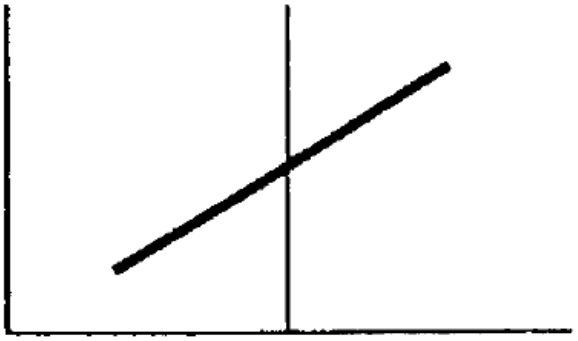
The optimal weights for combining the responses of individuals from clusters of size k are: 1/(1+(k-1)*R), where R=correlation of pairs of observations from the same cluster
i.e., Downweight individuals from large clusters, in proportion to the intracluster correlation.

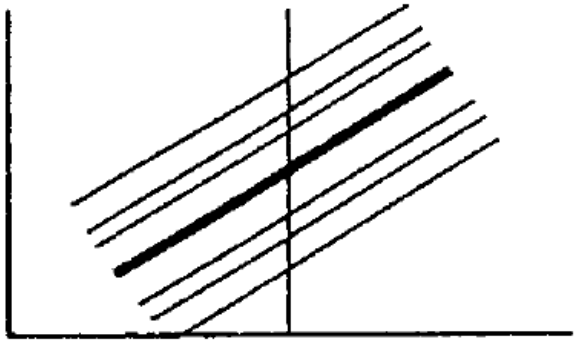
Estimator*	Form	Variance of estimator†			
		In general‡	If $R = 0$	If $R = 0.5$	If $R = 1.0$
$\bar{y}_{1:1:1}$	$\frac{1}{3} y_{\text{singleton}} + \frac{1}{3} y_{\text{sib1}} + \frac{1}{3} y_{\text{sib2}}$	$\frac{3+2R}{9}\sigma^2$	$\frac{\sigma^2}{3}$	$\frac{4\sigma^2}{9}$	$\frac{5\sigma^2}{9}$
$\bar{y}_{1:1}$	$\frac{1}{2} y_{\text{singleton}} + \frac{1}{2} y_{\text{sib1 or sib2}}$	$\frac{1}{2}\sigma^2$	$\frac{\sigma^2}{2}$	$\frac{\sigma^2}{2}$	$\frac{\sigma^2}{2}$
$\bar{y}_{2:1:1}$	$\frac{2}{4} y_{\text{singleton}} + \frac{1}{4} y_{\text{sib1}} + \frac{1}{4} y_{\text{sib2}}$	$\frac{3+R}{8}\sigma^2$	$\frac{3\sigma^2}{8}$	$\frac{7\sigma^2}{16}$	$\frac{\sigma^2}{2}$
$\bar{y}_{\text{optimal}}$	$\frac{1+R}{3+R} y_{\text{singleton}} + \frac{1}{3+R} y_{\text{sib1}} + \frac{1}{3+R} y_{\text{sib2}}$	$\frac{1+R}{3+R}\sigma^2$	$\frac{\sigma^2}{3}$	$\frac{3\sigma^2}{7}$	$\frac{\sigma^2}{2}$

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

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) and everyone in the same place has this same little bit of offset, plus their own personal differences.

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

Wide vs Tall Data sets

- Wide

id	X_time1	X_time2	X_time3
1	27	32	45
2	18	17	20
3	41	44	46

- Tall

id	time	X
1	1	27
1	2	32
1	3	45
2	1	18
2	2	17
2	3	20
3	1	41
3	2	44
3	3	46

Basic mixed model for clustered data without ordering

Two level random intercepts model for person *i* in place *j*:

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

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

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

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:

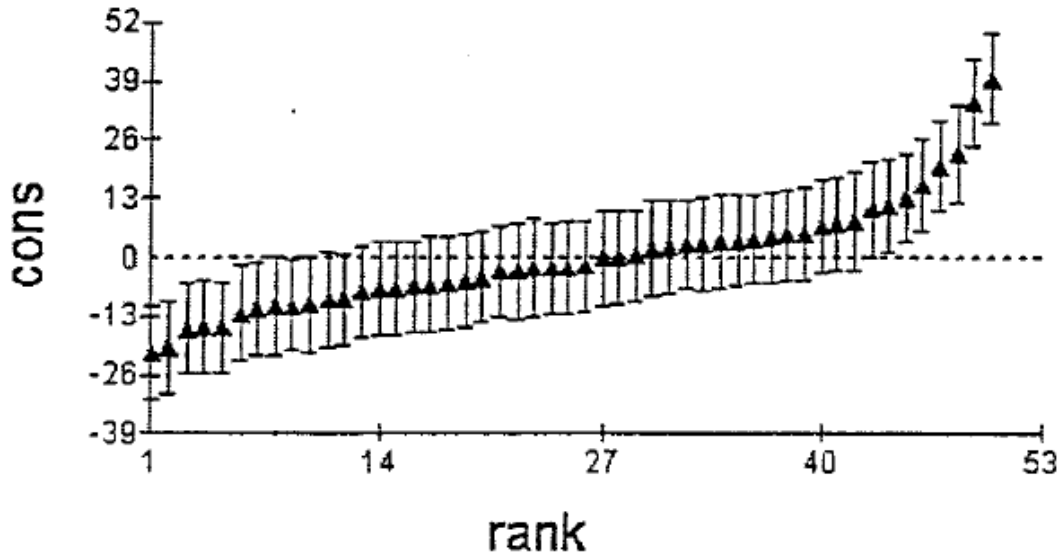
$$[\text{Overall model}] = y_{ij} = \beta_0 + \beta_1 x_{ij} + (\mu_{0j} + \varepsilon_{ij})$$

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

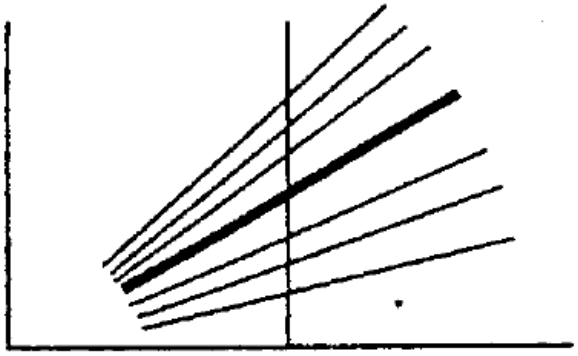
$\mu_{0j} \sim N(0, \sigma^2_{\mu 0})$

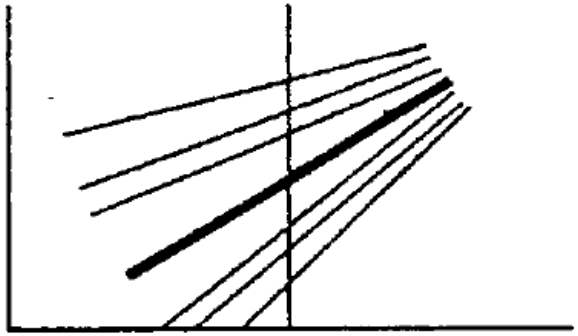
$\varepsilon_{ij} \sim N(0, \sigma^2_e)$

Can imagine rank ordering the μ for each place. Each μ is shrunk towards zero based on uncertainty in place specific estimate.



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” is different.

Random slopes model

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

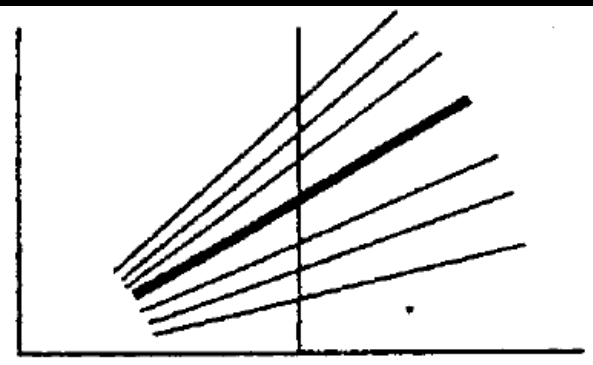
$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \varepsilon_{ij} \text{ [Micro model]}$$

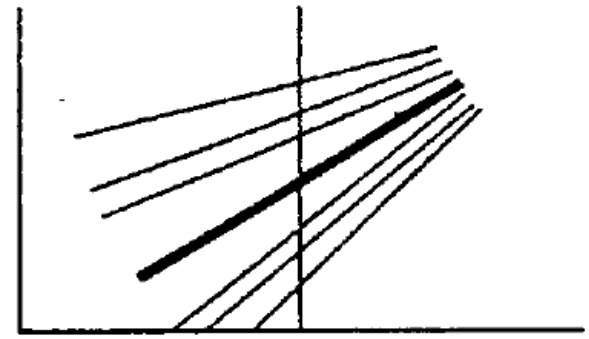
$$\beta_{0j} = \beta_0 + \mu_{0j} \text{ [Macro model for intercept]}$$

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

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (\mu_{1j} + \mu_{0j} + \varepsilon_{ij}) \text{ [Overall model]}$$

