Longitudinal Data Analysis

Linear Mixed Effects Model

S. M. Tafsir Hasan, MBBS, MS (Applied Statistics)

Assistant Scientist

Maternal and Child Nutrition

Nutrition and Clinical Services Division

icddr,b

Clustered and Repeated Measures Data

Clustered data

- Dependent variable is measured once for each subject, but the subjects themselves are somehow grouped (math scores of students from different classes)

Repeated measures data

- Dependent variable is measured more than once for each subject
- They can be repeated over space (the right knee gets the control operation and the left knee gets the experimental operation)
- They can be repeated over condition (each subject gets both the high and low cognitive load condition)
- They can be repeated over time (longitudinal data) – most common in public health

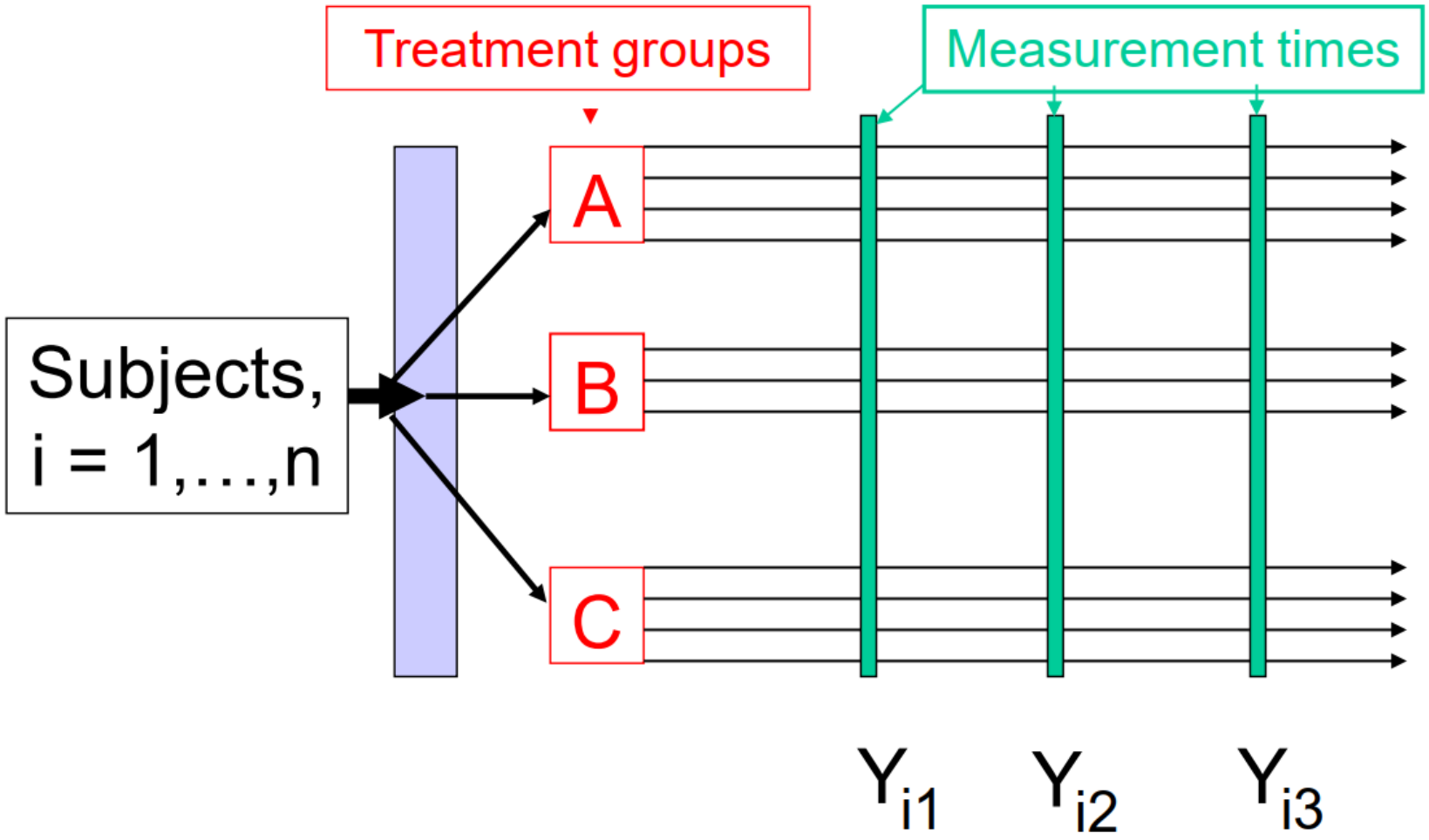
Similarity

- Each unit has multiple observations
- Not necessarily require the same number of observations per unit
- The missingness should be random
- Dependent variables are likely to be correlated (not independent)
- Longitudinal data can also be clustered (analyzing those students' math scores over three years)
- Analysis needs to take this into account this dependence, otherwise standard errors are likely to be underestimated, leading to inflated Type 1 error
- From an analysis point of view, all three are multilevel data and can be analyzed using some sort of mixed or multilevel modeling

