



# Longitudinal Data Analysis: Missing Data

Sergio R. Muñoz, PhD

[sergio.munoz.n@ufrontera.cl](mailto:sergio.munoz.n@ufrontera.cl)

Departamento de Salud Pública-Centro de Excelencia CIGES



**I Jornada de Cohortes Poblacionales Latinoamericanas  
para el Estudio de Enfermedades Crónicas –COPLAS**

**4-6 abril 2018**

**Hotel Termas de Quinamávica, Linares, Chile**





## Outline

- Longitudinal studies
- Case study
- Data analysis strategies
- Analytical tools
- Investigating missingness
- Classification of missing data
- Methods
- Recommendations



# Longitudinal studies

---

- Repeatedly collect data on the same individuals over time
- Advantages
  - Record incident cases
  - Evaluate exposure prospectively
  - Identify changes over time within individuals
  - Help to determine causal effect of exposure on outcome



# Longitudinal studies

---

- Repeatedly collect data on the same individuals over time
- Challenges
  - Determine causality when covariates vary over time
  - Decide about exposure lag when varies over time
  - Data is correlated (needs special statistical methods)
  - How to account for incomplete subject follow-up



## Outline

- Longitudinal studies
- **Case study**
- Data analysis strategies
- Analytical tools
- Investigating missingness
- Classification of missing data
- Methods
- Recommendations



# Case Study

- **Effects of intermittent exposure to high altitude on the health of workers**
  - First occupational cohort study of miners in Chile with 5 years follow-up
  - Objective:
    - Evaluate the effect of labor conditions on health of miners, over time, working on different levels of altitude in Chile
  - Specific objectives:
    - Determine the effects of intermittent exposure to high altitude on physiological, cognitive and social variables.
    - Determine the effect of the exposure to high altitude on the prevalence and incidence of occupational diseases and accidents
    - Establish the foundations to identify the best options of work cycles and daily activities for the population of workers exposed to high altitude



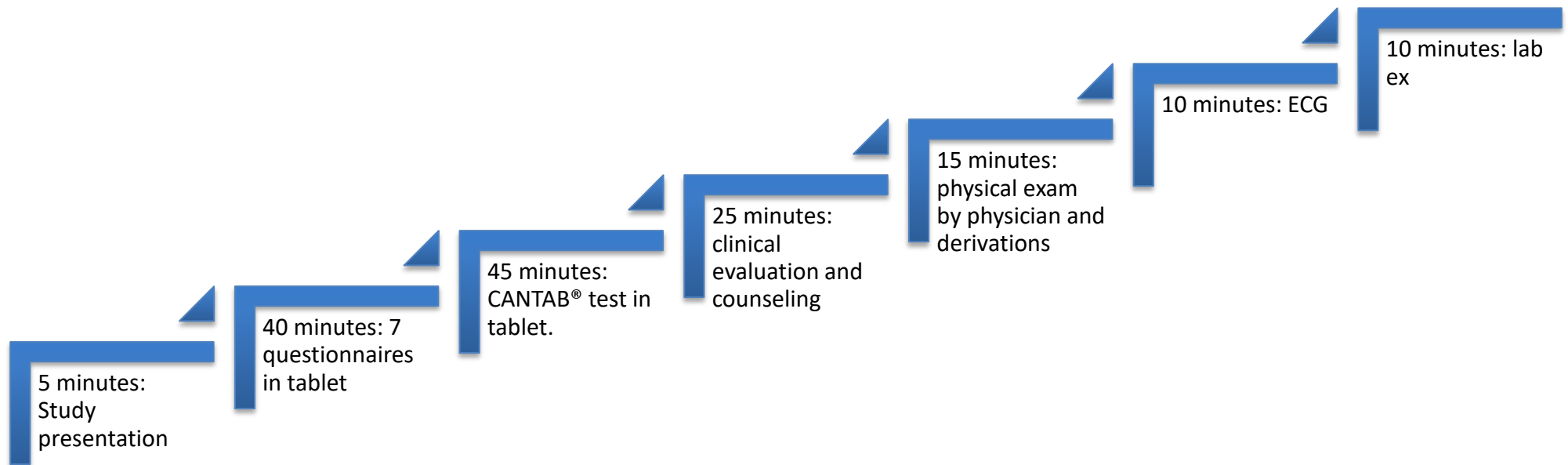
# Design

---

- Longitudinal study (5 years follow-up)
- Cohort study with two levels of exposure
  - Exposed group: Miners working over 3,000 meters over sea level
  - Un-exposed group: Miner working under 2,400 meters over sea level



