Analyses with Clustered Data and Lifecourse Models

Course outline

1. Causal inference from observational data and threats to validity.
   - ✓ Threats to validity: integrating perspectives
   - ✓ Data sources: finding what's available, and evaluating tradeoffs
   - ✓ Sampling: basic
2. Analyses with clustered Data:
   - ✓ Cluster randomized trials
   - ✓ Neighborhoods as clusters
3. Interactions
   - Review of interaction, effect modification, heterogeneous treatment effects
   - Individuals as clusters (longitudinal data analysis)
4. Evaluating Uncertainty
   - p-values and Confidence Intervals
   - Combining evidence
5. Mediation and Effect Decomposition
6. Power calculations and Publication bias
7. Evaluating theoretically motivated hypotheses

What Kinds of Questions Does Epidemiology Answer?

- **Description**
  - How frequent/common are risk factors/exposures/conditions/diseases?
- **Prediction**
  - Do A, B, and C predict presence or occurrence of Y? e.g., diagnosis or prognosis
- **Causation**
  - **Total effect**
    - What would changing some exposure – any biological, behavioral, environmental, social (etc.) factor – do to the incidence or prevalence of the health outcome in a population?
    - Will intervening upon X change occurrence of Y?
  - **Public health importance of effect/ attribution**
    - What fraction or how many cases of disease Y can be eliminated if a causal exposure X is eliminated or reduced? This reflects both the effect of X and the prevalence of X.
  - **Mediation**
    - Understanding the mechanisms of causation
    - How does X cause Y?
- **Interaction- which spans prediction and causation**
  - When and for whom does X cause Y? Are the effects different for different kinds of people or in different settings?

Outline

- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Interactions
- Options for age effects: cross-sectional, GEE, or growth curves
- Lifecourse timing

To understand growth curves, you must understand interactions

Effect (measure) modification

- Effect of an exposure on an outcome differs by levels of a third variable
- Decide which is of interest: marginal effect in whole population or conditional effect in a subgroup?
- What is the effect of a low phenylalanine diet in:
  - People who are homozygous for PKU?
  - Everyone else?
- What is the effect of obesity on SBP for
  - Men
  - Women?

Effect modification in linear models

- Typically this type of question is answered by
  - First, stratify and examine the association within subgroups.
  - Second, pool the data and test an interaction term:

$$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$$

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

b_0=average SBP in non-obese females

Because when

Male=0

Obese=0 then

$E(SBP|non\text{-}obese\ female)=b_0+b_1*0+b_2*0+b_3*0*0$

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

b_1=the difference in SBP for non-obese males vs non-obese females

Because when

Male=1

Obese=0 then

$E(SBP|\ non\text{-}obese\ male)=b_0+b_1*1+b_2*0+b_3*1*0= b_0+b_1$

And if male=0 and obese=0 then

$E(SBP|\ non\text{-}obese\ female)=b_0+b_1*0+b_2*0+b_3*0*0=b_0$

And the difference is: $b_0+b_1 -b_0 = b_1$

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

b_2=the difference in SBP for obese females vs non-obese females

Because when

Male=0

Obese=1 then

$E(SBP|\ obese\ female)=b_0+b_1*0+b_2*1+b_3*0*1= b_0+b_2$

And if male=0 and obese=0 then

$E(SBP|\ non\text{-}obese\ female)=b_0+b_1*0+b_2*0+b_3*0*0=b_0$

And the difference is: $b_0+b_2 -b_0 = b_2$

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

b_3=the extra difference in SBP for obese vs non-obese males compared to obese vs non-obese females

Because the difference in SBP for obese vs non-obese males is:

$E(SBP|obese\ male)=b_0+b_1*1+b_2*1+b_3*1*1= b_0+b_1+b_2+b_3$

$-\ E(SBP|non\text{-}obese\ male)=b_0+b_1*1+b_2*0+b_3*1*0= b_0+b_1$

$=b_2+b_3$

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

b_3=the extra difference in SBP for obese vs non-obese males compared to obese vs non-obese females

And the difference in SBP for obese vs non-obese females is:

$E(SBP|obese\ female)=b_0+b_1*0+b_2*1+b_3*0*1= b_0+b_2$

$-E(SBP|non\text{-}obese\ female)=b_0+b_1*0+b_2*0+b_3*0*0= b_0$

$=b_2$

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

b_3=the extra difference in SBP for obese vs non-obese males compared to obese vs non-obese females

Effect of obesity in men $= b_2+b_3$

Effect of obesity in women $= b_2$

The *extra* effect of obesity among men.

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

Or:

b_3=the extra difference in SBP for obese males vs obese females compared to non-obese males vs non-obese females

$E(SBP|obese\ male)=b_0+b_1*1+b_2*1+b_3*1*1= b_0+b_1+b_2+b_3$

$E(SBP|non\text{-}obese\ male)=b_0+b_1*1+b_2*0+b_3*1*0= b_0+b_1$

$E(SBP|obese\ female)=b_0+b_1*0+b_2*1+b_3*0*1= b_0+b_2$

$E(SBP|non\text{-}obese\ female)=b_0+b_1*0+b_2*0+b_3*0*0= b_0$

Effect of being male vs fem among obese= $b_0+b_1+b_2+b_3-(b_0+b_2)=b_1+b_3$

Effect of being male vs fem | non-obese= $b_0+b_1-b_0)=b_1$

The *extra* effect of being male among obese.

Effect modification in linear models

$E(SBP)=b_0+b_1*male+b_2*obese+b_3*male*obese$

Is a saturated model. This is sometimes called contrast coding. You can also specify the regression model as:

$E(SBP)=a_0*non\text{-}obese*female+a_1*obese*male$

$+a_2*obese*female+a_3*non\text{-}obese*male$

The predicted values for each group will be the same as with contrast coding. In other words:

$a_0=b_0$

$a_1=b_0+b_1+b_2+b_3$

$a_2=b_0+b_2$

$a_3=b_0+b_1$

Simulation example

```
. clear

. set seed 0707

. set obs 1000
number of observations (_N) was 0, now 1,000

. gen id=_n

. gen male=runiform()<.4

. gen obese=runiform()<.2

. gen SBP_e=rnormal()*10

. gen SBP=120+SBP_e+5*male+10*obese+7*obese*male

. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	1,000	500.5	288.8194	1	1000
male	1,000	.4	.4901431	0	1
obese	1,000	.218	.413094	0	1
SBP_e	1,000	-.2171756	9.790366	-31.87436	36.02893
SBP	1,000	124.5508	11.76823	89.53706	166.0289

What is the true effect of obesity on SBP?

What is the estimated effect of obesity if we ignore gender?

```
. reg SBP obese
```

Source	SS	df	MS	Number of obs	=	1,000
				F(1, 998)	=	303.11
Model	32230.7574	1	32230.7574	Prob > F	=	0.0000
Residual	106122.073	998	106.334743	R-squared	=	0.2330
				Adj R-squared	=	0.2322
Total	138352.831	999	138.491322	Root MSE	=	10.312

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	13.75003	.7897797	17.41	0.000	12.20021	15.29985
_cons	121.5533	.3687519	329.63	0.000	120.8297	122.2769

```
. reg SBP obese male
```

Source	SS	df	MS	Number of obs	=	1,000
				F(2, 997)	=	215.05
Model	41696.5779	2	20848.289	Prob > F	=	0.0000
Residual	96656.2527	997	96.9470939	R-squared	=	0.3014
				Adj R-squared	=	0.3000
Total	138352.831	999	138.491322	Root MSE	=	9.8462

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	13.86793	.7542063	18.39	0.000	12.38792	15.34794
male	6.280987	.6356472	9.88	0.000	5.033627	7.528346
_cons	119.0152	.4358329	273.08	0.000	118.16	119.8705

Why doesn't controlling for male change the effect estimate for obesity (by much)?

What is the estimated effect of obesity stratified by gender?