Dissimilarity

- Time itself is often an important independent variable in longitudinal studies, but in repeated measures studies, it is usually confounded with some independent variable
- Time may be measured with some proxy like Age or Order

Longitudinal Data



Wide Format

- Longitudinal data can be in a wide format
- One row of data per person
- Observations on a variable each occasion are recorded in different columns

id	age	sex	group	dep1	dep2	dep3
1	19	1	1	44	43	40
2	21	2	1	50	47	45
3	30	1	2	43	43	40
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Long Format

- Longitudinal data can be in a long format
- One row per observation (or measure) rather than one row per person
- A single column for each of the repeated outcome measures

id	age	sex	group	dep
1	19	1	1	44
1	19	1	1	43
1	19	1	1	40
2	21	2	1	50
2	21	2	1	47
2	21	2	1	45
3	30	1	2	43
3	30	1	2	43
3	30	1	2	40
…	…	…	…	…
…	…	…	…	…

Example Dataset

- A study was conducted to investigate the effects of smoking on birth outcomes using the Natality dataset derived from birth certificates by the US National Center for Health Statistics
- A total of 8604 births from 3978 mothers are in the dataset with up-to 3 births per mother

Research question

Does smoking during pregnancy affect birth weight?

Load the Dataset

- Let's load the **birthweight** dataset

- The dataset contains the following variables
  - ✓ momid: mother's identification
  - ✓ Idx: order of childbirth
  - ✓ mage: mother's age at the birth of the child (in years)
  - ✓ smoke: mother's smoking status during pregnancy (1: yes; 0: no)
  - ✓ married: mother being married (1: yes; 0: no)
  - ✓ black: mother being black (1: black; 0: white)
  - ✓ male: baby being male (1: male; 0: female)
  - ✓ birwt: birthweight (in grams)

Identify Data Structure (Long/Wide)

describe

```
Contains data from C:\Users\tafsi\Desktop\smoking.dta
  obs:         8,604
 vars:             8                          29 Nov 2021 23:21
 size:       275,328
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
momid           float   %9.0g
idx             float   %9.0g
mage            float   %9.0g
smoke           float   %9.0g      s
married         float   %9.0g
black           float   %9.0g      b
male            float   %9.0g      m
birwt           float   %9.0g
-------------------------------------------------------------------------------
Sorted by: momid
```

Identify Data Structure

browse

	momid	idx	mage	smoke	married	black	male	birwt
1	14	1	16	Nonsmoker	0	Black	Female	2790
2	14	2	17	Nonsmoker	0	Black	Female	2693
3	14	3	20	Nonsmoker	0	Black	Female	3600
4	25	1	23	Nonsmoker	0	Black	Male	2778
5	25	2	24	Nonsmoker	0	Black	Male	2835
6	25	3	26	Nonsmoker	0	Black	Male	3402
7	39	1	36	Nonsmoker	1	White	Male	2948
8	39	2	38	Nonsmoker	1	White	Male	3487
9	39	3	39	Nonsmoker	1	White	Male	3345
10	48	1	30	Smoker	1	White	Male	2880
11	48	2	32	Smoker	1	White	Male	3275
12	60	1	23	Nonsmoker	0	Black	Female	3714
13	60	2	26	Nonsmoker	0	Black	Female	3799
14	71	1	20	Smoker	1	White	Male	3203
15	71	2	22	Smoker	1	White	Female	3742
16	71	3	22	Smoker	1	White	Female	3760
17	73	1	25	Nonsmoker	1	White	Male	3450
18	73	2	27	Nonsmoker	1	White	Male	3580
19	86	1	34	Nonsmoker	1	White	Male	3345
20	86	2	38	Nonsmoker	1	White	Female	2835