# Measurements



**Application of instruments  
(2,5 hours approximately)**





# Follow-up

Follow-up	2015	2016	2017	N (%)
2015-2016-2017	x	x	x	355 (59,8)
2015-2016	x	x		46 (7,7)
2015- 2017	x		x	42 (7,1)
2015	x			52 (8,8)
2016- 2017		x	x	41 (6,9)
2016		x		8 (1,2)
2017			x	50 (8,4)
Total	499	450	488	594 (100%)



## Outline

- Longitudinal studies
- Case study
- **Data analysis strategies**
- Analytical tools
- Investigating missingness
- Classification of missing data
- Methods
- Recommendations











# Data analysis strategies

---

- Graphics
  - Scatterplot Matrix
  - Subject specific outcome profile
  - Boxplot
  - Mean profile over time
- Analytical tools



# Data

File Edit View Data Tools										
       										
id[1]					1					
	id	grupo	estado	edad	med0	med3	med6	med9	med12	
1	1	1	0	52	.	21.6	21.5	18.6	.	
2	2	1	0	54	26.4	27.9	23.2	22	19.5	
3	3	1	0	56	27.3	26.6	23.5	24.4	22.2	
4	4	1	0	57	27.4	27.1	28.1	.	21.4	
5	5	1	0	63	27.9	27	26	23	23.5	
6	6	1	0	57	29.7	25.6	24.6	21.4	20.6	
7	7	1	0	53	28.3	.	22.3	24	20.7	
8	8	1	0	59	29.7	24.6	25.6	18.9	25.1	
9	9	1	0	57	29.4	26	23.8	22	20.3	
10	10	1	0	48	23.3	23.7	22.1	18.6	19	
11	11	1	0	56	27.9	24.9	24.5	.	21	
12	12	1	0	47	23.6	22.7	21.7	.	17.5	
13	13	1	0	47	26.8	26.6	25.7	25.2	20.9	
14	14	1	0	53	25.6	22.5	21	20.4	16.2	
15	15	1	0	55	26.5	24.7	25.2	.	20.5	
16	16	1	0	56	23	21.3	25.1	20.3	16.6	
17	17	1	0	51	25.4	25.2	23.4	22.5	21.5	
18	18	1	0	76	32.9	31.4	31.5	.	26.4	
19	19	1	0	65	27.2	29.3	.	.	23.8	
20	20	1	0	43	20.5	.	16.5	16.3	14.8	
21	21	1	0	54	23.7	25.4	22.2	19.7	20.6	