Random slopes model

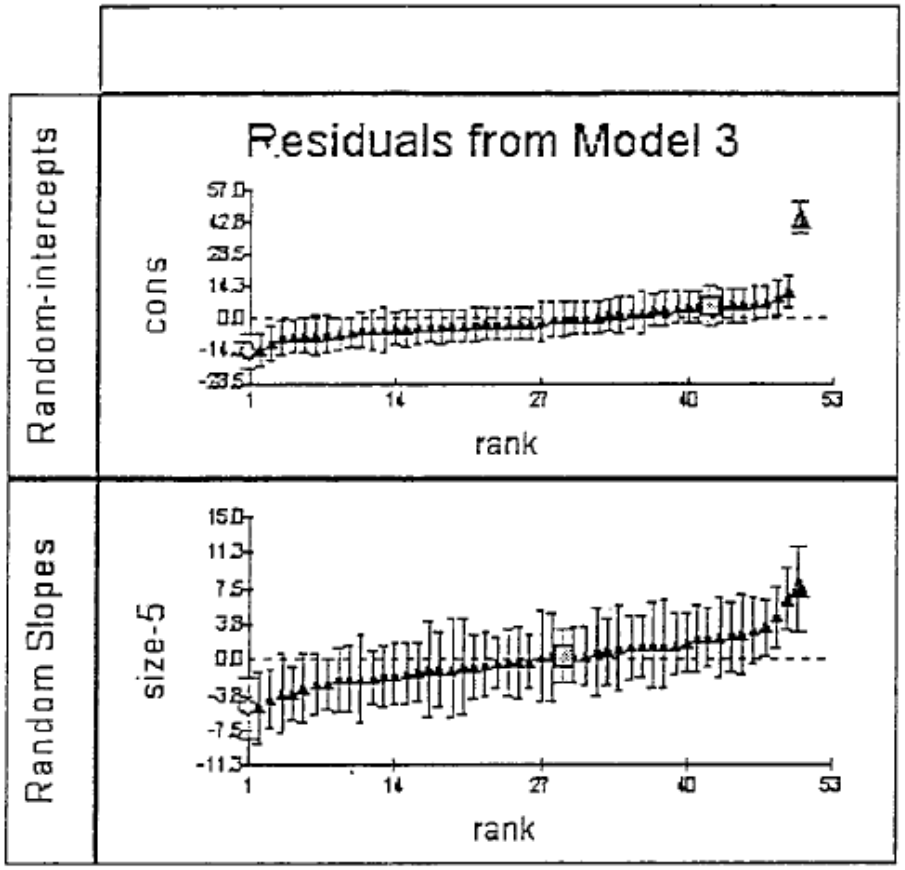


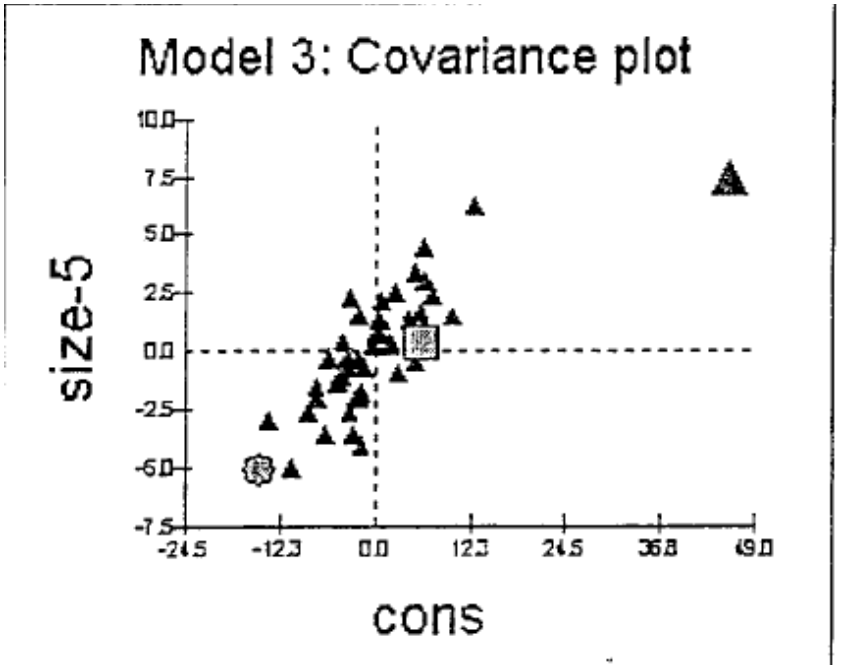


Now you have to consider the correlation of μ_{0j} and μ_{1j} for each place j.

- μ_{0j} and μ_{1j} positively correlated: places that have high values for the reference group have larger effect sizes of exposure.
- μ_{0j} and μ_{1j} negatively correlated: places that have high values for the reference group have smaller effect sizes of exposure.
- μ_{0j} and μ_{1j} uncorrelated

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

Random slopes model

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (\mu_{1j} + \mu_{0j} + \varepsilon_{ij})$$ **[Overall model]**

Random part as a variance/covariance matrix

Level-1

Variance: σ^2_{ε}

Level-2

Variance for slopes: $\sigma^2_{\mu 1}$

Variance for intercepts: $\sigma^2_{\mu 0}$

Covariance: $\sigma_{\mu 0 \mu 1}$

Random slopes model

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (\mu_{1j} + \mu_{0j} + \varepsilon_{ij})$$ **[Overall model]**

Level-2 (Between-dsitricts)

$$\begin{bmatrix} \mu_{0j} \\ \\ \mu_{1j} \end{bmatrix} \sim N(0,\ \Omega_\mu): \Omega_\mu \quad \begin{matrix} \\ (x_0) \\ (x_1) \end{matrix} \begin{bmatrix} (x_0) & (x_1) \\ \sigma^2_{\mu 0} & \\ \sigma_{\mu 0 \mu 1} & \sigma^2_{\mu 1} \end{bmatrix}$$

Level-1 (Between-houses within districts)

$$\begin{bmatrix} \varepsilon_{ij} \end{bmatrix} \sim N(0, \Omega_\varepsilon): \Omega_\varepsilon \quad \begin{matrix} (x_0) \end{matrix} \begin{bmatrix} \sigma^2_\varepsilon \end{bmatrix}$$

Distinguish the $\sigma_{\mu 0 \mu 1}$ covariance between random intercept and random slope from the correlation between observations within the same cluster. Syntax is very confusing (esp in SAS)

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

When the unit of clustering is a person

- We are usually interested in determinants of within person change, thus the descriptor "growth curves"
- Typically motivated by specifying change term using a time dimension (conceptually, age) and examining interactions with the time term
- Can be a time-constant modifier: Sex*age
- Or a time-varying modifier: depressive symptoms*age

Questions potentially of interest: repeated measures of a person

- Growth curve models are specified as interactions between time and (usually) a person-level variable (sometimes time-varying variable)
- Person level exposure → time-specific outcome
  - 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)

Classic growth curve: Willis et al Pmed S2: trajectories of SBP

$$SBP_{ij} = \beta_0 + \mu_{0i} + (\beta_1 + \mu_{1i})\, age_{ij} + (\beta_2 + \mu_{2i})\, age^2_{ij} + (\beta_3 + \mu_{3i})\, age^3_{ij} + \varepsilon_{ij}$$

- Where SBP_{ij} is the systolic blood pressure for individual *i* at observation time *j*.
- β_0 and β_1 are the fixed intercept and slope, μ_{0i} to μ_{3i} are the random intercept and slope coefficients for each individual *i*, and ε_{ij} is the residual error term for individual *i*, at time *j*.
- Covariance between the random coefficients was allowed . Age was centred at the baseline age in each cohort .

Classic growth curve: Willis et al Pmed S2: trajectories of SBP

$$SBP_{ij} = \beta_0 + \mu_{0i} + (\beta_1 + \mu_{1i})\, age_{ij} + (\beta_2 + \mu_{2i})\, age^2_{ij} + (\beta_3 + \mu_{3i})\, age^3_{ij} + \gamma_1\, zBMI_{ij} + \gamma_2\, (age_{ij}\, zBMI_{ij}) + \gamma_3\, (age^2_{ij}\, zBMI_{ij}) + \delta_1\, zbHT_i + \delta_2\, (zbHT_i\, age_{ij}) + \delta_3\, (zbHT_i\, age^2_{ij}) + \delta_4\, (zbHT_i\, age^3_{ij}) + \varepsilon_{ij}$$

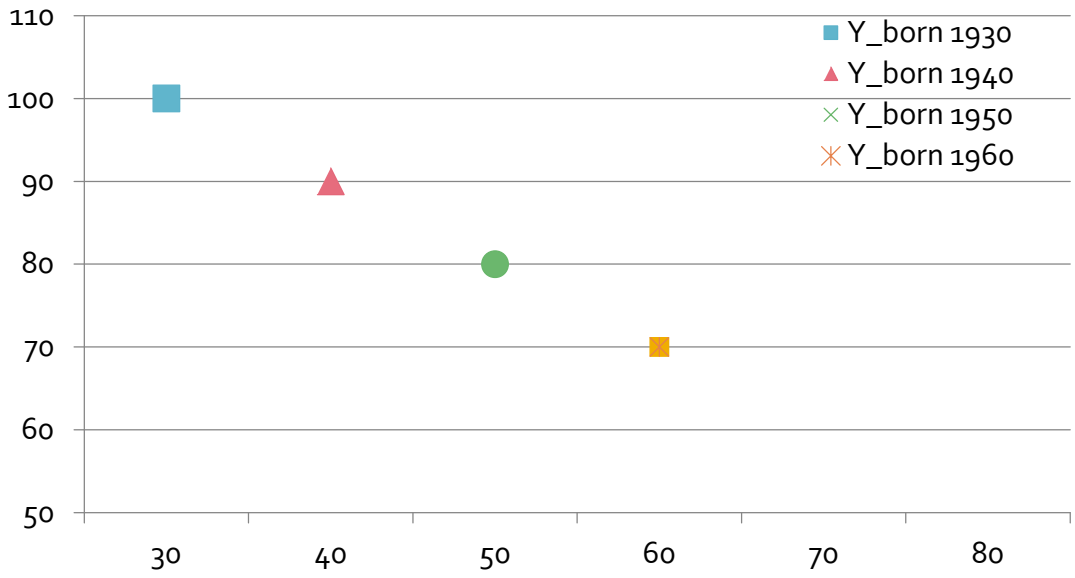
- $zBMI_{ij}$ is the BMI z-score externally referenced to the UK 1990 sex and age-specific growth charts for subject *i* at observation time *j*.
- γ_1 represents the cross sectional association between BMI and SBP
- γ_2 and γ_3 allow the association between current BMI and SBP to vary by chronological age.
- $zbHT_i$ represents the height z-score at baseline referenced to the UK 1990 sex and age-specific growth charts for subject *i*.
- δ_1, δ_2 and δ_3 thus allow baseline height to affect the SBP intercept and slope.
- Non significant γ's and δ's were removed from the final model.

What's the best time dimension?