```
. reg SBP obese if male==0

      Source |       SS           df       MS      Number of obs   =       600
-------------+----------------------------------   F(1, 598)       =    152.99
       Model |  14306.8175         1  14306.8175   Prob > F        =    0.0000
    Residual |  55920.9495       598  93.5132935   R-squared       =    0.2037
-------------+----------------------------------   Adj R-squared   =    0.2024
       Total |   70227.767       599  117.241681   Root MSE        =    9.6702

------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       obese |     11.7247   .9479096    12.37   0.000     9.863065    13.58634
       _cons |     119.4939   .4479646   266.75   0.000     118.6141    120.3737
------------------------------------------------------------------------------

. reg SBP obese if male==1

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(1, 398)       =    198.44
       Model |  19698.6234         1  19698.6234   Prob > F        =    0.0000
    Residual |  39507.5117       398  99.2651049   R-squared       =    0.3327
-------------+----------------------------------   Adj R-squared   =    0.3310
       Total |  59206.1351       399  148.386304   Root MSE        =    9.9632

------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       obese |     17.22919   1.223053    14.09   0.000     14.82473    19.63364
       _cons |     124.5903   .5604731   222.29   0.000     123.4885    125.6922
------------------------------------------------------------------------------
```

What is the estimated effect of obesity stratified by gender?

```
. reg SBP obese if male==0
```

Source	SS	df	MS		
Model	14306.8175	1	14306.8175	Number of obs	= 600
Residual	55920.9495	598	93.5132935	F(1, 598)	= 152.99
				Prob > F	= 0.0000
				R-squared	= 0.2037
				Adj R-squared	= 0.2024
Total	70227.767	599	117.241681	Root MSE	= 9.6702

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	11.7247	.9479096	12.37	0.000	9.863065	13.58634
_cons	119.4939	.4479646	266.75	0.000	118.6141	120.3737

```
. reg SBP obese if male==1
```

Source	SS	df	MS		
Model	19698.6234	1	19698.6234	Number of obs	= 400
Residual	39507.5117	398	99.2651049	F(1, 398)	= 198.44
				Prob > F	= 0.0000
				R-squared	= 0.3327
				Adj R-squared	= 0.3310
Total	59206.1351	399	148.386304	Root MSE	= 9.9632

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	17.22919	1.223053	14.09	0.000	14.82473	19.63364
_cons	124.5903	.5604731	222.29	0.000	123.4885	125.6922

Hey, why is the effect estimate among women nearly 12!?? Isn't it supposed to be 10? (check the data generating rule).

Maybe it's a fluke? How could I tell?

What is the model with an interaction?

```
. gen ObeseMale=obese*male

. reg SBP obese male ObeseMale
```

Source	SS	df	MS		
Model	42924.3693	3	14308.1231	Prob > F	0.0000
Residual	95428.4613	996	95.8117081	R-squared	0.3103
				Adj R-squared	0.3082
Total	138352.831	999	138.491322	Root MSE	9.7883

Number of obs	=	1,000
F(3, 996)	=	149.34

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	11.7247	.959488	12.22	0.000	9.841851	13.60755
male	5.096468	.7133065	7.14	0.000	3.696713	6.496224
ObeseMale	5.504484	1.537672	3.58	0.000	2.487036	8.521932
_cons	119.4939	.4534364	263.53	0.000	118.6041	120.3837

What if I switch the reference category?

```
. gen NotObese=1-obese

. gen female=1-male

. gen NotObeseFemale=NotObese*female

. tab obese NotObese
```

obese	NotObese 0	1	Total
0	0	782	782
1	218	0	218
Total	218	782	1,000

```
. tab male female
```

male	female 0	1	Total
0	0	600	600
1	400	0	400
Total	400	600	1,000

```
. reg SBP NotObese female NotObeseFemale
```

Source	SS	df	MS		
Model	42924.3693	3	14308.1231	Number of obs	= 1,000
Residual	95428.4613	996	95.8117081	F(3, 996)	= 149.34
				Prob > F	= 0.0000
				R-squared	= 0.3103
				Adj R-squared	= 0.3082
Total	138352.831	999	138.491322	Root MSE	= 9.7883

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NotObese	-17.22919	1.201589	-14.34	0.000	-19.58712	-14.87125
female	-10.60095	1.362214	-7.78	0.000	-13.27409	-7.927813
NotObeseFemale	5.504484	1.537672	3.58	0.000	2.487036	8.521932
_cons	141.8195	1.067996	132.79	0.000	139.7237	143.9153

```
.
```

What if I switch the reference category?

```
. reg SBP obese male ObeseMale
```

Source	SS	df	MS		
Model	42924.3693	3	14308.1231	Number of obs	= 1,000
Residual	95428.4613	996	95.8117081	F(3, 996)	= 149.34
				Prob > F	= 0.0000
				R-squared	= 0.3103
				Adj R-squared	= 0.3082
Total	138352.831	999	138.491322	Root MSE	= 9.7883

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	11.7247	.959488	12.22	0.000	9.841851	13.60755
male	5.096468	.7133065	7.14	0.000	3.696713	6.496224
ObeseMale	5.504484	1.537672	3.58	0.000	2.487036	8.521932
_cons	119.4939	.4534364	263.53	0.000	118.6041	120.3837

Predicted value for:

1. Obese male?
2. Non obese male?
3. Obese female?
4. Non obese female?

Predicted value for:

1. Obese male?
2. Non obese male?
3. Obese female?
4. Non obese female?

```
. reg SBP NotObese female NotObeseFemale
```

Source	SS	df	MS		
Model	42924.3693	3	14308.1231	Number of obs	= 1,000
Residual	95428.4613	996	95.8117081	F(3, 996)	= 149.34
				Prob > F	= 0.0000
				R-squared	= 0.3103
				Adj R-squared	= 0.3082
Total	138352.831	999	138.491322	Root MSE	= 9.7883

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NotObese	-17.22919	1.201589	-14.34	0.000	-19.58712	-14.87125
female	-10.60095	1.362214	-7.78	0.000	-13.27409	-7.927813
NotObeseFemale	5.504484	1.537672	3.58	0.000	2.487036	8.521932
_cons	141.8195	1.067996	132.79	0.000	139.7237	143.9153

What if I switch the reference category?