Identify Data Structure

tab1 idx smoke

-> tabulation of idx

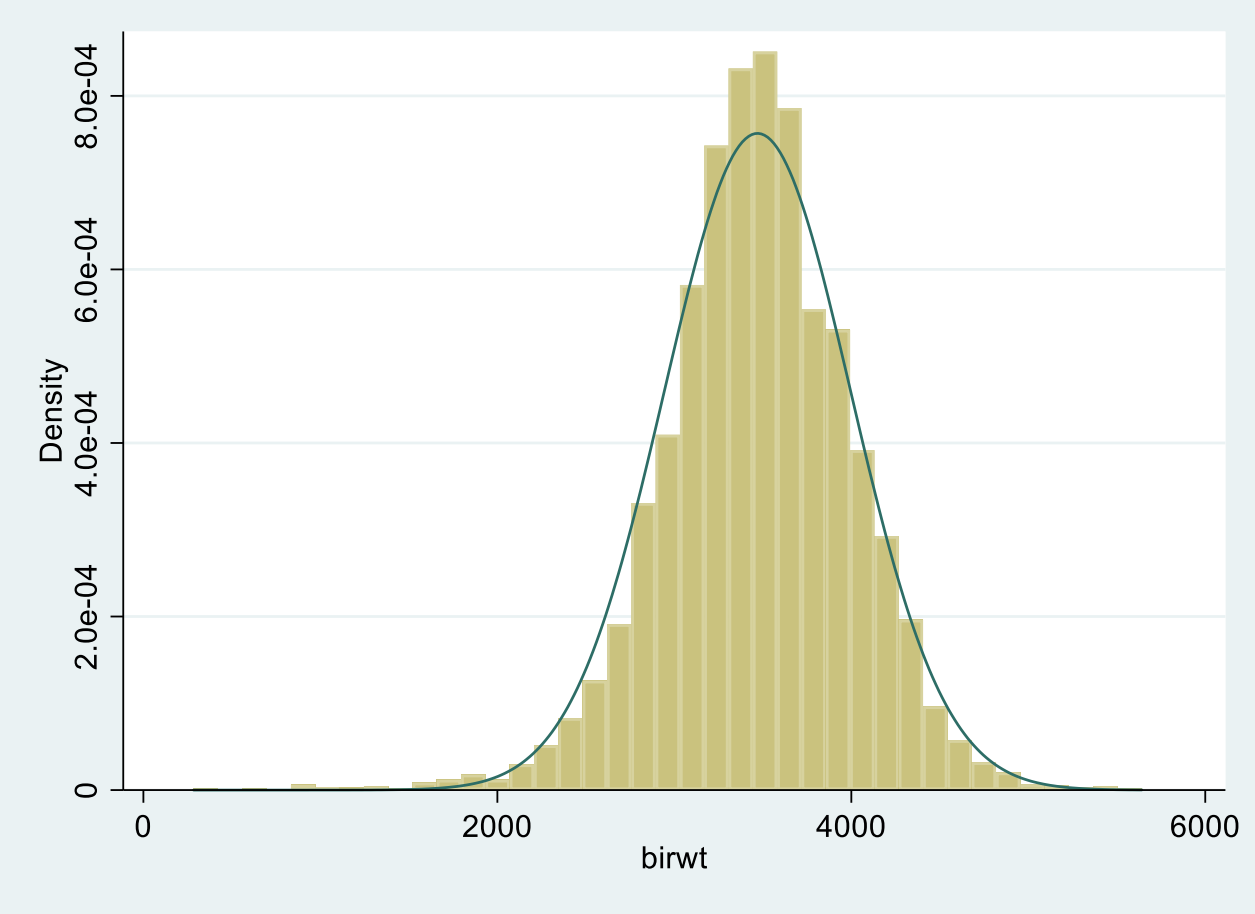
idx	Freq.	Percent	Cum.
1	3,978	46.23	46.23
2	3,978	46.23	92.47
3	648	7.53	100.00
Total	8,604	100.00	

-> tabulation of smoke

smoke	Freq.	Percent	Cum.
Nonsmoker	7,400	86.01	86.01
Smoker	1,204	13.99	100.00
Total	8,604	100.00	