- Age is conceptually natural (?)
  - But using age as the time-dimension conflates between person differences and within person changes
  - For studies of old people, cohort effects can bias age-based growth curves
  - Especially problematic if exposure of interest is highly correlated with age
- Using "time from baseline" and adjusting for baseline age eliminates cohort problems, but introduces potential period and practice effect problems
- Best practice: estimate separately (baseline age + time from baseline) and understand why they diverge (Fitzmaurice, Laird, and Ware, pg 420)

It's possible to estimate an age slope with cross-sectional data

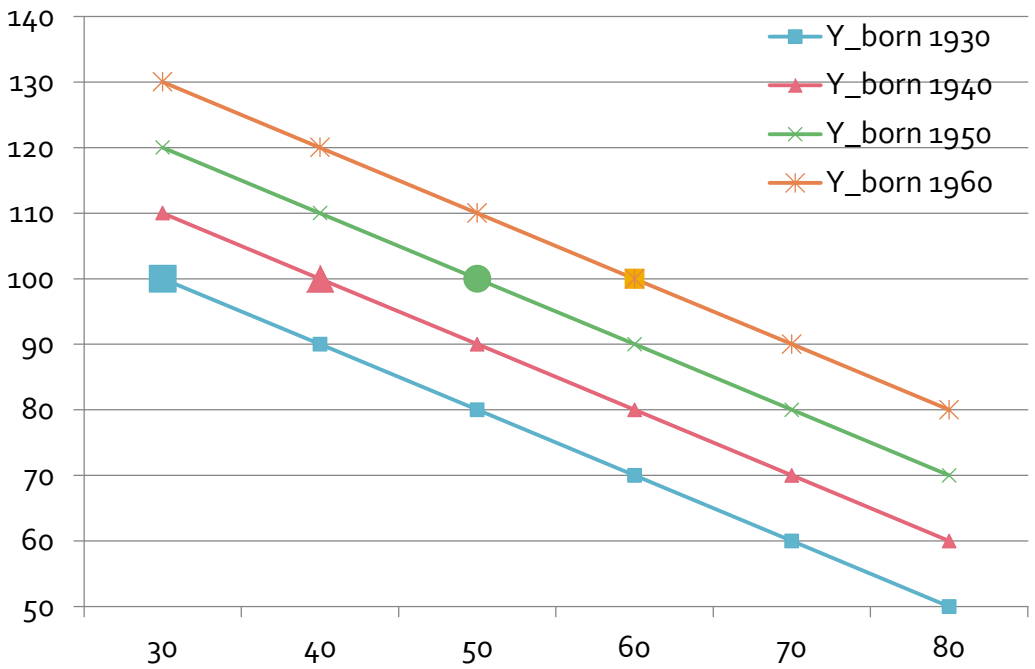
Age	Y_born 1930	Y_born 1940	Y_born 1950	Y_born 1960
30	100	100	100	100
40	90	90	90	90
50	80	80	80	80
60	70	70	70	70
70	60	60	60	60
80	50	50	50	50



But the XS data conflates age, period , and cohort differences

Age	Y_born 1930	Y_born 1940	Y_born 1950	Y_born 1960
30	100	110	120	130
40	90	100	110	120
50	80	90	100	110
60	70	80	90	100
70	60	70	80	90
80	50	60	70	80

Imagine a large cohort effect (10 points higher for every decade later birth), but identical age slopes

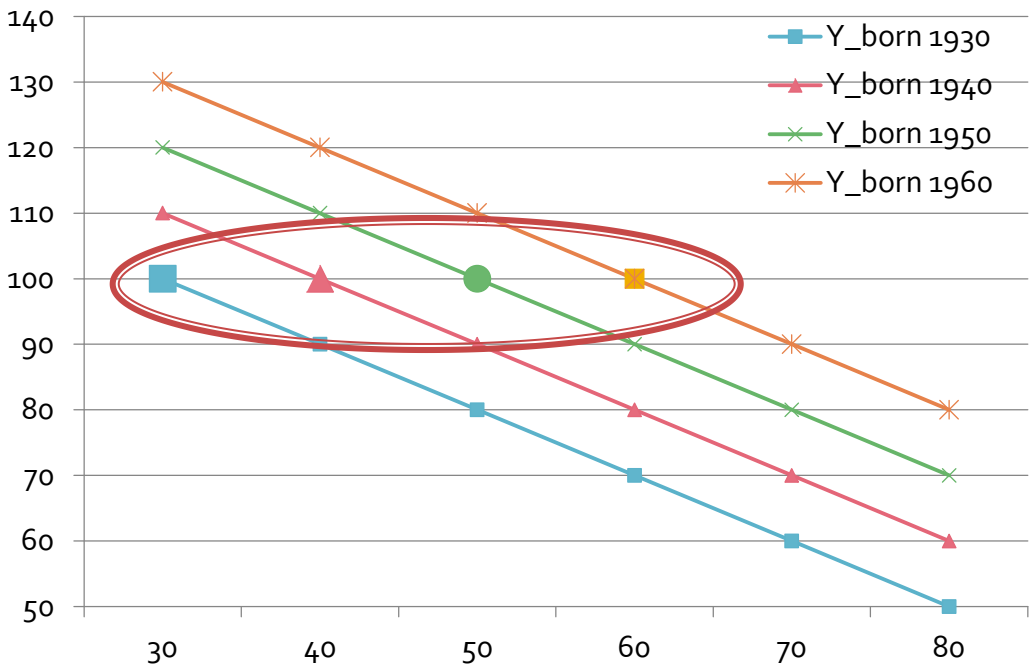


Regressing Y on age in a cross-sectional data set leads to a flattened age slope

But the XS data conflates age, period , and cohort differences

Age	Y_born 1930	Y_born 1940	Y_born 1950	Y_born 1960
30	100	110	120	130
40	90	100	110	120
50	80	90	100	110
60	70	80	90	100
70	60	70	80	90
80	50	60	70	80

Imagine a large cohort effect, but identical age slopes

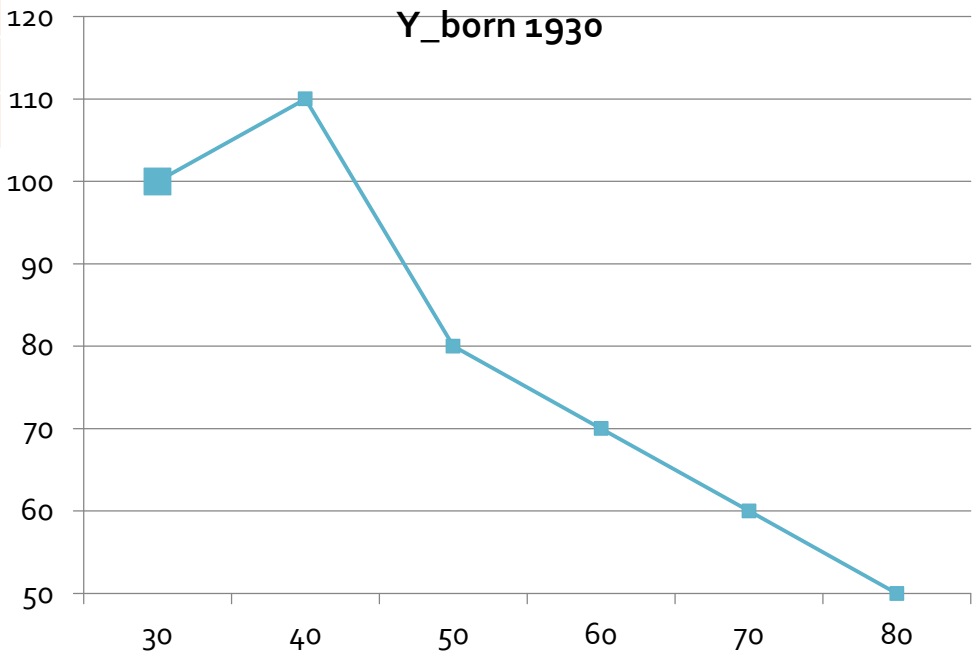


Regressing Y on age in a cross-sectional data set leads to a flattened age slope

But the XS data conflates age, period , and cohort differences

Age	Y_born 1930	Y_born 1940	Y_born 1950	Y_born 1960
30	100	101	102	103
40	110	91	92	93
50	80	101	82	83
60	70	71	92	73
70	60	61	62	83
80	50	51	52	53

Imagine almost no cohort effects, but large period effect (added 20 points in 1970)

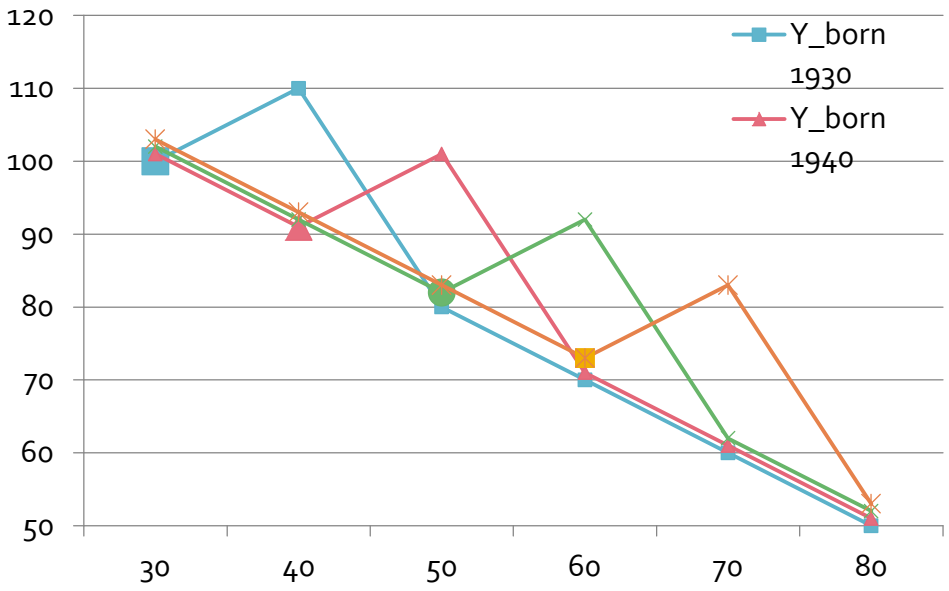


Fitting a straight line to age in a single birth cohort will lead to a much flatter age-slope estimate

But the XS data conflates age, period , and cohort differences

Age	Y_born 1930	Y_born 1940	Y_born 1950	Y_born 1960
30	100	101	102	103
40	110	91	92	93
50	80	101	82	83
60	70	71	92	73
70	60	61	62	83
80	50	51	52	53

Imagine almost no cohort effects, but large period effect (added 20 points in 1970)



Fitting a straight line to age in a single birth cohort will lead to a much flatter age-slope estimate- XS data could actually get the right estimate

What's the best time dimension?