```
. reg SBP obese male ObeseMale
```

Source	SS	df	MS		
Model	42924.3693	3	14308.1231	Number of obs	= 1,000
Residual	95428.4613	996	95.8117081	F(3, 996)	= 149.34
				Prob > F	= 0.0000
				R-squared	= 0.3103
				Adj R-squared	= 0.3082
Total	138352.831	999	138.491322	Root MSE	= 9.7883

SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
obese	11.7247	.959488	12.22	0.000	9.841851	13.60755
male	5.096468	.7133065	7.14	0.000	3.696713	6.496224
ObeseMale	5.504484	1.537672	3.58	0.000	2.487036	8.521932
_cons	119.4939	.4534364	263.53	0.000	118.6041	120.3837

Predicted value for:

1. Obese male? 141.8
2. Non obese male? 124.6
3. Obese female? 131.2
4. Non obese female? 119.5

Predicted value for:

1. Obese male? 141.8
2. Non obese male? 124.6
3. Obese female? 131.2
4. Non obese female? 119.5

```
. reg SBP NotObese female NotObeseFemale
```

Source	SS	df	MS		
Model	42924.3693	3	14308.1231	Number of obs	= 1,000
Residual	95428.4613	996	95.8117081	F(3, 996)	= 149.34
				Prob > F	= 0.0000
				R-squared	= 0.3103
				Adj R-squared	= 0.3082
Total	138352.831	999	138.491322	Root MSE	= 9.7883

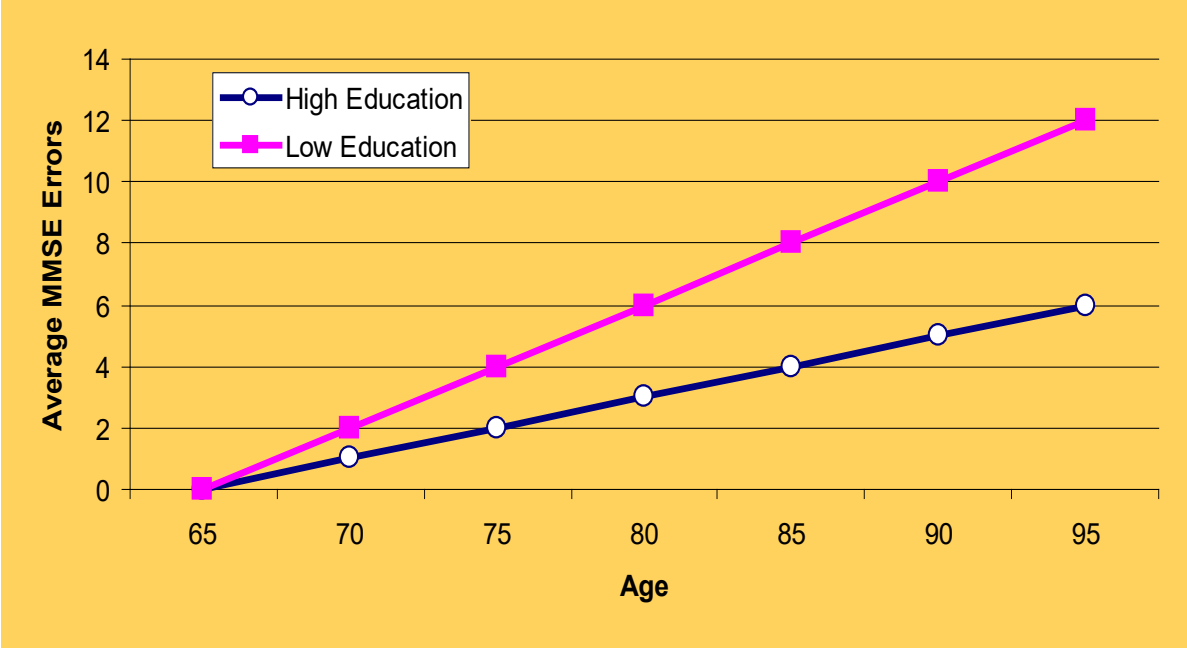
SBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NotObese	-17.22919	1.201589	-14.34	0.000	-19.58712	-14.87125
female	-10.60095	1.362214	-7.78	0.000	-13.27409	-7.927813
NotObeseFemale	5.504484	1.537672	3.58	0.000	2.487036	8.521932
_cons	141.8195	1.067996	132.79	0.000	139.7237	143.9153

Effect modification in linear models with a continuous variable

In linear models, an interaction between a binary and a continuous variable implies the slope of the line relating the continuous variable to the outcome is different depending on the value of the binary variable

"different slopes for different folks"

$$E(MMSE)=b_0+b_1*HighEd+b_2*Age+b_3*HighEd*Age$$



Effect modification in linear models with a continuous variable

$$E(\text{CognitiveScore}) = b_0 + b_1 * \text{Male} + b_2 * \text{Age} + b_3 * \text{Male} * \text{Age}$$

$b_0 = 25$

$b_1 = -3$

$b_2 = -.1$

$b_3 = .05$

What is the effect of 1 year of age among women?
What is the effect of 5 years of age among women?
What is the effect of 1 year of age among men?
What is the effect of 5 years of age among men?

Effect modification in linear models with a continuous variable

$$E(\text{CognitiveScore}) = b_0 + b_1 * \text{Male} + b_2 * \text{Age} + b_3 * \text{Male} * \text{Age}$$

$b_0 = 25$

$b_1 = -3$

$b_2 = -.1$

$b_3 = .05$

What is the effect of 1 year of age among women?	-0.1