Check Distribution of Birthweight

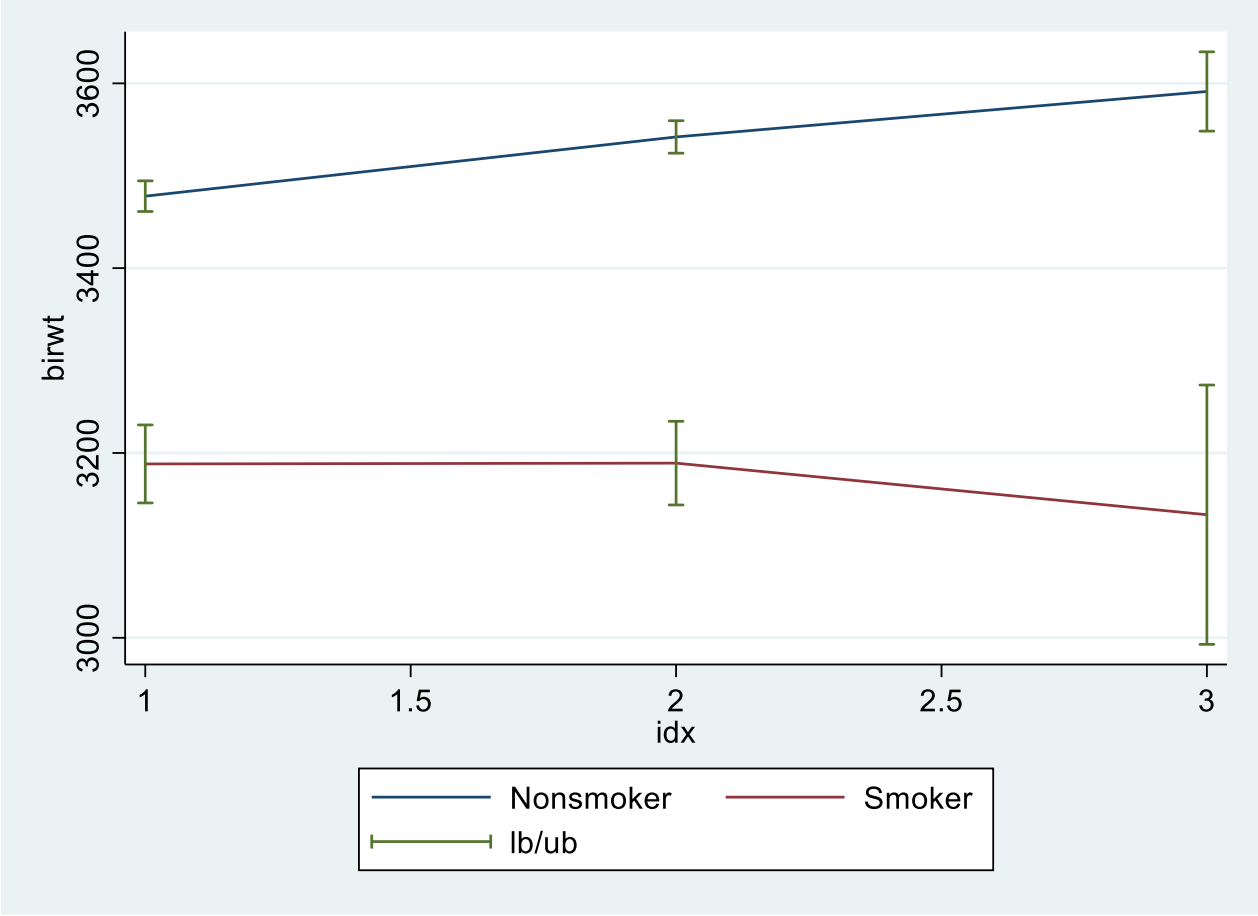
hist birwt, norm



Check Distribution of Birthweight

ssc install xtgraph, replace

xtgraph birwt, t(idx) i(momid) group(smoke)



Change Data Format - Long to Wide

- Conversion from one form to the other can be done using ***reshape*** command
- From Long to Wide form

  reshape wide birwt mage smoke male, i(momid) j(idx)

```
. reshape wide birwt mage smoke male, i(momid) j(idx)
(note: j = 1 2 3)

Data                                long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      8604   ->    3978
Number of variables                    8   ->      15
j variable (3 values)                idx   ->   (dropped)
xij variables:
                                   birwt   ->   birwt1 birwt2 birwt3
                                    mage   ->   mage1 mage2 mage3
                                   smoke   ->   smoke1 smoke2 smoke3
                                    male   ->   male1 male2 male3
```

Wide Data

describe

```
Contains data
  obs:         3,978
 vars:            15
 size:       238,680
```

variable name	storage type	display format	value label	variable label
momid	float	%9.0g		
mage1	float	%9.0g		**1 mage**
smoke1	float	%9.0g	s	**1 smoke**
male1	float	%9.0g	m	**1 male**
birwt1	float	%9.0g		**1 birwt**
mage2	float	%9.0g		**2 mage**
smoke2	float	%9.0g	s	**2 smoke**
male2	float	%9.0g	m	**2 male**
birwt2	float	%9.0g		**2 birwt**
mage3	float	%9.0g		**3 mage**
smoke3	float	%9.0g	s	**3 smoke**
male3	float	%9.0g	m	**3 male**
birwt3	float	%9.0g		**3 birwt**
married	float	%9.0g		
black	float	%9.0g	b	

```
Sorted by: momid
     Note: Dataset has changed since last saved.
```