- Best practice: estimate separately (baseline age + time from baseline) and understand why they diverge (Fitzmaurice, Laird, and Ware, pg 420)
- $Y_{ij}=\beta_{00}+\beta_1*age\_base+\beta_2*(age\_now-age\_base)+\mu_i+\varepsilon_{ij}$
- $\beta_1 \sim=\beta_2$? If so, then:
- $Y_{ij}=\beta_{00}+\beta_1*age\_now+\mu_i+\varepsilon_{ij}$
- If not, why?
- Can examine practice, period, etc, e.g., :
- $Y_{ij}=\beta_{00}+\beta_1*age\_base+\beta_2*(age\_now-age\_base)+\beta_3*(1^{st}\ wave)+\mu_i+\varepsilon_{ij}$

Alternative time dimensions

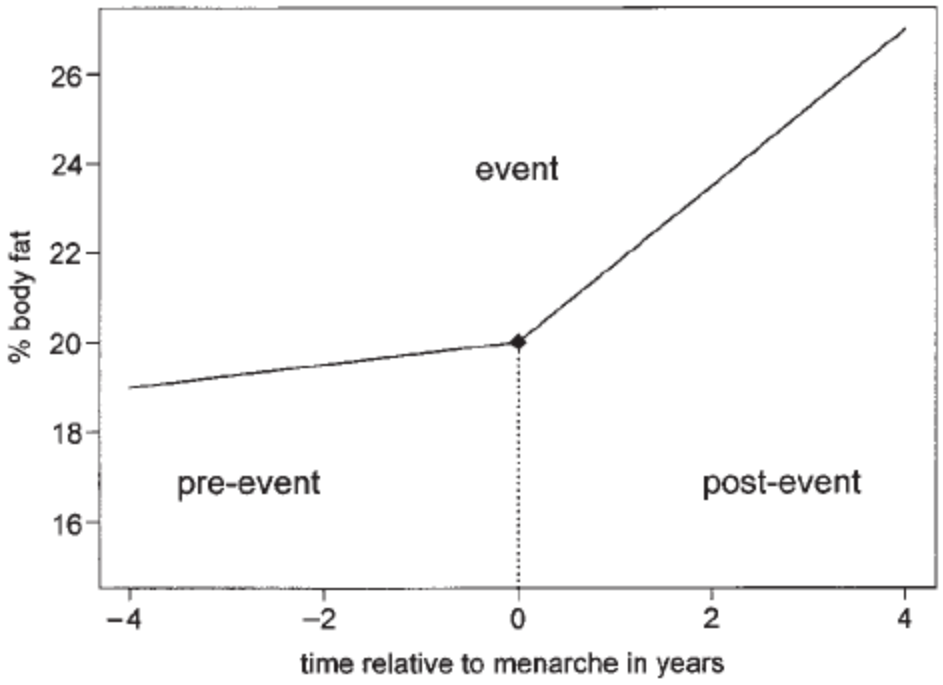


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

Choose a time-dimension:

- Age
- Time before or after a major event
- Secular time

Specify a model with time as a linear predictor

- How does body fat change per year of age?

Add a factor that modifies the rate of change with time

- Is the rate of change different before (δ=1) or after (δ=0) menarche?

First the model for before menarche (δ=1)

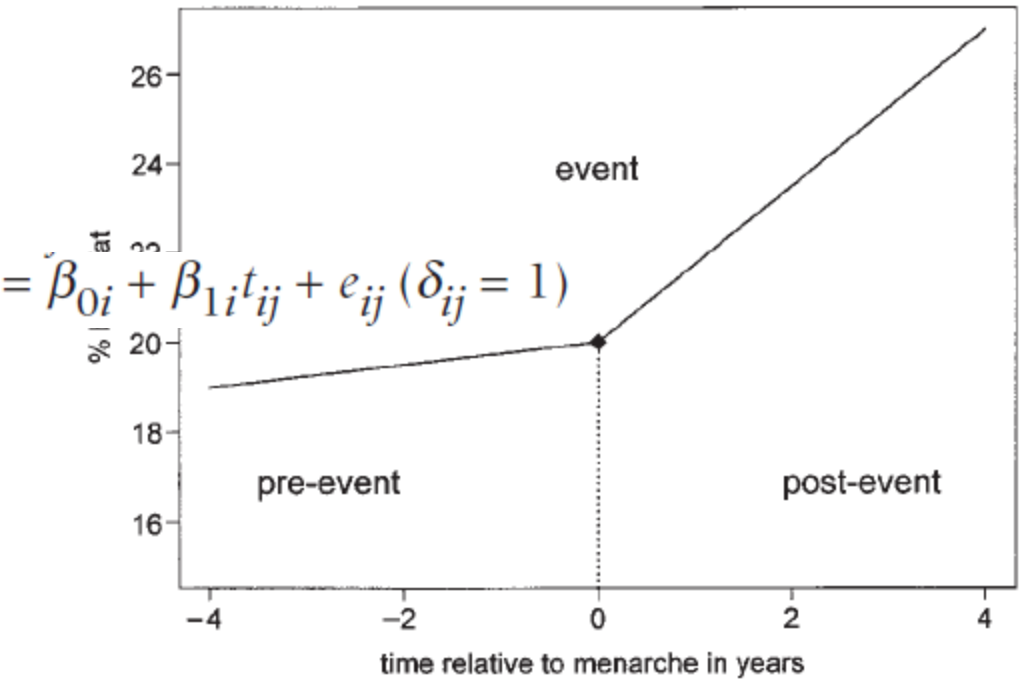


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

Choose a time-dimension:

- Age
- Time before or after a major event
- Secular time

Specify a model with time as a linear predictor

- How does body fat change per year of age?

Add a factor that modifies the rate of change with time

- Is the rate of change different before (δ=1) or after (δ=0) menarche?

Then the model for after menarche (δ=0)

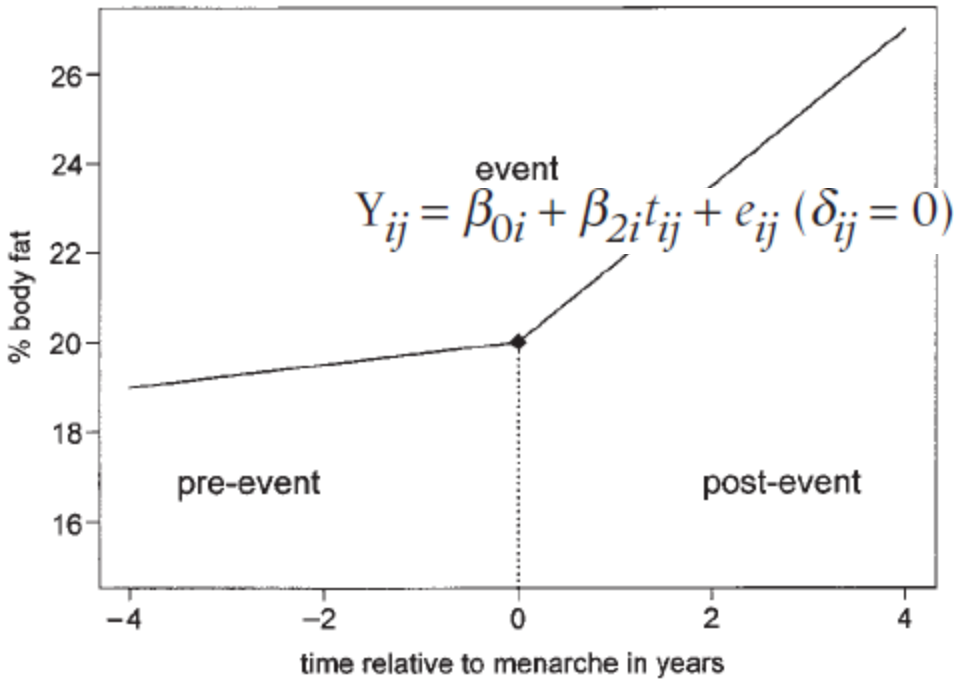


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

Choose a time-dimension:

- Age
- Time before or after a major event
- Secular time

Specify a model with time as a linear predictor

- How does body fat change per year of age?

Add a factor that modifies the rate of change with time

- Is the rate of change different before (δ=1) or after (δ=0) menarche?

Then the combined model

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} \delta_{ij} + \beta_{2i} t_{ij} (1 - \delta_{ij}) + e_{ii,}$$

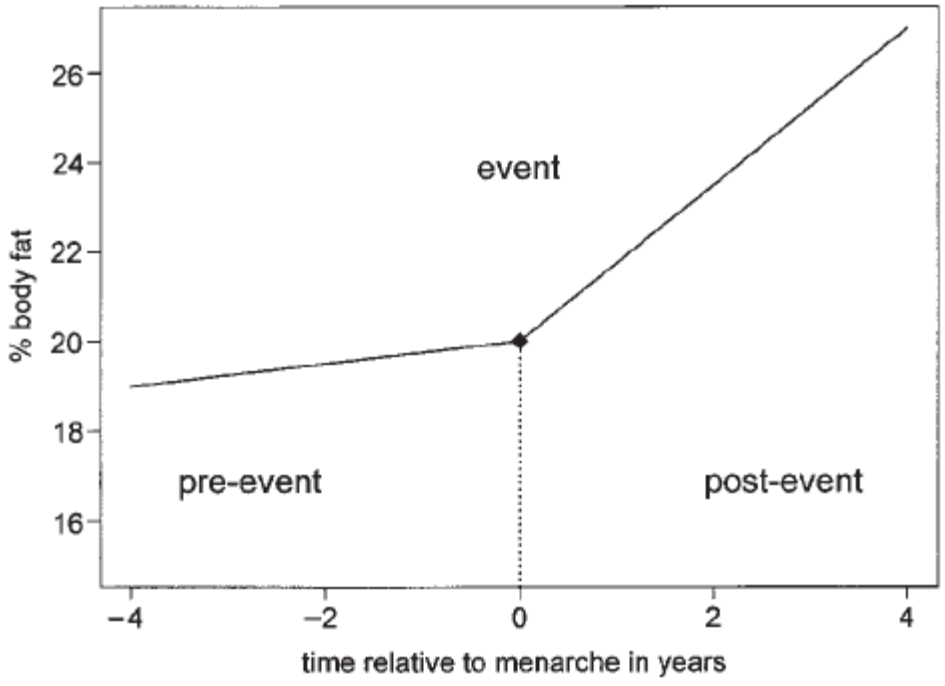


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

Choose a time-dimension:

- Age
- Time before or after a major event
- Secular time

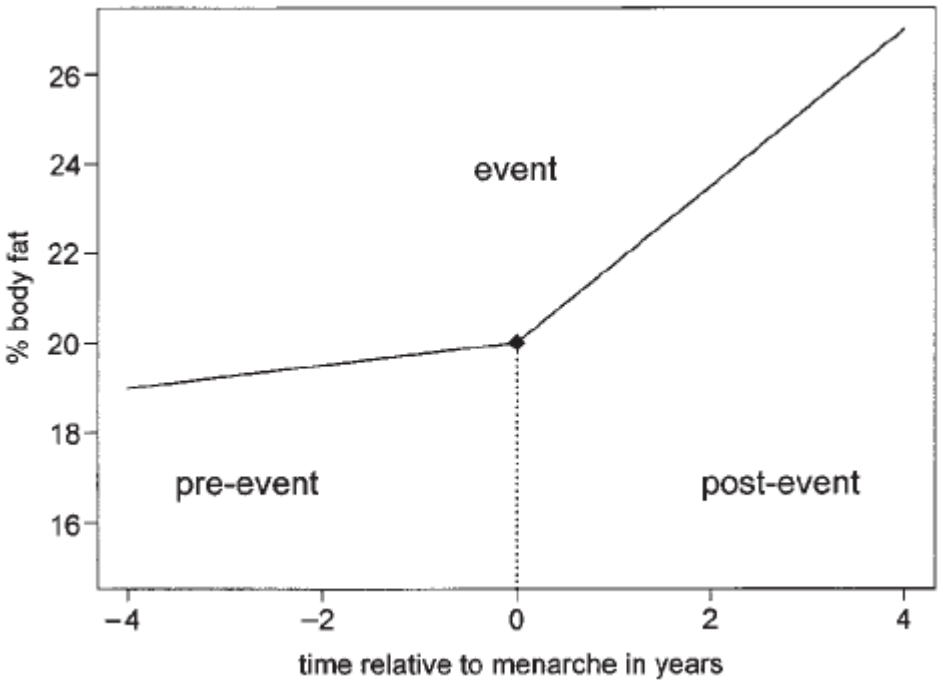
Specify a model with time as a linear predictor

- How does body fat change per year of age?

Add a factor that modifies the rate of change with time

- Is the rate of change different before (δ=1) or after (δ=0) menarche?

Provides a flexible test of multiple questions



Does rate of change differ before or after menarche?

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} \delta_{ij} + \beta_{2i} t_{ij} (1 - \delta_{ij}) + e_{ii,}$$

Does rate of change differ after menarche depend on parental obesity (v_i)?

$$Y_{ij}=\beta_0+\beta_1 t_{ij}\delta_{ij}+\beta_2 t_{ij}(1-\delta_{ij})+\beta_3 v_i+ \beta_4 t_{ij}(1-\delta_{ij})v_i+\varepsilon_{ij}$$

This equation estimates a slope for pre-menarche and post-menarche, but lets the post-menarche slope differ by parental obesity

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
- Lifecourse timing

Lifecourse timing

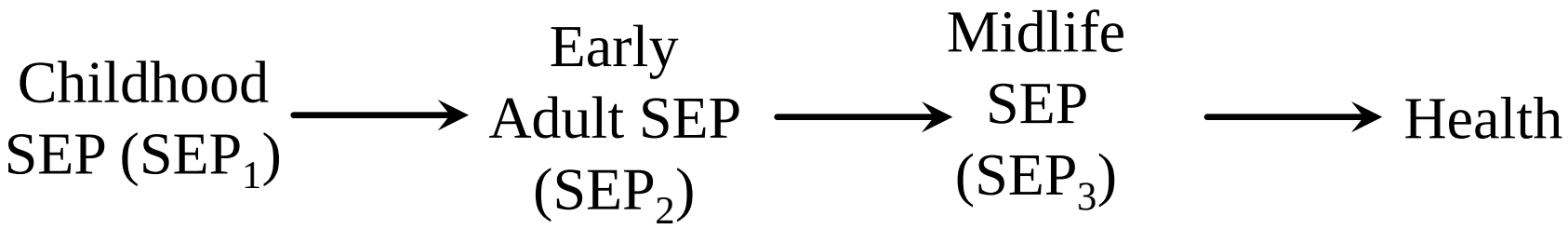
- Alternative models: DAGs and statistical claims corresponding with the DAGs
- How you could distinguish and why you would bother
- Conventional approaches and limitations of those approaches

Etiologic Model: Immediate Risk

- After the exposure is removed, the risk starts to decline.

Etiologic Models: Immediate Risk + Social Trajectory

- After the exposure is removed, the risk starts to decline.
- Social conditions at each moment set the stage for the next period: the trajectory is sticky. The initial insult does not directly harm later health – but the first exposure leads to subsequent exposures that then directly affect health.

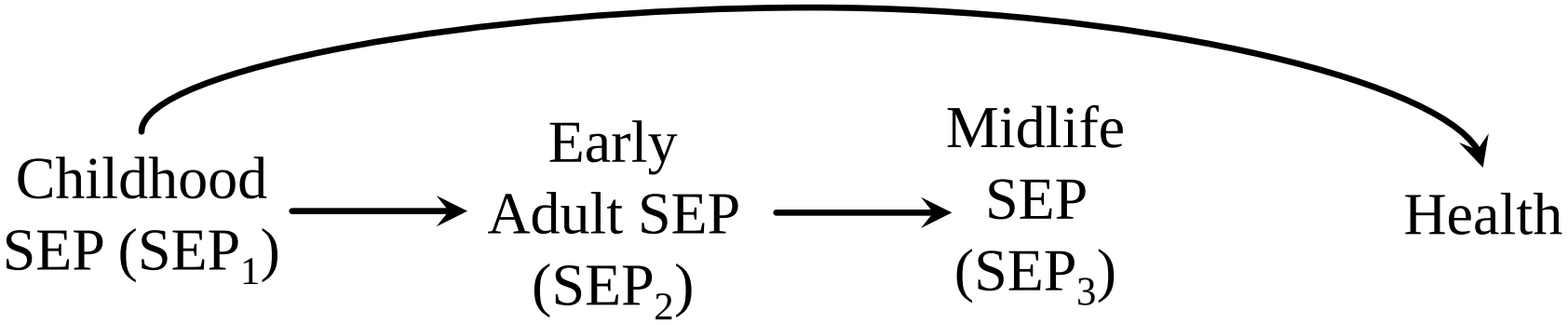


- Consistent with chain of risk model

71

Critical Period/Latency (+ Social Trajectory)

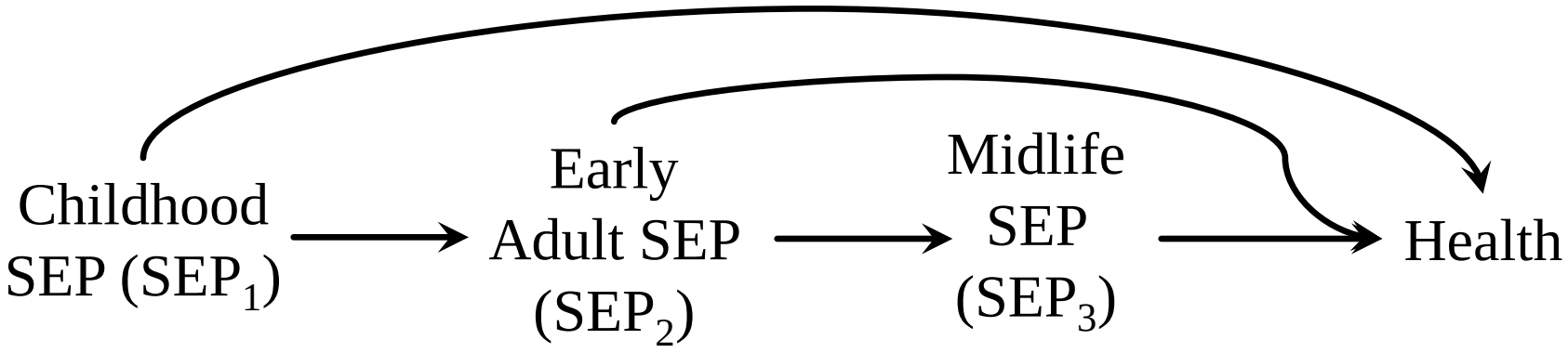
- Exposure in one period makes all the difference: subsequent exposures do not influence health.



- Also called programming models
- Example: fetal origins of disease
- Critical periods may occur in adulthood/old age, may be social, rather than biological

Etiologic Models: Cumulative Risk

- Each exposure period has an independent effect on risk.