What is the effect of 5 years of age among women?	5*-0.1=-0.5
What is the effect of 1 year of age among men?	-.1+.05=-.05
What is the effect of 5 years of age among men?	5*-.05=-.25

Coding reminder

Always recode dichotomous variables to be 0 (NO) or 1 (YES) and give them a meaningful name (like "male", not "sex"; or "white", not "race")

Note that the mean of a binary variable coded as 0/1 equals the prevalence of the "YES" category, so if you look at the mean of 'male', it should tell you the percent of the sample that is male.

Logistic Regression ?

Probabilities vs Odds

- How does the odds relate to the probability?
- P=.01, O=.01/.99=.010
- P=.05, O=.05/.95=.053
- P=.10, O=.10/.90=.11
- P=.25, O=.25/.75=.33
- P=.50, O=.50/.50=1

- Very similar for rare events

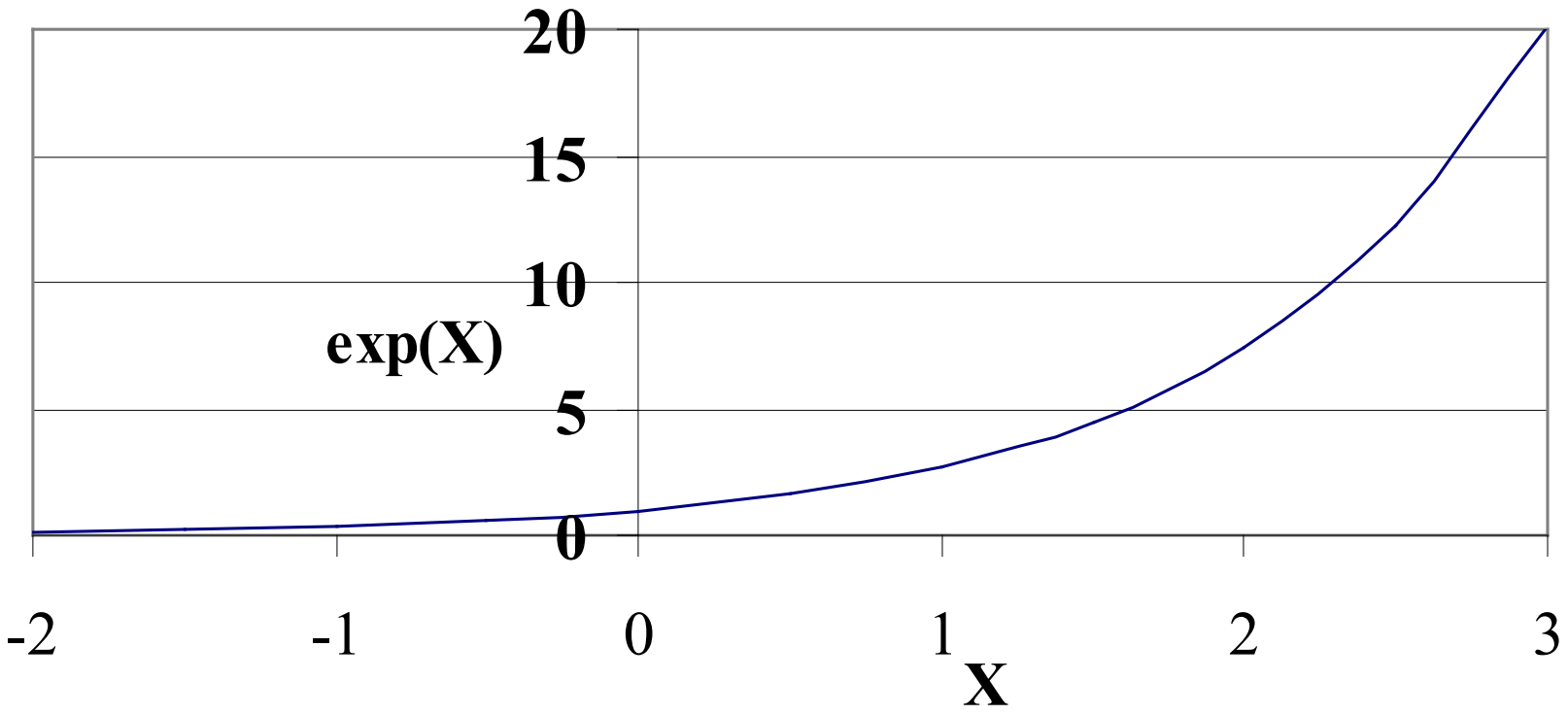
Probabilities vs Odds

- If you know the probability, you can calculate the odds, and vice versa
- If you know the probability, you can calculate the logit=ln(p/(1-p)
- e~=2.7
- e^2=7.4
- $e^{0.5}$=1.65
- ln(k)=what you exponentiate to get k
- ln(1.65)=0.5
- ln(7.40)=2
- ln(2.70)=1

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Plot of e



Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Probability ranges from 0 to 1

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

$E(Cognitive\_Impairment)=b_0+b_1*Male+b_2*Age+b_3*Male*Age$

$b_0=.15$ $b_1=-.3$ $b_2=.1$ $b_3=-.05$

What is the predicted probability of cognitive impairment for women age=0?

What is the predicted probability of cognitive impairment for women age=20?

Using probability as the dependent variable in a linear model can lead to out-of-possible-range predictions.

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^l/(1+e^l)$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Ln(odds) ranges from negative infinity to positive infinity

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
[illegible]	[illegible]	[illegible]	[illegible]	0.05
[illegible]	[illegible]	[illegible]	[illegible]	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Let's REGRESS!!

Ln odds ranges from – infinity to + infinity

Basic logistic model

- Logit(p)= log(p/(1-p)) =α+βX
- α and β are unknown parameters to be estimated from the data (e.g. with MLE).
- This model assumes that the logit increases linearly with X.
- If that is not true, the equation is mis-specified (or possibly mispecified)

- probability=odds/(1+odds)

Basic logistic model

- **Logit(p)= ln(p/(1-p)) =α+βX**