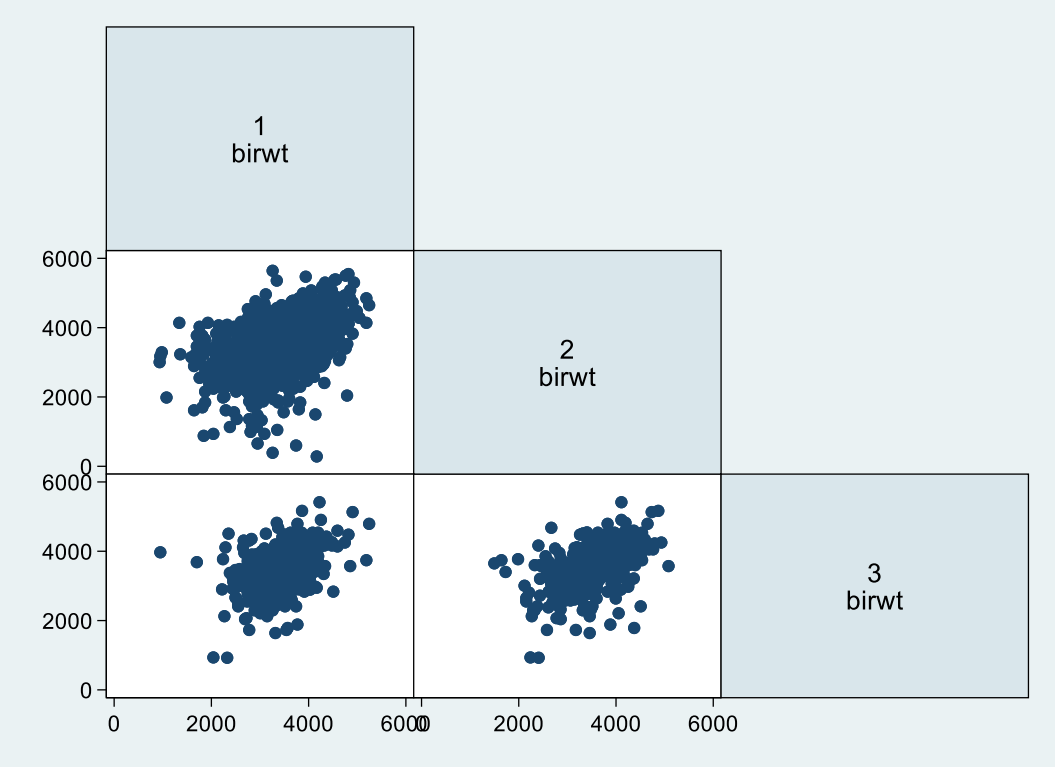
Wide Data

browse

	momid	mage1	smoke1	male1	birwt1	mage2	smoke2	male2	birwt2	mage3	smoke3	male3	birwt3
1	14	16	Nonsmoker	Female	2790	17	Nonsmoker	Female	2693	20	Nonsmoker	Female	3600
2	25	23	Nonsmoker	Male	2778	24	Nonsmoker	Male	2835	26	Nonsmoker	Male	3402
3	39	36	Nonsmoker	Male	2948	38	Nonsmoker	Male	3487	39	Nonsmoker	Male	3345
4	48	30	Smoker	Male	2880	32	Smoker	Male	3275	.	.	.	.
5	60	23	Nonsmoker	Female	3714	26	Nonsmoker	Female	3799	.	.	.	.
6	71	20	Smoker	Male	3203	22	Smoker	Female	3742	22	Smoker	Female	3760
7	73	25	Nonsmoker	Male	3450	27	Nonsmoker	Male	3580	.	.	.	.
8	86	34	Nonsmoker	Male	3345	38	Nonsmoker	Female	2835	.	.	.	.
9	110	30	Nonsmoker	Male	2892	32	Nonsmoker	Female	2948	.	.	.	.
10	115	28	Nonsmoker	Female	3880	30	Nonsmoker	Female	3260	.	.	.	.
11	138	29	Nonsmoker	Female	3997	32	Nonsmoker	Female	3402	.	.	.	.
12	142	26	Nonsmoker	Male	3685	29	Nonsmoker	Female	3685	31	Nonsmoker	Female	3969
13	151	21	Smoker	Male	3160	23	Nonsmoker	Male	3220	.	.	.	.
14	156	29	Nonsmoker	Male	2750	30	Nonsmoker	Female	3345	.	.	.	.
15	160	23	Nonsmoker	Male	3629	25	Nonsmoker	Female	3385	.	.	.	.
16	171	23	Smoker	Male	3005	25	Smoker	Male	2537	.	.	.	.
17	206	30	Nonsmoker	Male	4195	31	Nonsmoker	Male	3154	.	.	.	.
18	209	20	Nonsmoker	Female	2665	21	Nonsmoker	Female	3884	.	.	.	.
19	214	27	Nonsmoker	Male	2960	31	Nonsmoker	Female	3317	.	.	.	.
20	227	31	Smoker	Female	2975	32	Smoker	Male	1890	.	.	.	.

Check Correlations of Birthweight

graph matrix birwt*, half



corr birwt*

	birwt1	birwt2	birwt3
birwt1	1.0000		
birwt2	0.4839	1.0000	
birwt3	0.4370	0.4920	1.0000

Missing Values in LDA

- Missing completely at random (MCAR): Missingness is nothing to do with the person being studied, e.g., a questionnaire might be lost in the post, or a blood sample might be damaged in the lab
- Missing at random (MAR): Missingness is to do with the person but can be predicted from other information about the person. It is not specifically related to the missing information. For example, if a child does not attend an educational assessment because the child is (genuinely) ill, this might be predictable from other data we have about the child's health, but it would not be related to what we would have measured had the child not been ill
- Missing not at random (MNAR): Missingness is specifically related to what is missing, e.g., a person does not attend a drug test because the person took drugs the night before

Check Missing Values

mdesc

Variable	Missing	Total	Percent Missing
momid	0	3,978	0.00
mage1	0	3,978	0.00
smoke1	0	3,978	0.00
male1	0	3,978	0.00
birwt1	0	3,978	0.00
mage2	0	3,978	0.00
smoke2	0	3,978	0.00
male2	0	3,978	0.00
birwt2	0	3,978	0.00
mage3	3,330	3,978	83.71
smoke3	3,330	3,978	83.71
male3	3,330	3,978	83.71
birwt3	3,330	3,978	83.71
married	0	3,978	0.00
black	0	3,978	0.00