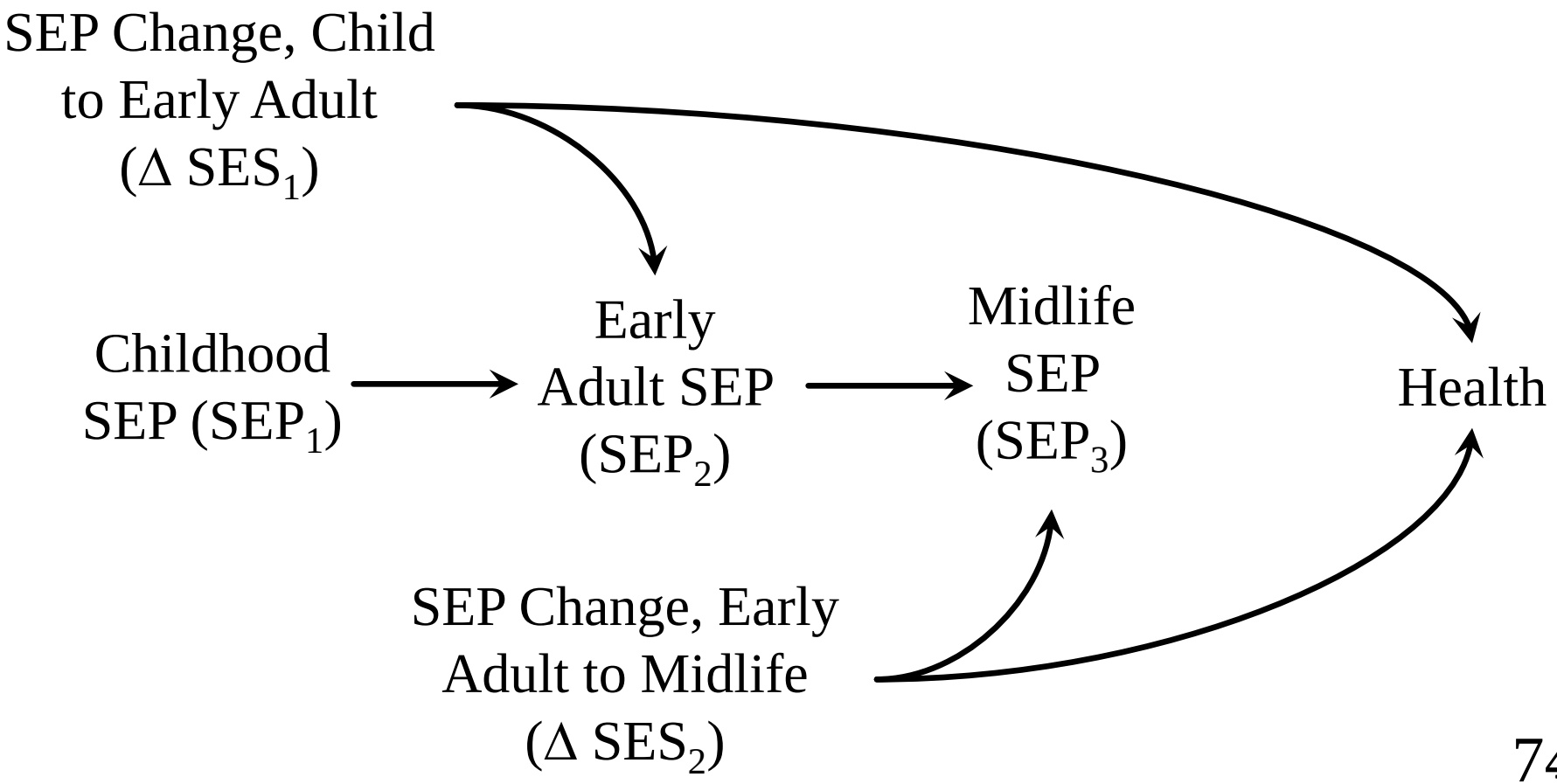


- Also consistent with chain of risk model
- Example: allostatic load

73

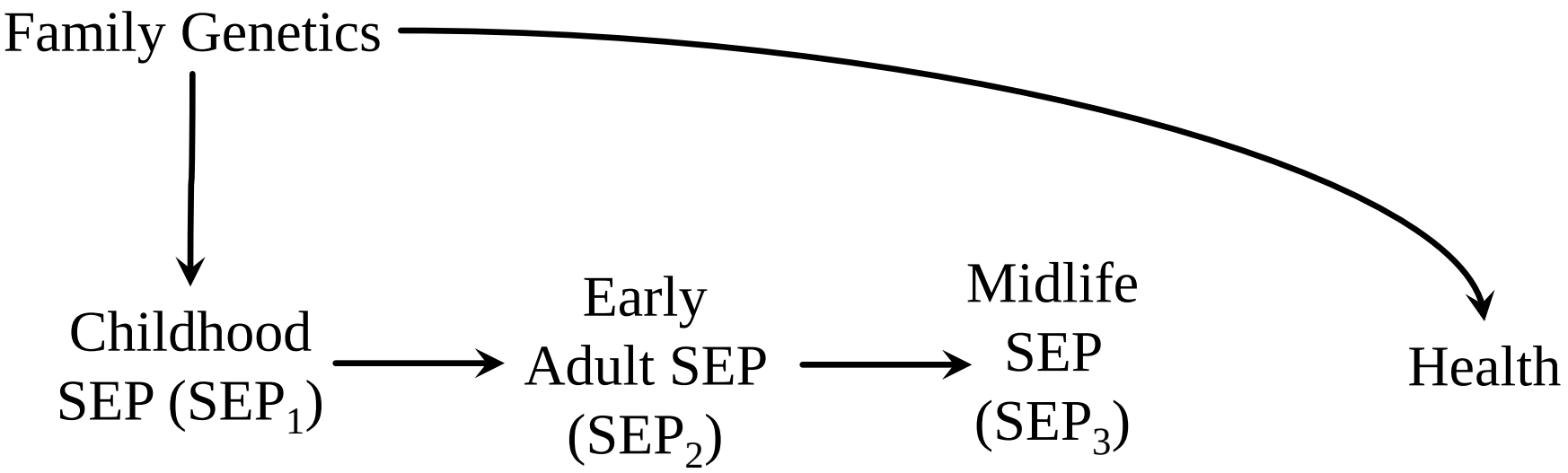
Physiological Effects of Trajectory/Change

- Change itself, rather than level of SEP, influences health.



Etiologic Models: All Confounding

- Social conditions have no effect on adult health.
- The association between social conditions and adult health is entirely attributable to confounding by unmeasured characteristics such as genetic background.



75

Implications of Alternatives

- All confounding: no effect of SES on adult health.
- Immediate risk: childhood SES does not affect adult health
- Social trajectory: childhood SES has no direct effect on adult health other than that mediated by adult SES
- Latency: adult SES has no effect on adult health, but the association between adult SES and adult health may be confounded by childhood SES
- Cumulative: childhood SES indirectly affects adult health, via adult SES, and directly affects adult health, via pathways not mediated by adult SES.

Implications for Prioritizing Resources

- Under all but the "immediate risk" and "all confounding" model, intervening to change childhood SEP would improve adult health outcomes
- Under the latency and cumulative risk model, it is not possible to completely remediate the harm from childhood disadvantage in adulthood: action must be taken in childhood.
- Under the latency/critical childhood period model, intervening to improve adult SES doesn't even improve adult health.

Lifecourse Models Guide Intervention Design/ Policy Priorities

- Policy/financial tradeoffs between:
  - Timing of educational investments (pre-school, secondary school completion, university attendance, adult training, continuing education)
  - Timing of financial security programs: pregnancy, early childhood, old age
  - Housing policies: moves away from concentrated poverty areas to low/mixed poverty communities

Lifecourse Models Guide Intervention Design/ Policy Priorities

- "In contrast to the documentation of significant long-term effects from model preschool interventions, later remediation efforts have been shown to be considerably less effective … Similarly, public job training programs, adult literacy services, prisoner rehabilitation programs, and education programs for disadvantaged adults have yielded low economic returns, with the returns for males often being negative (19)." Knudson, et al. 2006

Conventional Approach: Total Effect of Child SES on Adult Health

- To distinguish latency and cumulative from immediate risk:
- E(health2)=b0+b1*SES1

- If you have some measured confounders (common causes or factors on a common cause pathway):
- E(health2)=b0+b1*SES1+b2*Age+b3*Birth Regn

Conventional Approach: Total Effect of Child SES on Adult Health

- E(health2)=b0+b1*SES1
- E(health2)=b0+b1*SES1+b2*Age+b3*Birth Regns

- Important: to estimate the effect of an exposure on an outcome, do not adjust for things that are affected by the exposure!
- Especially do not adjust for mediators on the pathway between the exposure and the outcome!

Conventional Approach: Direct Effect of Child SES on Adult Health

- $E(health_2)=a_0+a_1*SES_1+a_2*SES_2$

- Challenge: should not adjust for mediators of childhood SES but most measured variables are not temporally prior to early life SES

Conventional Approach: Indirect Effect of Child SES on Adult Health

- Indirect Effect=Total Effect-Direct Effect
  - Indirect Effect=b1-a1

- This decomposition assumes:
  - Indirect Effect + Direct Effect = 1

Conventional Approach: Cumulative vs Latency

- Does adult SES predict health conditional on child SES:
  - $E(health_2)=a_0+a_1*SES_1+a_2*Age+a_3*SES_2$
- Does a "cumulative" indicator, such as a sum of times in life "disadvantaged" predict health?
  - $E(health_2)=a_0+a_1*(SES_1+SES_2+SES_3)$

When The Classic Approach Works

- There are no unmeasured confounders of childhood SES and no unmeasured confounders of adult SES
- When the sample is not conditioned (restricted, weighted) on any factor affected by childhood SES, adult SES (or, it follows, adult health).
- When childhood and adult SES are measured perfectly
- Linear models
- Adult SES doesn't modify the effect of childhood SES
- There are no confounders of adult SES that are affected by childhood SES

Difficult to Assess Empirically

- There are no unmeasured confounders of childhood SES and no unmeasured confounders of adult SES
  - Impossible to prove in observational data.