- p/(1-p)=exp(α+βX)
- p=(1-p)*exp(α+βX)
- p=exp(α+βX)-p*exp(α+βX)
- p+p*exp(α+βX)=exp(α+βX)
- p*(1+exp(α+βX))=exp(α+βX)
- p=exp(α+βX)/ (1+exp(α+βX))
- **$p=e^{l}/(1+e^{l})$**

Basic logistic model

- Logit(p)= ln(p/(1-p)) =$\alpha+\beta X$
- β is the ln of the odds ratio
- Exponentiate β to get the odds ratio
- If X is dichotomous:
  - OR=odds in the exposed/ odds in the unexposed

- If X is continuous:
  - OR=proportional increase in odds per one unit change in X.
  - Implies that odds increase by the same factor for each one unit change in X.
  - This implies the effect is multiplicative

Basic logistic model

- Logit(p)= ln(p/(1-p)) $=\alpha+\beta X$
- Logit(p)= ln(p/(1-p)) $=-3+.5*X$
  - OR=exp(.5)=1.65
- X=0,
  - logit=-3,
  - odds=exp(-3)=.05
  - p=exp(-3)/(1+exp(-3))=.05
- X=1
  - logit=-3+.5=-2.5
  - odds=exp(-2.5)=.082
  - .082/.05=1.64
  - p=exp(-2.5)/(1+exp(-2.5))=.075
  - .075/.05=1.5
- X=2
  - logit=-3+.5*2=-2
  - odds=exp(-2)=.135
  - .135/.05=2.7
  - p=exp(-2)/(1+exp(-2))=.119
  - .119/.05=2.38
- X=2 vs X=0
  - Exp(.5*2)=exp(1)=2.7

Basic logistic model

- Logit(p)= ln(p/(1-p)) $=\alpha+\beta X$
- Logit(p)= ln(p/(1-p)) $=-3+.5*X+1.2*Z$
  - $OR_x=\exp(.5)=1.65$
  - $OR_z=\exp(1.2)=3.32$
- X=0, Z=0
  - logit=-3,
  - odds=exp(-3)=.05
  - p=exp(-3)/(1+exp(-3))=.05
- X=1, Z=1
  - logit=-3+.5+1.2=-1.3
  - odds=exp(-1.3)=.27
  - .27/.05=5.5
  - Exp(.5+1.2)=exp(1.7)=5.5
  - 1.65*3.32=5.5
  - p=exp(-1.3)/(1+exp(-1.3))=.214
  - .214/.05=4.3

Basic logistic model

- Logit(p)= ln(p/(1-p)) =$\alpha+\beta X$
- This is a multiplicative model: the odds ratios multiply
- If someone is already really "high risk", a doubling of that risk (OR~2) is much greater in absolute terms than if someone is "low risk"
  - Risk of .01, doubled to .02, absolute change of .01
  - Risk of .2, doubled to .4, absolute change of .2
- If a population is very high risk, cutting that risk in half has much greater benefits in absolute terms (i.e. prevents more cases) than if a population is low risk.
- If you test interaction terms in logistic models, you are testing deviations from multiplicative relationships

How do interactions work in a logistic model?

- Logit(p)= ln(p/(1-p)) =−3+.5*X+1.2*Z+.2*X*Z
  - OR_x=exp(.5)=1.65 for Z=0
  - OR_z=exp(1.2)=3.32 for X=0
- X=0, Z=0
  - logit=-3,
  - odds=exp(-3)=.05
  - p=exp(-3)/(1+exp(-3))=.05
- X=1, Z=1
  - logit=-3+.5+1.2+0.2=-1.1
  - odds=exp(-1.1)=0.33
  - Odds for X=1, Z=1 / odds for X=0, Z=0= .33/.05=6.7