Change Data Format – Wide to Long

- Conversion from one form to the other can be done using ***reshape*** command
- From Wide to Long form

  reshape long birwt mage smoke male, i(momid) j(idx)
- ***idx*** variable is not in the dataset, we must create it
- Check the data in Long form

  describe

  browse

Longitudinal Data Analysis

Subject-specific models (mixed effects/multilevel model)

- Estimate subject-specific effects over time
- Average of the slopes
- Focus usually on individual trajectories
- Assumptions are made and may matter

Population-averaged models (GEE)

- Estimate regression coefficients for the population rather than for any individual
- Slope of the averages
- Focus usually on averages (their trajectories)
- Robust

Which Model Is Useful for What?

Subject-specific models (mixed effects/multilevel model)

- To compare individual changes over time (trajectories)
- To study natural history

Population-averaged models (GEE)

- To compare group of populations over time
- To inform public policy

Fixed and Random Effects Model

Fixed effects model

- A regression model in which the group means are fixed (non-random)

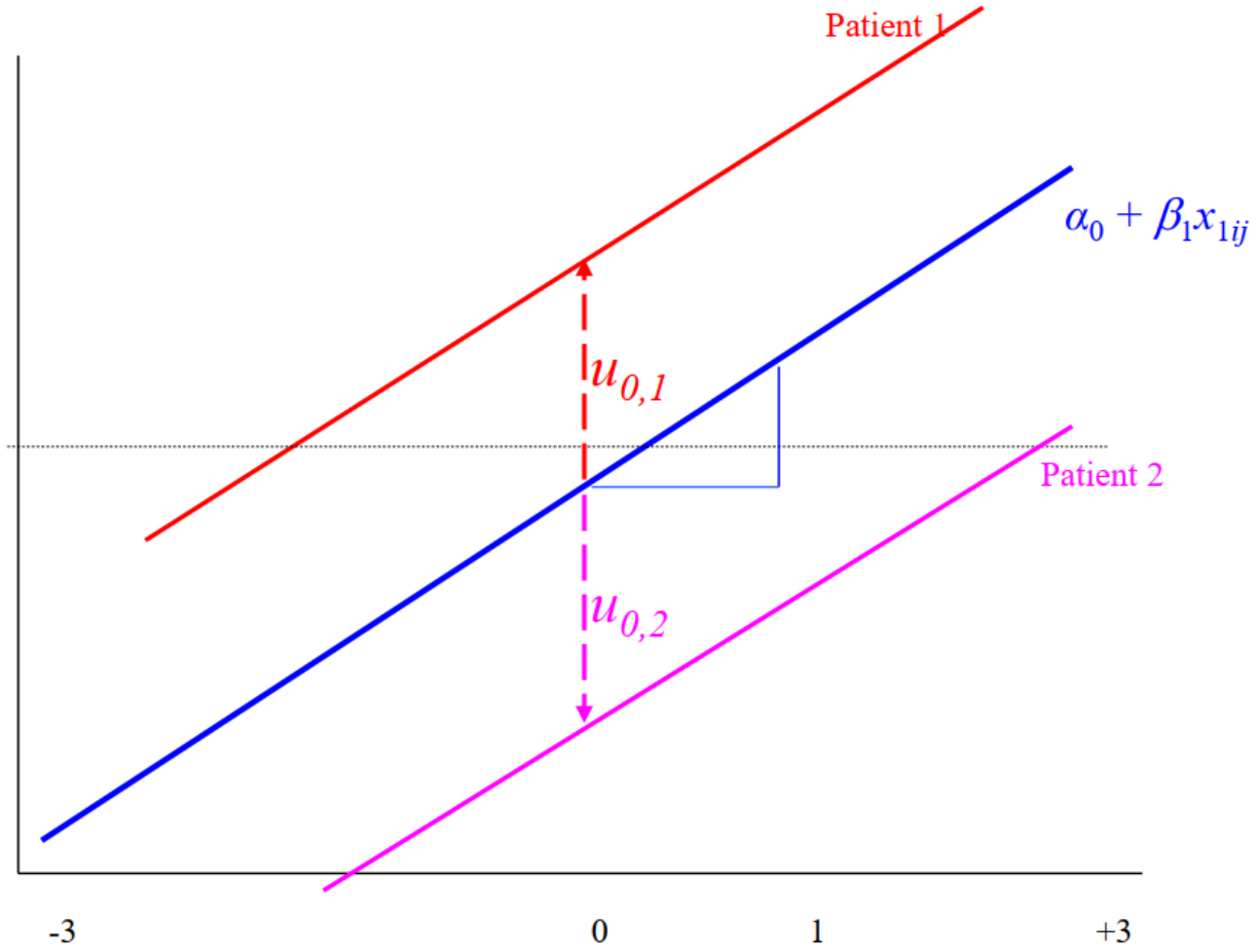
Random effects model

- A regression model in which the group means are a random sample from a population