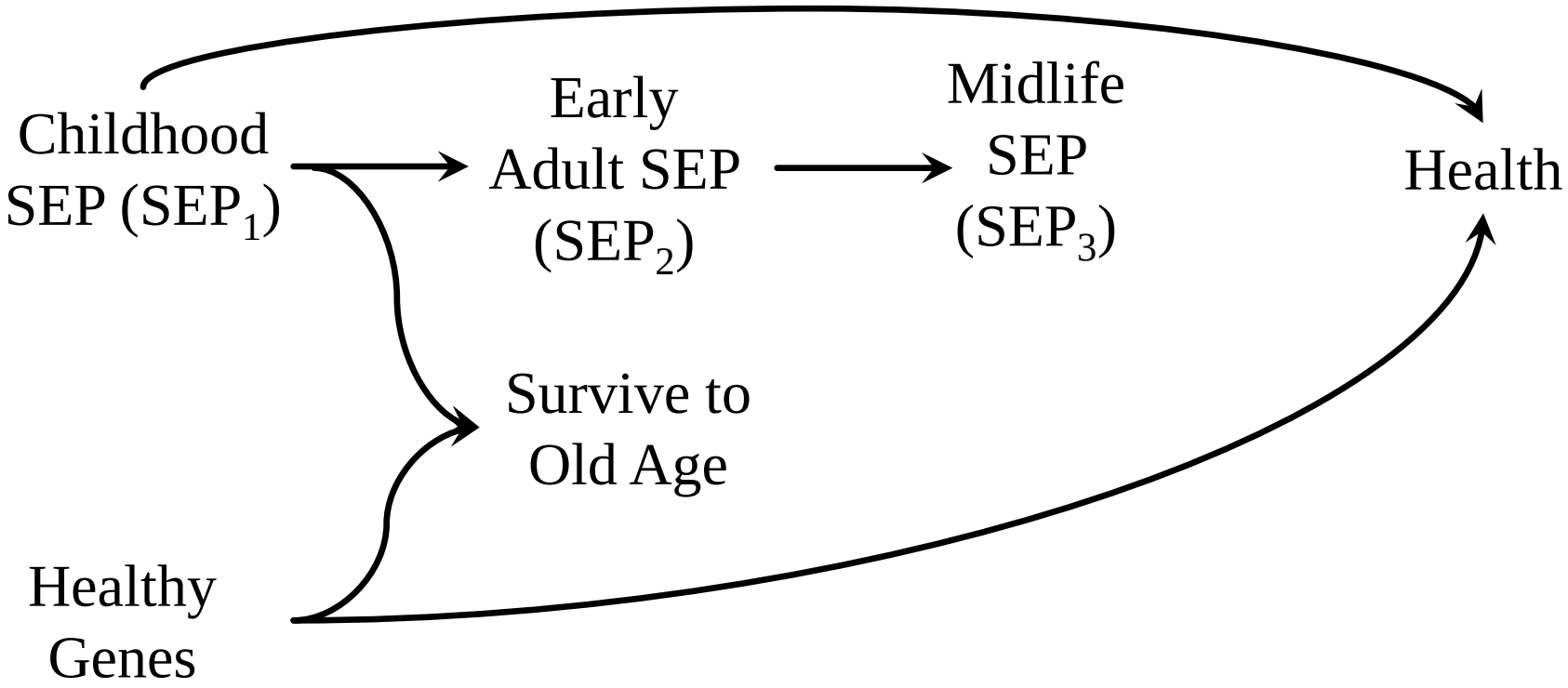
86

When The Classic Approach Works

- 2. When the sample is not conditioned (restricted, weighted) on any factor affected by childhood SES, adult SES (or, it follows, adult health).
  - Most studies start in adulthood (or old age)
  - Who survives to adulthood?

Critical Period/Latency & Survivor Bias

Exposure in one period makes all the difference: subsequent exposures do not influence health.



Who survived to be old?

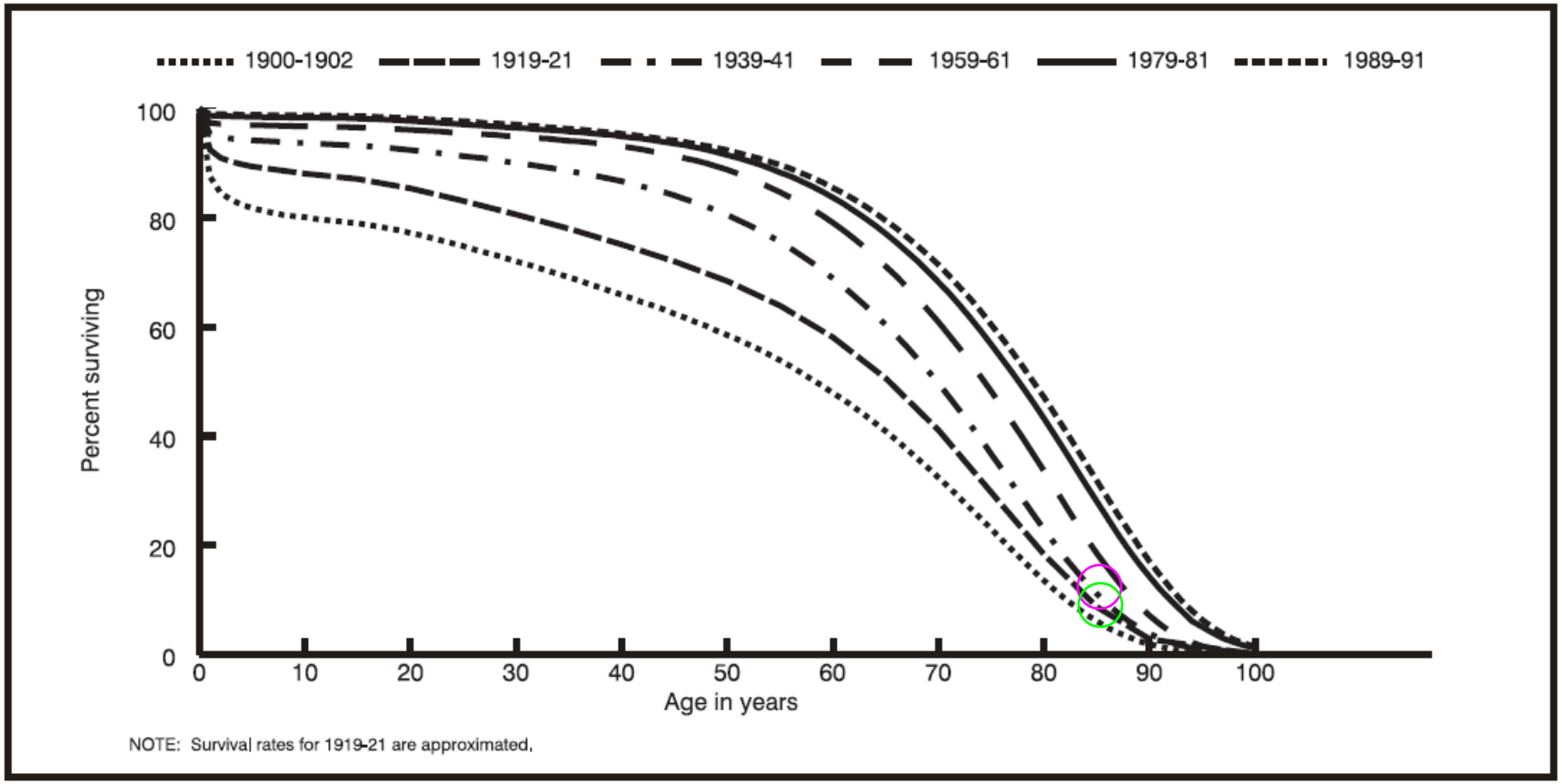


Figure 2. Percent survival by age: Death-registration States, 1900–1902 to 1919–21, and United States, 1939–41 to 1989–91

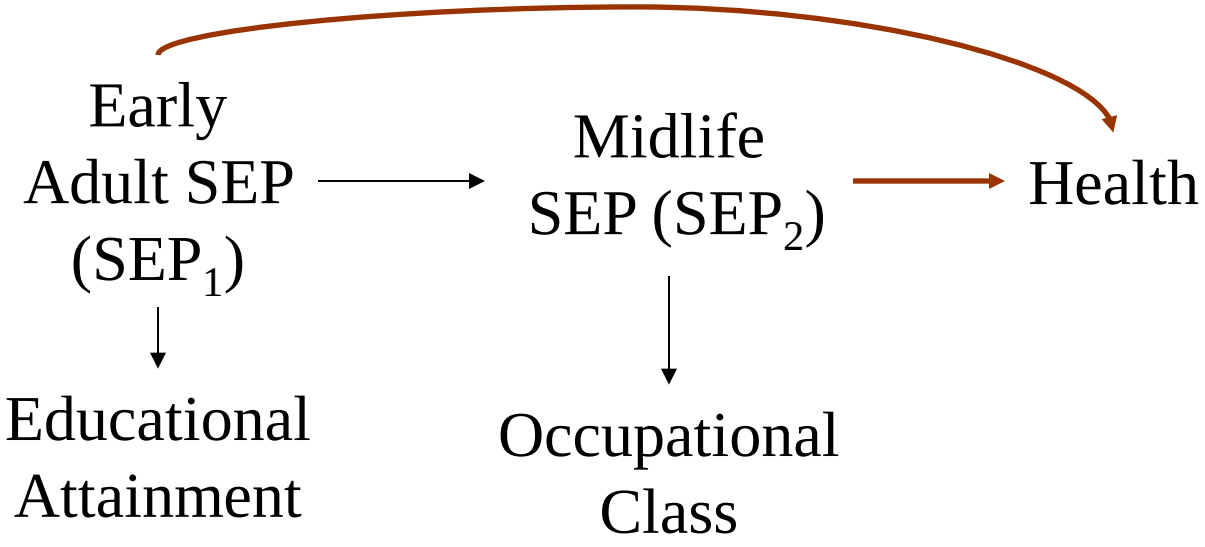
Survivor Bias in Lifecourse Studies

- If the study enrolls adults, it is by design conditioned on factors that influence survival to adulthood.
- Early childhood adversity is very likely to influence survival
- Survivor Bias is not just a generalizability problem: it can obscure (or inflate) causal effects
- Very similar issue if you are examining a patient population: by definition they are selected into having the disease. Relevant for obesity paradox.

Problems with Conventional Approach: Measurement Error

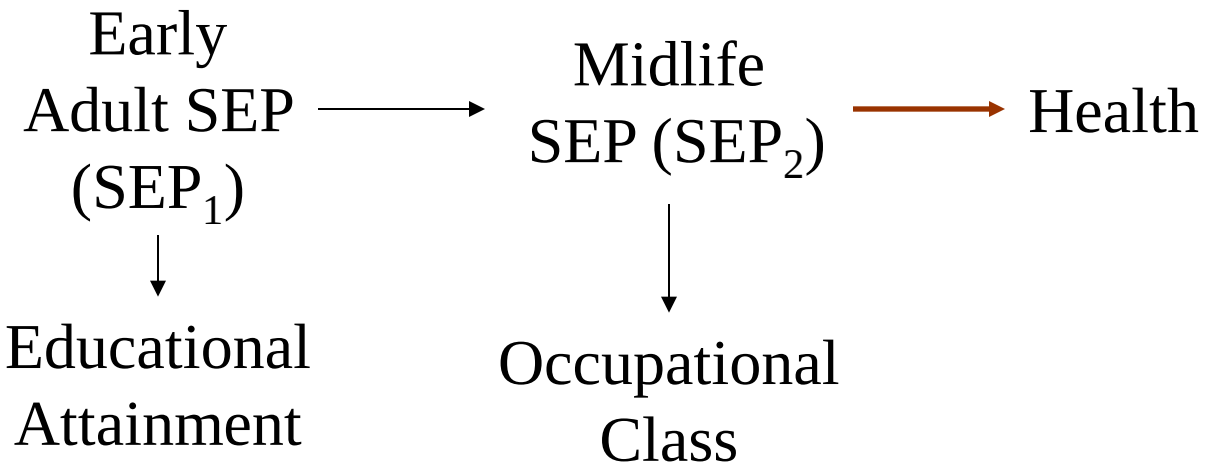
- 3. When childhood and adult SES are measured perfectly: but what is the causal factor?
  - What level? Individual? Family? Neighborhood? State? Region?
  - What time period? Infancy? Early childhood? Age 13?
  - What domain? Education? Income? Status? Psychological traumas?
  - How to measure? Retrospective? Linking outside data sets?
  - The effect of adjusting for a mediator attenuates as the square of the correlation between the measured value and the true value of the mediator.