  - Could also calculate: Exp(.5+1.2+0.2)=exp(1.9)=6.7

Suppose Z is continuous

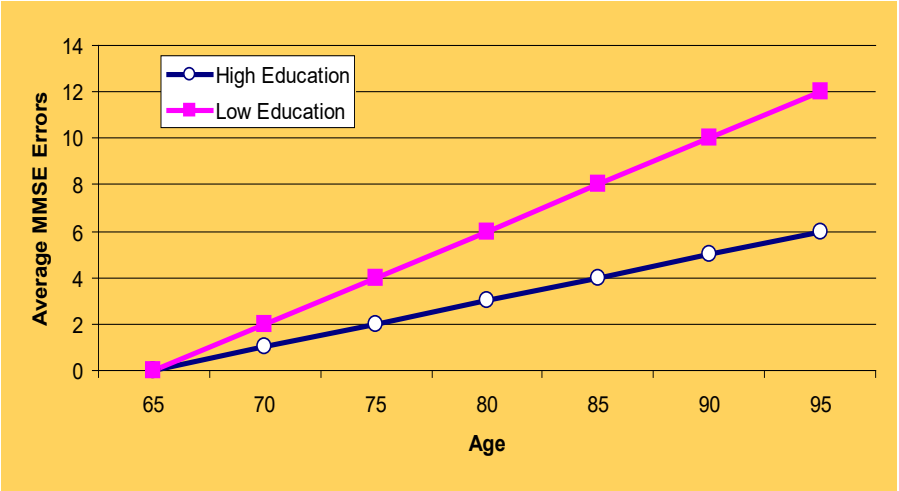
- Logit(p)= ln(p/(1-p)) =–3+.5*X+1.2*Z+.2*X*Z
  - OR_x=exp(.5)=1.65 for Z=0
  - OR_z=exp(1.2)=3.32 for X=0
- X=1, Z=0
  - logit=-3+.5
  - odds=exp(-2.5)=.082
  - p=exp(-2.5)/(1+exp(-2.5))=.076
- X=1, Z=1
  - logit=-3+.5+1.2+0.2=-1.1
  - odds=exp(-1.1)=0.33
  - Odds for X=1, Z=1 / odds for X=1, Z=0= .33/.082=4.02
  - Could also calculate: Exp(1.2+0.2)=exp(1.4)=4.06
- X=1, Z=2
  - logit=-3+.5+2*1.2+2*0.2=0.3
  - odds=exp(0.3)=1.35
  - Odds for X=1, Z=2 / odds for X=1, Z=0= 1.35/.082=16.46
  - Could also calculate: Exp(2*1.2+2*0.2)=exp(2.8)=16.4

Outline

- Sources of clustering
- Conceptual questions about clustered data
- Consequences of ignoring clustering
- Basic mixed model for clustered data without ordering
- Interactions
- Options for age effects: cross-sectional, GEE, or growth curves
- Lifecourse timing

Substantive questions about health are often about changes with age

- Do rates of cognitive decline differ between people with versus without Type 2 Diabetes?
- Does weight gain occur more rapidly in infants fed with breast milk versus formula?
  - Some longitudinal questions are about the time to an event: is the time to dementia diagnosis longer for people with versus without Type 2 Diabetes?
  - But we focus here on changes in a continuous or dimensional outcome.
  - i.e., comparing slopes.



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

Imagine complete data shows all birth cohorts average the same at age 30, but decrease by 10 units per decade of age

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

Imagine complete data shows all birth cohorts average the same at age 30, but decrease by 10 units per decade of age

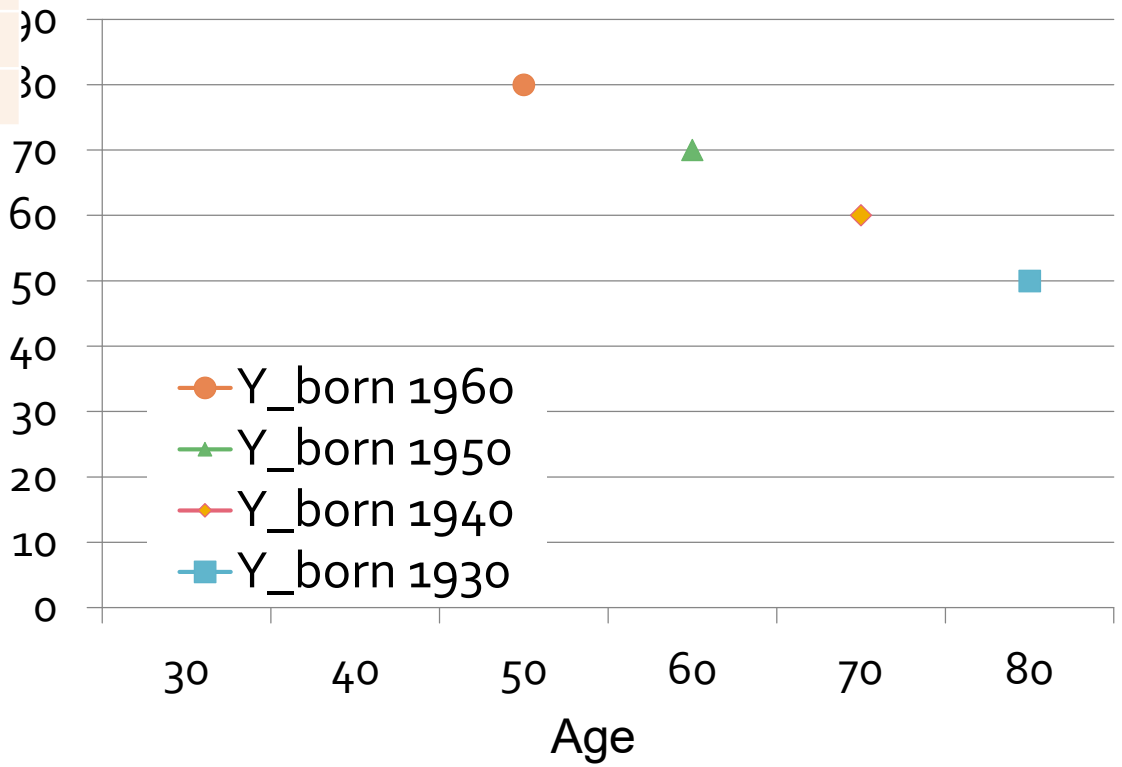
It's *possible* in special cases to estimate an age slope with cross-sectional data

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

In this case, you could get the right answer about the age trajectory from cross-sectional data.

- There are no period or cohort effects
- All cohorts have the same slope
- There is no selective mortality

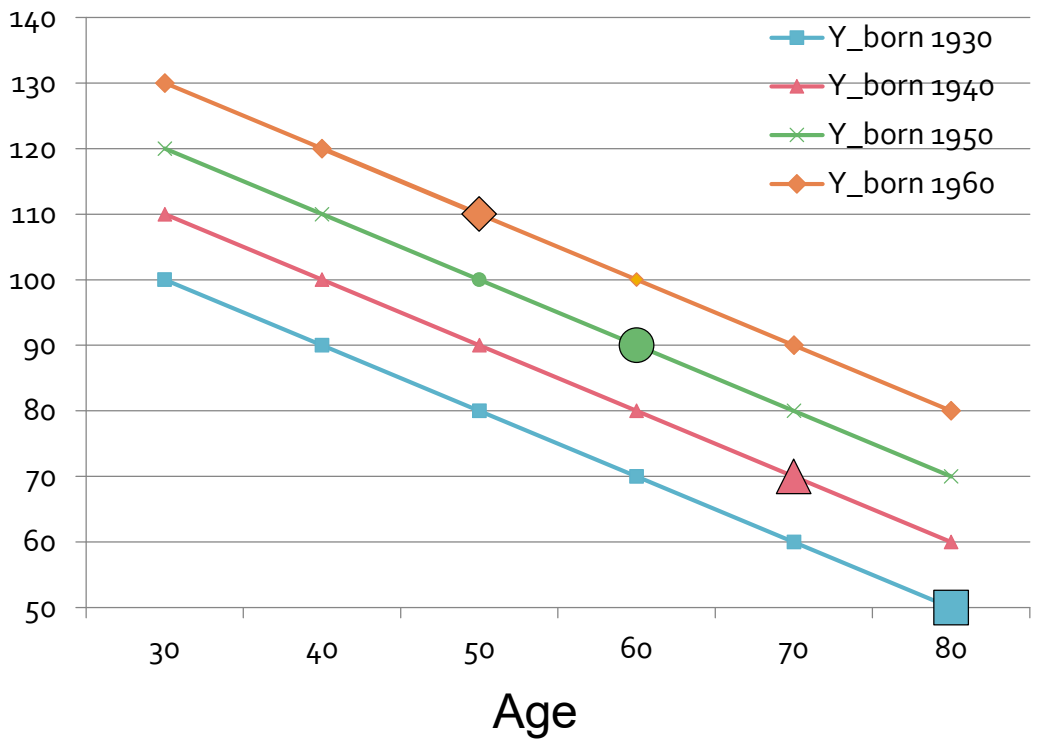
Imagine cross-sectional data collected in 2010



But the cross-sectional 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 (10 points lower for each year of age)

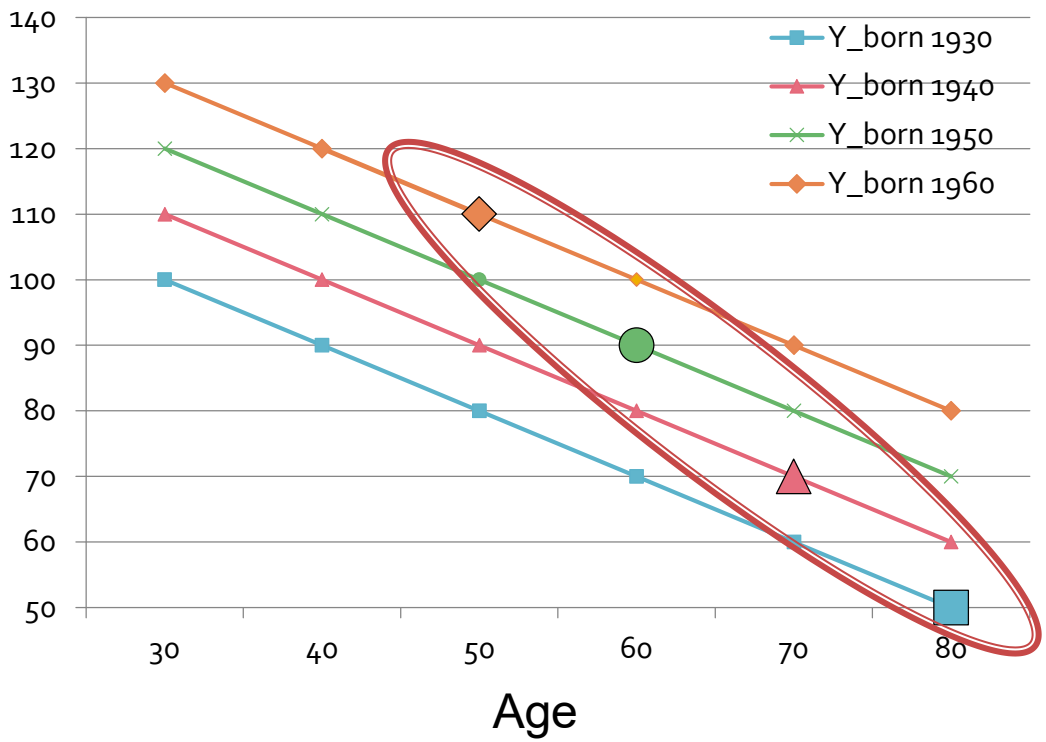


Regressing Y on age in a cross-sectional data set leads to a steeper age slope because it's combining age and cohort differences.