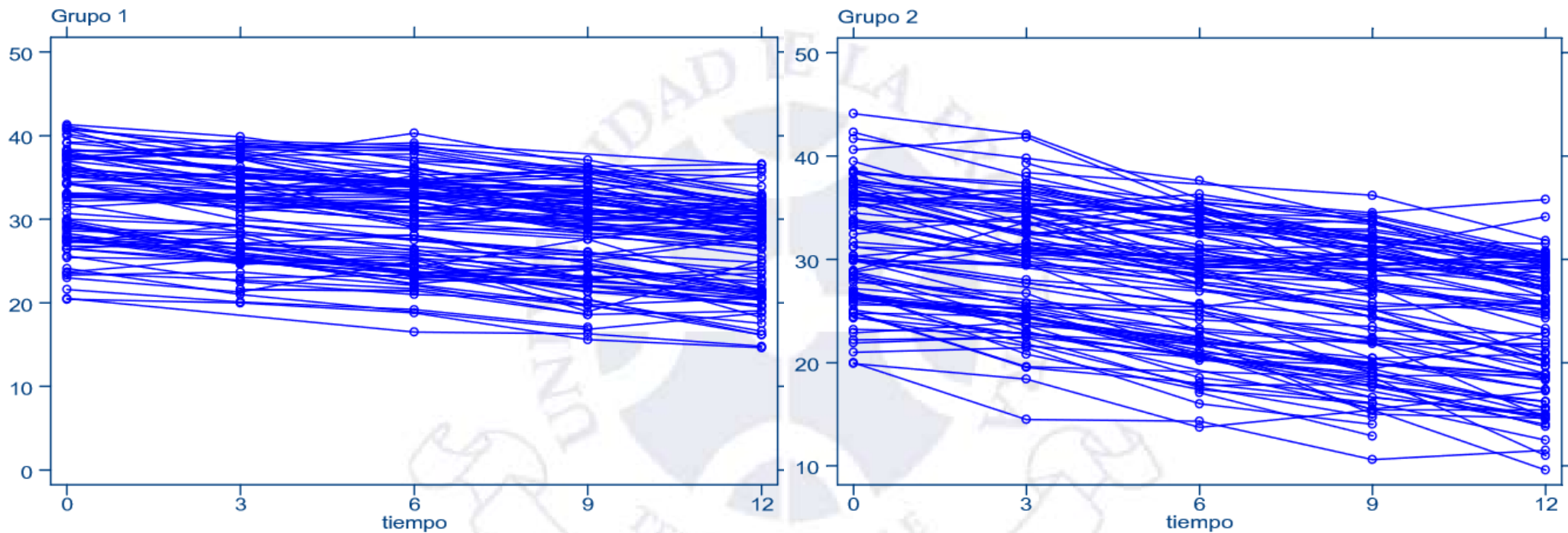


# reshape long med, i ( id) j( tiempo)

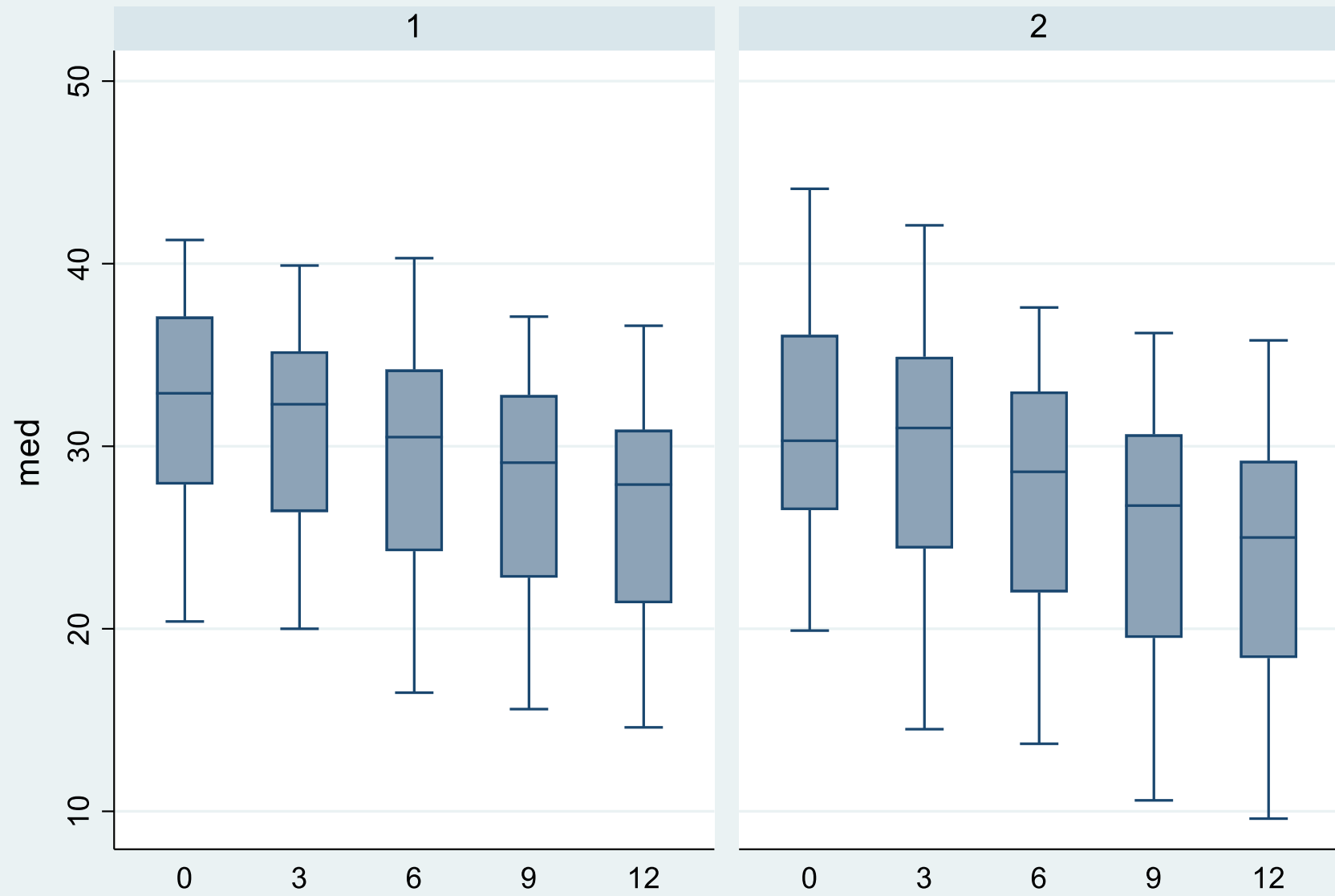
	id	tiempo	grupo	estado	edad	med
1	1	0	1	0	52	.
2	1	3	1	0	52	21.6
3	1	6	1	0	52	21.5
4	1	9	1	0	52	18.6
5	1	12	1	0	52	.
6	2	0	1	0	54	26.4
7	2	3	1	0	54	27.9
8	2	6	1	0	54	23.2
9	2	9	1	0	54	22
10	2	12	1	0	54	19.5
11	3	0	1	0	56	27.3
12	3	3	1	0	56	26.6
13	3	6	1	0	56	23.5
14	3	9	1	0	56	24.4
15	3	12	1	0	56	22.2
16	4	0	1	0	57	27.4
17	4	3	1	0	57	27.1
18	4	6	1	0	57	28.1
19	4	9	1	0	57	.
20	4	12	1	0	57	21.4