Distinguishing Etiologic Models: Testing for Latency or Cumulative vs Immediate Risk



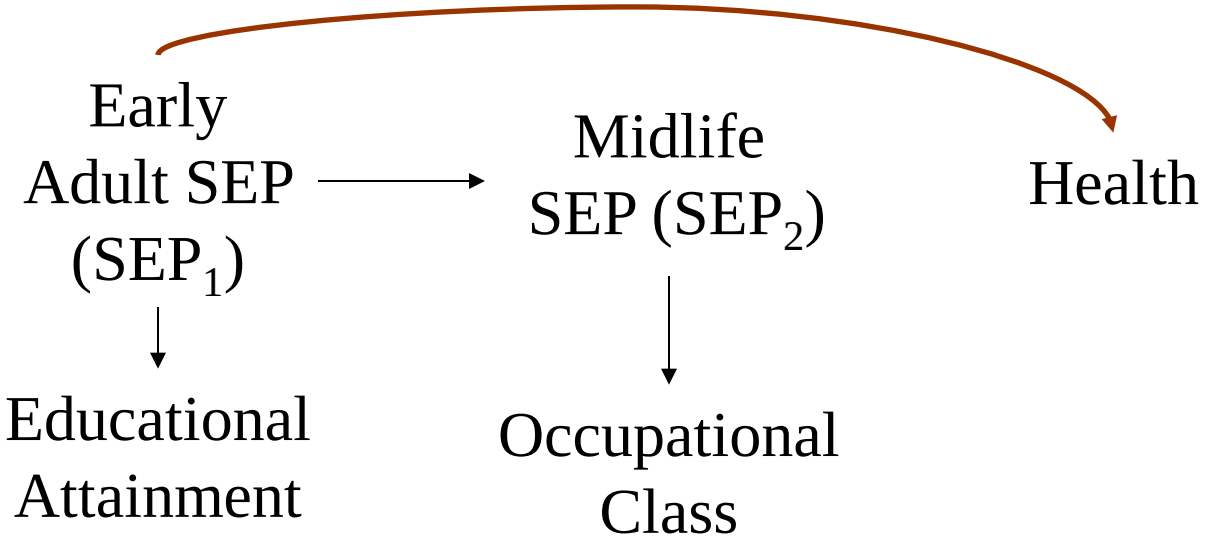
To test for a direct effect of SEP_1 on Health, we would compare the regression of Health on Education with and without adjustment for occupational class.

Distinguishing Etiologic Models: Testing for Latency or Cumulative vs Immediate Risk



Even if there were **no direct effect** of SEP_1 on health, educational attainment would predict health independently of adult occupational class, due to measurement error.

Distinguishing Etiologic Models: Testing for Latency or Cumulative vs Immediate Risk



Even if there were **no effect of SEP_2 on health**, occupational class would predict health independently of education, because occupational class is a correlate of the latent variable, SEP_1.

94

Alternative Sources of Evidence

- Some types of randomization
- Natural experiments/policy changes influencing childhood or adult SES
- Biological studies/animal models
- Better regression models
  - Better measurement of SES
  - Better measurement of confounders
  - Modeling modification of direct pathways by mediators
  - Weighting to deal with measured confounders
  - of adult SES that are mediators of child SES effects

Alternative Sources of Evidence

- Some types of randomization
  - A few formal trials that should be compared to observational studies to see how many of the "potential" biases are serious problems.
- Moving To Opportunity
  - Subgroup analyses to understand mechanisms
  - Analyses by age of randomization
- Romanian Orphans' Study
  - Evidence for differential specificity of timing for specific cognitive/behavioral outcomes

Alternative Sources of Evidence

- Natural experiments/policy changes influencing childhood or adult SES
  - Migration studies: where you're born strongly influences lifecourse trajectory (research limited by data availability)
- Instrumental variables
  - Policy changes: school policies, health care access policies, income support policies, racial segregation policies (research limited by data availability)
  - Institutional quirks: age of enrollment cutoff for primary school
  - Mendelian Randomization studies: genotypes control phenotypes, use genotypes as natural experiments (not clear what phenotypes of interest are genetically controlled)

Alternative Sources of Evidence

- Better regression models
  - Better measurement of exposures
  - Better measurement of confounders
  - Subgroup assessments with improved quality
  - Modeling modification of direct pathways by mediators

Alternative Sources of Evidence

- Better measurement of exposures:
  - Improvements in ideas, data, and analysis
  - important to recognize when we are looking under the lamppost and try to move out.

Old notes on interactions

Basic logistic model

- If you test interaction terms, you are testing deviations from multiplicative relationships
- This issue comes up again when you consider marginal versus conditional models
  - What happens to MY risk if I'm exposed?
  - What happens to the population's risk if the population is exposed?

Effect Measure Modification

- Effect measure modification: is the effect measure the same in different strata?
  - Is the relative risk of stroke associated with obesity the same in men and women? Is it the same in blacks and whites?
  - Is the risk difference for stroke associated with obesity the same in men and women? Is it the same in blacks and whites?
- If there is a main effect (e.g. obesity predicts stroke and race also predicts stroke), usually no more than one effect measure will show *no* modification.
  - If the RRs are the same for blacks and whites, but blacks have a higher overall rate of stroke than whites, the RDs *must be different.*

Effect Modification vs Interaction

- If the RRs are the same for men and women, but men have a higher overall rate of stroke than women, the RDs *must be different.*
- Suppose:
  - Obesity is associated with an RR of 1.5 in both men and women.
  - Non-obese women have an absolute risk of stroke of 2 per 1,000, so obese women have a risk of 3/1000
  - Non-obese men have an absolute risk of stroke of 4 per 1,000, so obese men have a risk of 6/1000.
  - RD for women=1/1000. RD for men=2/1000.

Scale Dependence of Effect Modification

- If there is a main effect (e.g. obesity predicts stroke and race also predicts stroke), usually no more than one effect measure (e.g. RR, OR, RD) will show *no* modification.
- If one of the exposures doesn't predict at all, then there shouldn't be effect measure modification

$$E(Y|X=x,M=1)-E(Y|X=x',M=1) \neq E(Y|X=x,M=0)-E(Y|X=x',M=0)$$

Implies:

$$E(Y|X=x,M=1) \neq E(Y|X=x,M=0)$$

Which means:

$$E(Y|M=1) \neq E(Y|M=0)$$

Scale Dependence of Effect Modification

- If there is a main effect (e.g. obesity predicts stroke and race also predicts stroke), usually no more than one effect measure (e.g. RR, OR, RD) will show *no* modification.
- If one of the exposures doesn't predict at all, then there shouldn't be effect measure modification
- There could be effect measure modification on many scales (e.g. subgroup differences in both additive effects and also multiplicative effects).

Effect Modification vs Interaction

- The scale dependence of interaction is essential to recognize when interpreting disparities, and especially trends in disparities.
- Relative effect measures will tend to be closer to the null in groups with higher baseline risk.
  - RRs in young routinely more extreme than RRs in old.
  - RRs tend to become more extreme over time if there is progress in reducing the population risk of a disease
- Which is more important: relative or absolute changes?

Relative vs Absolute Measures

Some claims that depend on comparing risk factor coefficients:

- The black-white disparity in infant mortality changed little during the 20th century, despite the advances in civil rights.
- Although the medical system may not work ideally for younger African-Americans, these problems are ameliorated among the elderly.
- After age 70, the role of socioeconomic factors in determining morbidity and mortality declines, and genetic risks predominate.
- From 1968-1996, Alabama made greater strides in reducing stroke mortality than Oregon.

Some Rules

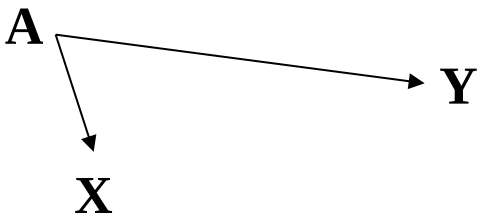
- Don't drop the lower order variables (the variables used to create the interaction tern) when you have an interaction term in the model.
- Don't test all possible interactions in your data set: use theory guiding what you'll test.
  - You'll get tired
  - You'll definitely find something (unless it's your dissertation, in which case you will defy the laws of probability and never find a $p<.05$ even after examining 5 million possible interactions).
- But remember: theory is only as good as … well, it's often not very good.

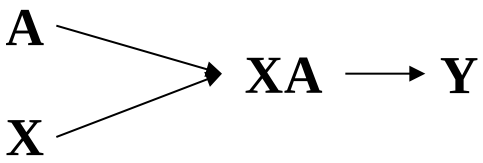
Some Rules

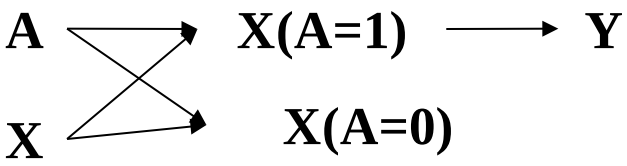
- It's easy to mess up the interpretation
- Categorizing variables simplifies.
  - Start with stratification
  - Pool and examine dichotomized versions for one or both of your variables unless you really have a strong reason not too.
  - Steve says: always categorize one of your variables if you possibly can
  - I say: always start by looking at categorized versions, and then consider which story you want to tell.
- A plot can go a very long way to understanding what is happening for yourself and readers.
- Remember - interactions go both ways

Common Questions

What's the difference between evaluating confounding and effect modification?







Both will result in a change in the estimated association between X and Y when conditioning on a specific value of A.

Common Questions

- Do two variables have to be correlated to interact?

NO

Common Questions

- Does Z have to be associated with X to confound the relation between X and Y?

Yes

Common Questions

- Which do I test for first, confounding or interactions?
- Think about what we need to adjust (i.e., control) for and then look at interactions.
- If you have a very firm theory and need to stratify than you do the first step in the two separate (stratified) models.