But the cross-sectional 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 (10 points lower for each year of age)

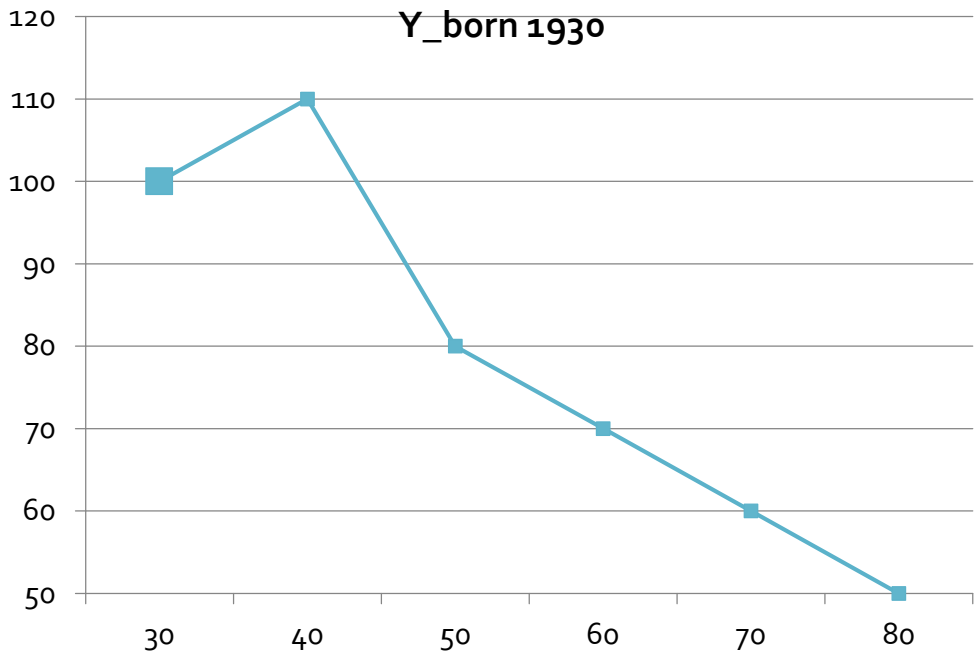


Regressing Y on age in a cross-sectional data set leads to a steeper age slope because cohort differences folded into age differences.

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, 10 point decline per decade of age, but large period effect (added 20 points in 1970)

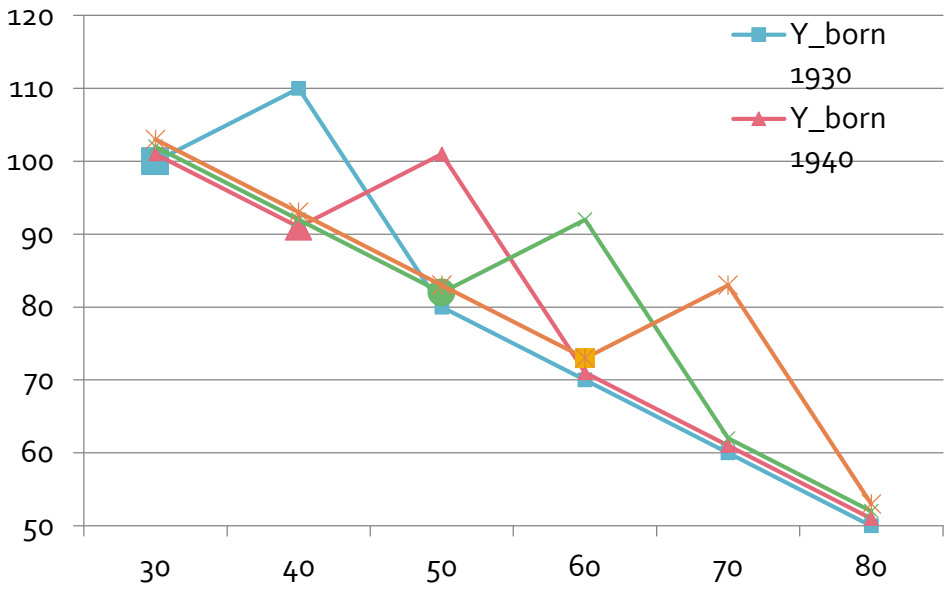


Fitting a straight line to age in a single birth cohort will lead to a much flatter age-slope estimate

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 for age

Longitudinal data preferred to estimate age/growth effects

- With longitudinal data, you can identify effects based on within-person change
- With cross-sectional data, all estimates of the effects of age are based on comparing between people
- Comparing within person in general will do a better job of accounting for potential confounders

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
  - Does nutritional supplementation improve growth of premature infants?

To get ready for mixed: reconceptualize the data as person-period

- Instead of one observation (row of data) per cluster, consider one observation at the lowest level of data available
- One row per year for each person
- Or one row per person for each neighborhood
- Or one row per year for each person in each neighborhood

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

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

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

Stata: reshape command
Or w/ arrays

- 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

Wide vs Tall Data sets: Same for people or places or items

- Wide

place	X_pers1	X_pers2	X_pers3
1	27	32	45
2	18	17	20
3	41	44	46

Stata: reshape command
Or w/ arrays

- Tall

place	person	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

Wide vs Tall Data sets: Same for people or places or items

- Wide

person	item1	item2	item3
1	27	32	45
2	18	17	20
3	41	44	46

Item 1= Difficulty getting out of bed
Item 2 = Difficulty feeding self
Item 3= Difficulty dressing self

- Tall

Person	Item	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

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

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

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