# Subject specific outcome trajectories



# Boxplot



Graphs by grupo

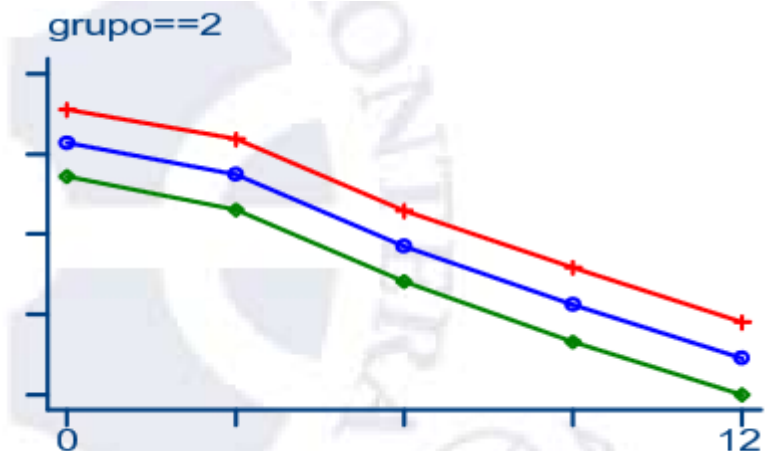
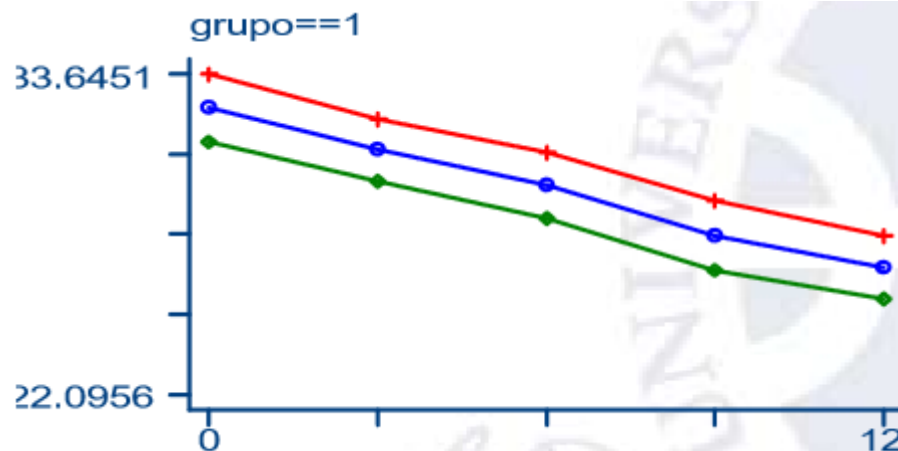