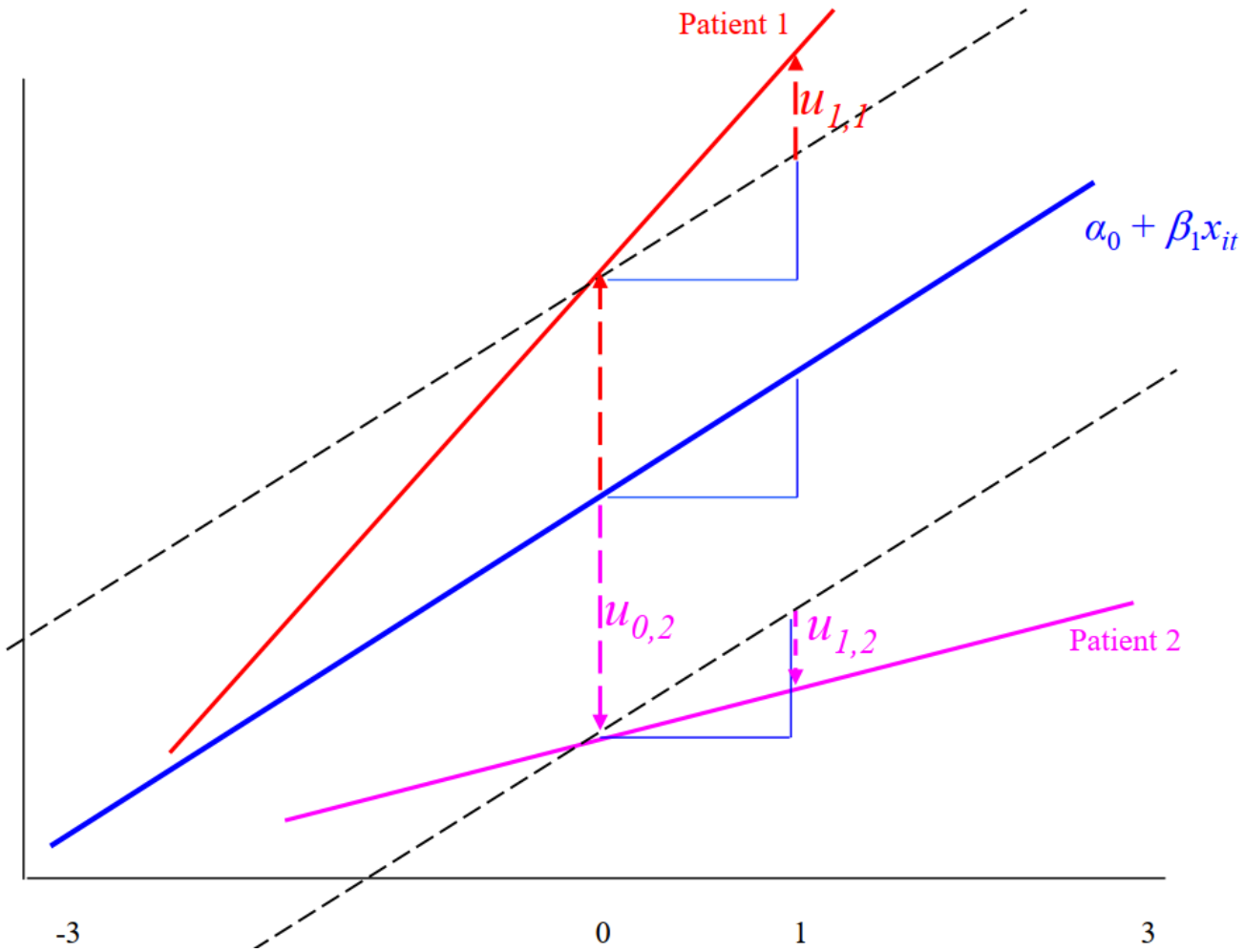
Mixed effects model

- Model containing both fixed effects and random effects–
  - ✓ Random intercept model: intercept terms vary by individuals (subject-specific)
  - ✓ Random slope model: both intercepts and slopes vary among individuals

Random Intercept Model



Random Slope Model



Fit Two Level Random Intercept Model

mixed birwt smoke male mage married black || momid:

```
Mixed-effects ML regression                     Number of obs      =      8,604
Group variable: momid                           Number of groups   =      3,978

                                                Obs per group:
                                                              min =          2
                                                              avg =        2.2
                                                              max =          3

                                                Wald chi2(5)       =     637.56
Log likelihood = -65171.429                     Prob > chi2        =     0.0000
```

birwt	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke	-233.6451	17.72467	-13.18	0.000	-268.3849	-198.9054
male	120.6073	9.578245	12.59	0.000	101.8343	139.3803
mage	9.707351	1.222748	7.94	0.000	7.310808	12.10389
married	83.23453	24.25818	3.43	0.001	35.68937	130.7797
black	-220.2905	28.24326	-7.80	0.000	-275.6463	-164.9347
_cons	3104.512	38.52752	80.58	0.000	3029	3180.025