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (\mu_{1j} + \mu_{0j} + \varepsilon_{ij})$$ **[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: $\sigma_{\mu 0 \mu 1}$

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)

To model age…

- Do you really want to model age, or are you substantively interested in period or cohort effects?
- Longitudinal data on multiple cohorts over time ideal
- If not available, consider age, cohort, and period (including practice) effects
- We focus on settings where cohort and period effects can be ignored, and longitudinal data are available

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 okay to use:
- $Y_{ij}=\beta_{00}+\beta_1*age\_now+ \mu_i+ \varepsilon_{ij}$
- If not, why?
- Can examine practice, period, etc, e.g., :
- $Y_{ij}=\beta_{00}+\beta_1*age\_base+\beta_2*(age\_now - age\_base)+\beta_3*(1^{st}\ wave)+\mu_i+\varepsilon_{ij}$

Growth curves and lifecourse

- Growth curves are very powerful to study human development.

Alternative time dimensions

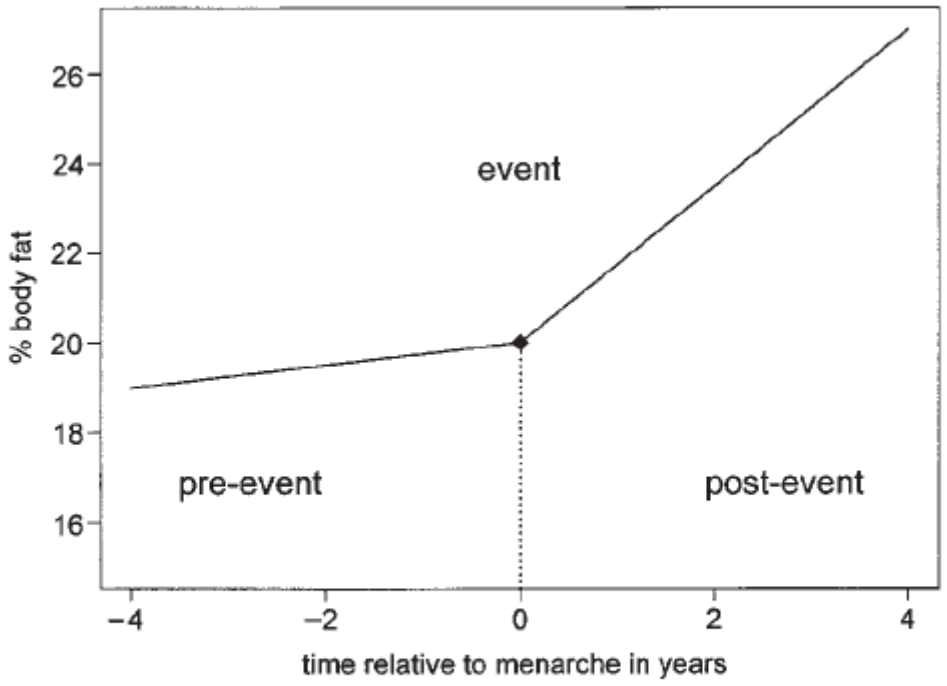


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

Choose a time-dimension:

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

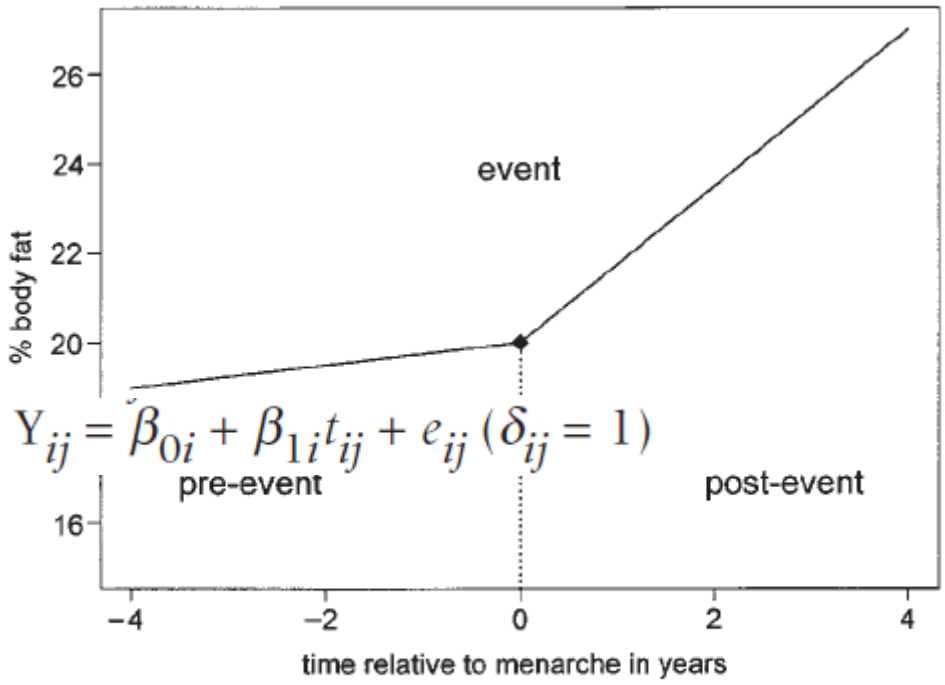
Specify a model with time as a linear predictor

- 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)



$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij} \; (\delta_{ij} = 1)$$

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)

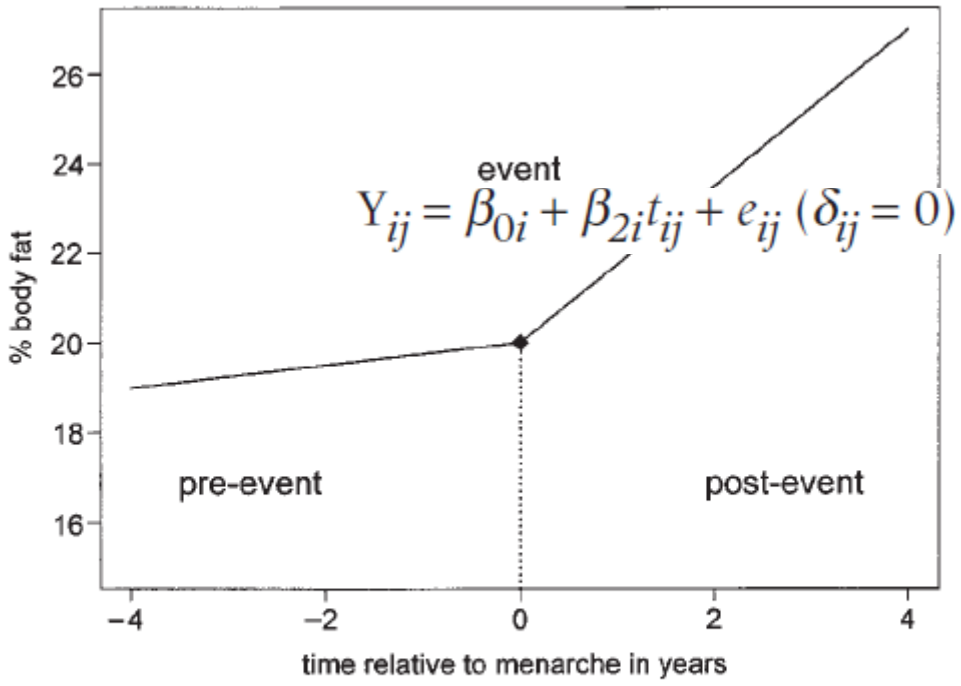


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

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

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

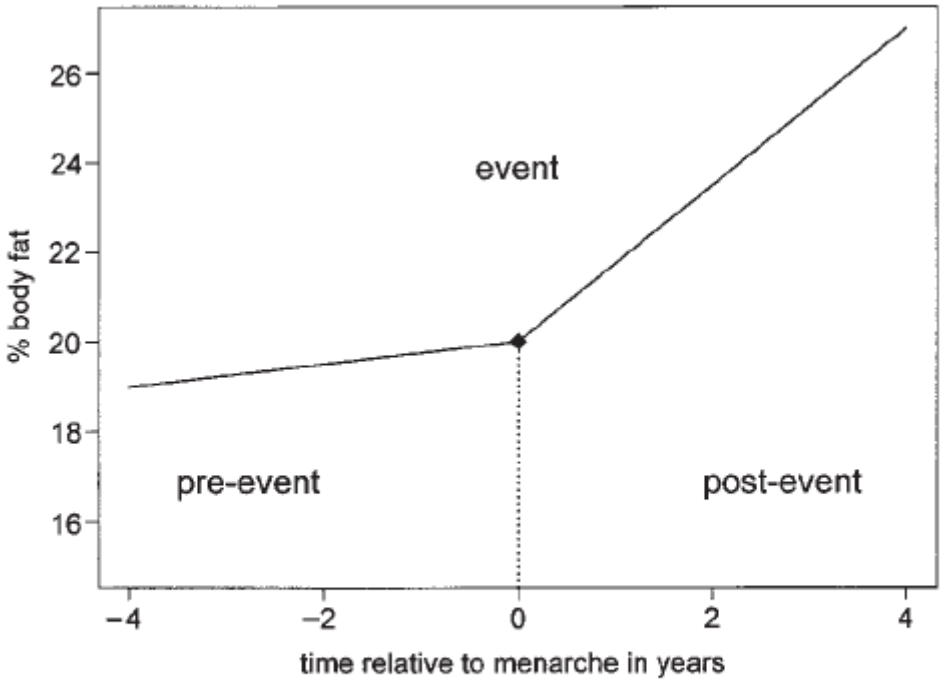


Figure 4 Graphical representation of the research hypothesis and modelling approach: the piecewise model assigns a pre-event slope β_1 and a post-event slope β_2 separated by the menarcheal event

Choose a time-dimension:

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

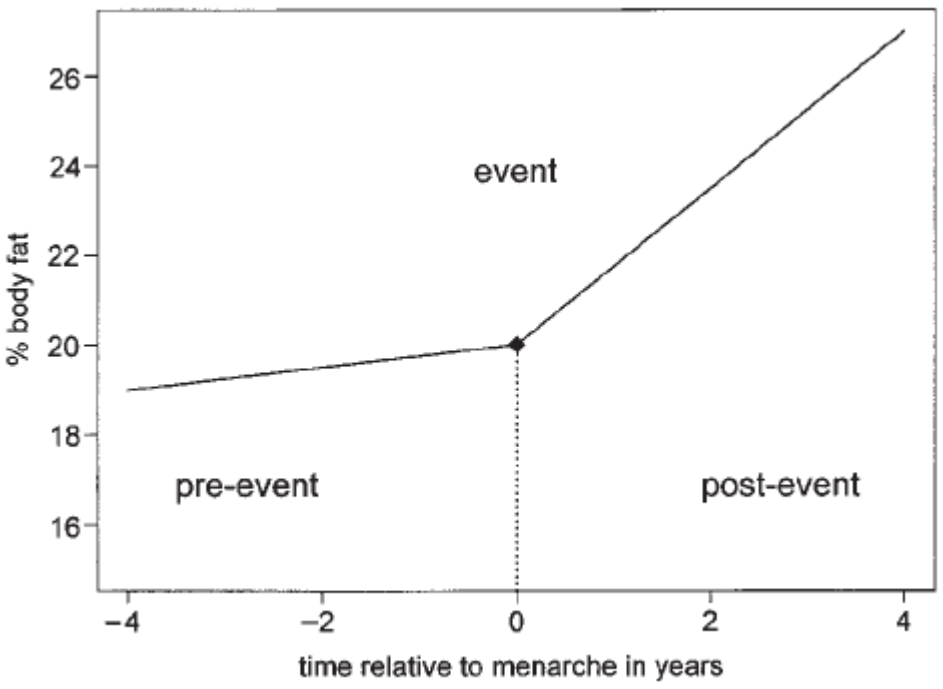
Specify a model with time as a linear predictor

- 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

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
- Interactions
- Options for age effects: cross-sectional, GEE, or growth curves
- Lifecourse timing

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

End may 3

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.

82

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?