# Mean profile over time

collapse (mean) med (sd) sdmed=med (count) n=med, by(tiempo grupo)

tiempo	grupo	med	sdmed	n
0	1	32.42912	5.514437	79
0	2	31.16385	5.683793	83
3	1	30.91798	5.388784	89
3	2	30.03605	6.024653	86
6	1	29.63636	5.663214	88
6	2	27.44253	6.084151	87
9	1	27.82785	5.713882	79
9	2	25.33295	6.390829	88
12	1	26.66966	5.50643	89
12	2	23.40562	6.305505	89





## Outline

- Longitudinal studies
- Case study
- Data analysis strategies
- **Analytical tools**
- Investigating missingness
- Classification of missing data
- Methods
- Recommendations



# Analytical Tools

- Simple situation
  - Baseline measurement ( $t=0$ )
  - Single follow-up measurement ( $t=1$ )
- Pre-post design
  - Ignore baseline, analyze  $t=1$
  - Analyze difference post-pre (change)
  - Analyze post adjusting for pre
  - Analyze change adjusting for pre



# More than two measurements

- Include all data in a statistical model
- Generalized estimating equations (GEE)
  - Contrast average outcome values across populations of subjects defined by covariates taking into account the correlated structure of the data
  - Assumptions:
    - Observations are independent across subjects
    - Observations may be correlated within subjects



# GEE (Liang-Zeger)

```
xtset id tiempo
```

```
panel variable: id (strongly balanced)
```

```
time variable: tiempo, 0 to 12, but with gaps
```

```
. xtgee med grupo estado edad, i(id) t( tiempo) corr(exc) link(iden) fam(gauss)
```

GEE population-averaged model

Group variable:

Link: identity

Family: Gaussian

Correlation: exchangeable

Scale parameter:

10.97193

Number of obs = 857

Number of groups = 200

Obs per group:

min = 2

avg = 4.3

max = 5

Wald chi2(3) = 2594.30

Prob > chi2 = 0.0000

med	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
grupo	-1.569764	.2168276	-7.24	0.000	-1.994738	-1.144789
estado	10.15155	.2193104	46.29	0.000	9.721705	10.58139
edad	.3064352	.0146242	20.95	0.000	.2777722	.3350982
_cons	8.078129	.8669802	9.32	0.000	6.378879	9.777379