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
momid: Identity				
var(_cons)	116018.9	4300.052	107889.8	124760.6
var(Residual)	137929.7	2878.461	132401.8	143688.3

LR test vs. linear model: chibar2(01) = 1118.15 Prob >= chibar2 = 0.0000

estimates store randinter

Fit Two Level Random Slope Model

mixed birwt smoke male mage married black || momid: mage

```
Mixed-effects ML regression                     Number of obs      =      8,604
Group variable: momid                           Number of groups   =      3,978

                                                Obs per group:
                                                              min =          2
                                                              avg =        2.2
                                                              max =          3

                                                Wald chi2(5)       =     648.09
Log likelihood = -65166.613                     Prob > chi2        =     0.0000
```

birwt	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke	-233.5315	17.57302	-13.29	0.000	-267.974	-199.089
male	120.0169	9.567708	12.54	0.000	101.2645	138.7692
mage	9.928066	1.227432	8.09	0.000	7.522342	12.33379
married	81.22849	23.75763	3.42	0.001	34.66438	127.7926
black	-219.5471	27.84409	-7.88	0.000	-274.1206	-164.9737
_cons	3100.66	37.99864	81.60	0.000	3026.184	3175.136

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
momid: Independent				
var(mage)	37.9779	12.35102	20.07736	71.83817
var(_cons)	84221.88	10702.79	65653.14	108042.4
var(Residual)	137679.6	2876.942	132154.8	143435.4

LR test vs. linear model: chi2(2) = 1127.78 Prob > chi2 = 0.0000

estimates store randslope

Compare Two Models

- Likelihood-ratio test to compare the models

  lrtest randslope randinter

```
Likelihood-ratio test                                 LR chi2(1)  =      9.63
(Assumption: randinter nested in randslope)           Prob > chi2 =    0.0019
```

- Near-zero significance level favors the more complex model over the simpler model

Fit Two Level Model with Correlated Random Effects

mixed birwt smoke male mage married black || momid: mage, cov(un)

```
Mixed-effects ML regression                     Number of obs      =      8,604
Group variable: momid                           Number of groups   =      3,978

                                                Obs per group:
                                                              min =          2
                                                              avg =        2.2
                                                              max =          3

                                                Wald chi2(5)       =     634.13
Log likelihood = -65163.904                     Prob > chi2        =     0.0000
```

birwt	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke	-233.1492	17.59949	-13.25	0.000	-267.6436	-198.6549
male	119.9587	9.563988	12.54	0.000	101.2136	138.7037
mage	9.779008	1.25601	7.79	0.000	7.317274	12.24074
married	80.47698	23.9384	3.36	0.001	33.55858	127.3954
black	-219.9888	27.97912	-7.86	0.000	-274.8268	-165.1507
_cons	3106.432	38.85466	79.95	0.000	3030.279	3182.586

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
momid: Unstructured				
var(mage)	296.5203	116.5281	137.2603	640.5664
var(_cons)	280652.8	89287.07	150441.5	523565.8
cov(mage,_cons)	-7253.911	3251.107	-13625.96	-881.8594
var(Residual)	136907.6	2895.636	131348.3	142702.2

```
LR test vs. linear model: chi2(3) = 1133.20                   Prob > chi2 = 0.0000
```

estimates store randcovar

Compare Two Models

- Likelihood-ratio test to compare the models

  lrtest randcovar randslope

```
Likelihood-ratio test                                 LR chi2(1)  =      5.42
(Assumption: randslope nested in randcovar)           Prob > chi2 =    0.0199
```

- Near-zero significance level favors the more complex model over the simpler model

Fit OLS and GEE Models

- Fit OLS model

  quietly reg birwt smoke male mage married black

  estimates store ols

- Fit GEE model

  quietly xtgee birwt smoke male mage married black, i(momid) family(normal) link(identity) cor(exc)

  estimates store gee

Compare Estimates from All Models

estimates table ols gee randinter randslope randcovar, equation(1)

Variable	ols	gee	randinter	randslope	randcovar
#1					
smoke	-278.76468	-235.31741	-233.64515	-233.53149	-233.14924
male	116.02169	120.4816	120.60729	120.01689	119.95868
mage	8.244738	9.6346846	9.7073513	9.9280656	9.7790082
married	77.212196	83.064998	83.234531	81.228489	80.47698
black	-227.10904	-220.53633	-220.29051	-219.54714	-219.98878
_cons	3162.0768	3107.0766	3104.5124	3100.6598	3106.4323
lns1_1_1					
_cons			5.8307544	1.8185022	2.8460579
lnsig_e					
_cons			5.9172497	5.9163424	5.9135307
lns1_1_2					
_cons				5.670605	6.2724369
atr1_1_1_2					
_cons					-1.0853388