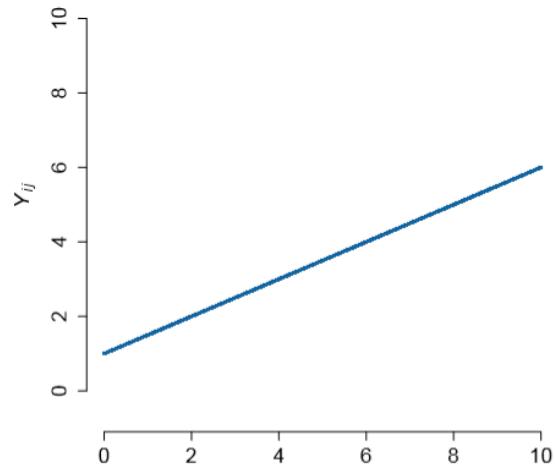


# Mixed effect model (Laird-Ware)

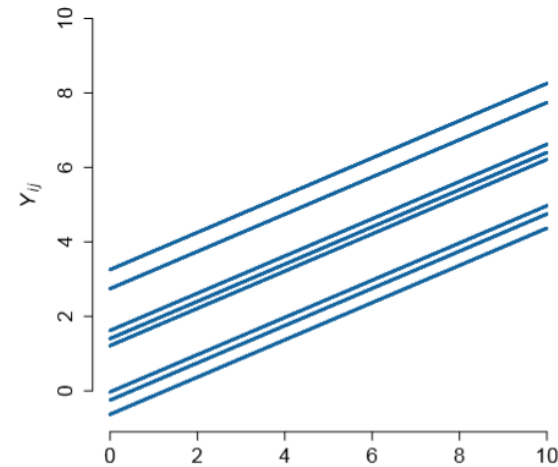
- Consider outcome measurements both between and within subjects
- Assume that each subject has a regression model
- Combine fixed-effect parameters common to all subjects in the population with random-effect parameters unique for each subject
- Subject-specific random effects induce a correlation structure
- Two stage estimation procedure

# Choices for modelling

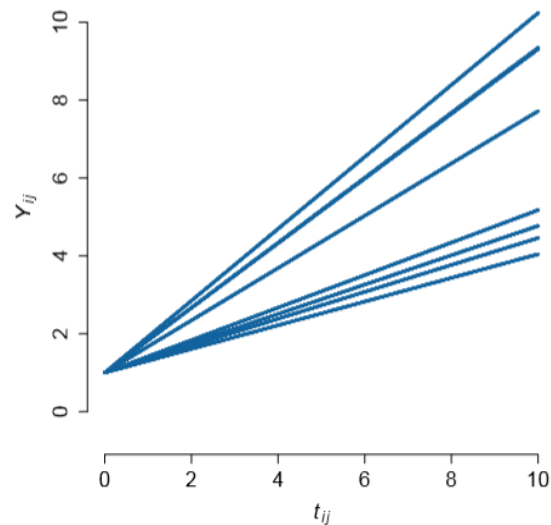
Fixed intercept, fixed slope



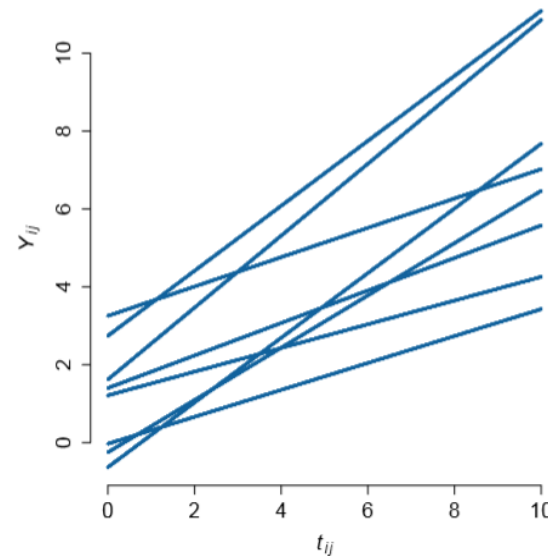
Random intercept, fixed slope



Fixed intercept, random slope



Random intercept, random slope



xtmixed



- Longitudinal studies
- Case study
- Data analysis strategies
- Analytical tools
- Investigating missingness
- Classification of missing data
- Methods
- Recommendations





# Missing

A measurement was intended to be taken and was not!

- Types:
  - Survey sampling
  - Item non-response from survey
  - Loss to follow-up
  - Discontinuous follow-up
- Situation:
  - Refused to participate (representativeness)
  - Subject refused or tired to answer
  - Informed consent retired
  - Loses eligibility, ...
  - Missing one follow-up, but returns for the next



# Is missing a problem?

- Standard Statistical methods are not aware of missing data
- If missing cases are deleted:
  - Reduces sample size and lower statistical power (lower SE and harder to detect sig relationships)
  - Biased estimates (sample selection bias) because analytic sample may not be representative of whole sample
- If we impute missing data
  - Risk of biased estimates: inadequate imputations
  - Biased standard errors (significance tests and confidence intervals)
- Publication of research: Journal editors and reviewers are increasingly strict about how you deal with missing data



# What should we do?

---

- Investigate quantity and patterns of missingness
  - On subject
  - On outcomes and covariates
- Investigate mechanism of missingness



# On subjects

```
. egen num_missing=rowmiss( med0 med3 med6 med9 med12)
```

```
. tab num_missing
```

num_missing	Freq.	Percent	Cum.
0	91	45.50	45.50
1	81	40.50	86.00
2	22	11.00	97.00
3	6	3.00	100.00
Total	200	100.00	



# Quantity and patterns

- Elaborate table of missingness over time

```
. misstable summarize med0 med3 med6 med9 med12
```

Obs<.

Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
med0	38		162	113	19.9	44.1
med3	25		175	118	14.5	42.1
med6	25		175	123	13.7	40.3
med9	33		167	120	10.6	37.1
med12	22		178	119	9.6	36.6



# Quantity and patterns

- Elaborate patterns of missingness

. misstable pattern, freq  
Missing-value patterns  
(1 means complete)

Frequency	Pattern				
	1	2	3	4	5
<b>91</b>	1	1	1	1	1
<b>19</b>	0	1	1	1	1
<b>19</b>	1	1	1	0	1
<b>15</b>	1	1	0	1	1
<b>14</b>	1	0	1	1	1
<b>14</b>	1	1	1	1	0
<b>4</b>	0	0	1	1	1
<b>4</b>	0	1	1	1	0
<b>3</b>	0	1	1	0	1
<b>3</b>	1	1	0	0	1
<b>2</b>	0	0	1	0	1
<b>2</b>	0	1	0	0	1
<b>2</b>	0	1	0	1	1
<b>2</b>	1	0	0	1	1
<b>2</b>	1	0	1	0	1
<b>1</b>	0	0	1	1	0
<b>1</b>	0	1	1	0	0
<b>1</b>	1	1	0	1	0
<b>1</b>	1	1	1	0	0
<b>200</b>					



# What should we do?

---

- Investigate quantity and patterns of missingness
  - On subject
  - On outcomes and covariates
- Investigate mechanism of missingness



# Understanding the nature of the missing data pattern

- Generate a dummy variable for missingness
- $M = \begin{cases} 1 & \text{if missing} \\ 0 & \text{if measured} \end{cases}$
- Fit a logistic model using  $m$  as outcome and relevant variables as covariates (OR should be equal to 1)
- May help in deciding about how to handle missing data





# Mechanism

---

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random



# MCAR

- Missing cases are unrelated to any variable in the analysis
- The probability that an observation is missing is a random event independent of subject factors.  
 $P[M=1 \mid \text{SBP, Age}] = P[M=1]$
- Complete cases are a random sample of the full dataset
- Analysis remains unbiased, but lose power
- Most missing data techniques will work well



# MAR

- Missingness depends on observed variables
- Overall estimates are biased in complete cases analysis
- $P(R = 1 \mid \text{SBP, Age}) = P(R = 1 \mid \text{Age})$
- Amongst subjects of the same age, missingness in SBP is independent of SBP



# Missing not at random

---

- Usually named as informative
- Sample looks like a convenient sample
- Rarely seen



# Methods

---

- Last observation carried forward (conservative if positive time trend in the outcome)
- Use complete cases
  - No problem if MCAR
  - Biased if MAR or informative



# Methods

- Single imputation
  - Estimate a predicted value for the missing value
  - Use imputed values in the analysis
  - Unbiased if MCAR or MAR
- Multiple imputation
  - Impute several times
  - Use multiple values to estimate variability
  - Unbiased if MCAR or MAR
- Inverse probability weighting
  - Inflate subjects by the inverse probability of being non-missing
  - Unbiased if MCAR or MAR



# Do we need to impute?

---

- It all depends!
- If  $<5\%$  , use complete case analysis
- If  $>50\%$ , you should quit and look for another job.
- Otherwise, imputation is best option



# Recommendations

---

- Try your best in doing a good job at all stages of your research
- Do not forget to consider a biostatistician at earliest stages of your project, but...



