



UNIVERSITY OF
LIVERPOOL

Department of Biostatistics

Introduction to Longitudinal Data Analysis

A Biostatistics PGR Development Course
FLHR 627

Dr Trevor Cox & Dr Graeme Hickey



25th April 2017



Course overview

Times	Topic	Instructor
Room: WHELAN-B09		
0945 to 0950	Welcome	Trevor Cox
0950 to 1015	Introduction to longitudinal data	Trevor Cox
1015 to 1045	Reduced data	Trevor Cox
1045 to 1100	Break	
1100 to 1200	Summary measures	Trevor Cox
1200 to 1300	Break	
Room: SHER-SR1		
1300 to 1345	Extending ANOVA	Graeme Hickey
1345 to 1415	Missing data	Graeme Hickey
1415 to 1430	Break	
1430 to 1500	Sample size calculations	Graeme Hickey
1500 to 1530	Linear mixed models	Graeme Hickey

Course learning objectives

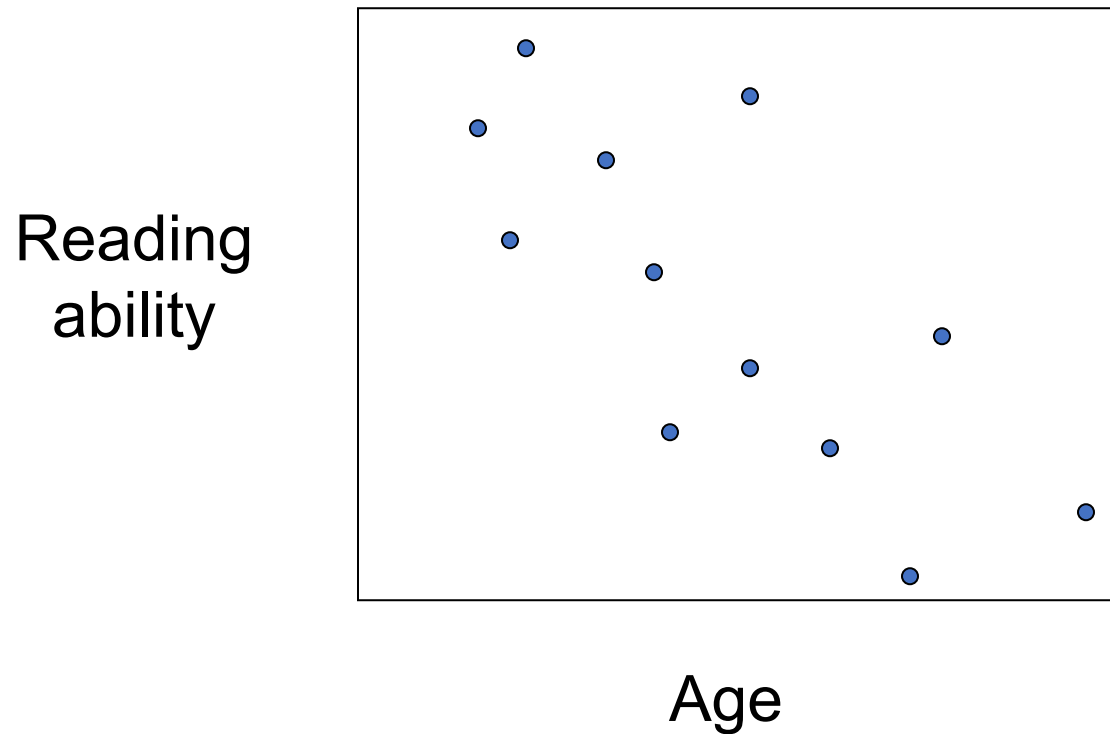
- To appreciate options available when analysing longitudinal data
- To relate choice of analysis to the clinical question
- To obtain some practical experience of analysing longitudinal data
- To identify what you can do yourself and when to ask a statistician for help

Introduction to longitudinal data

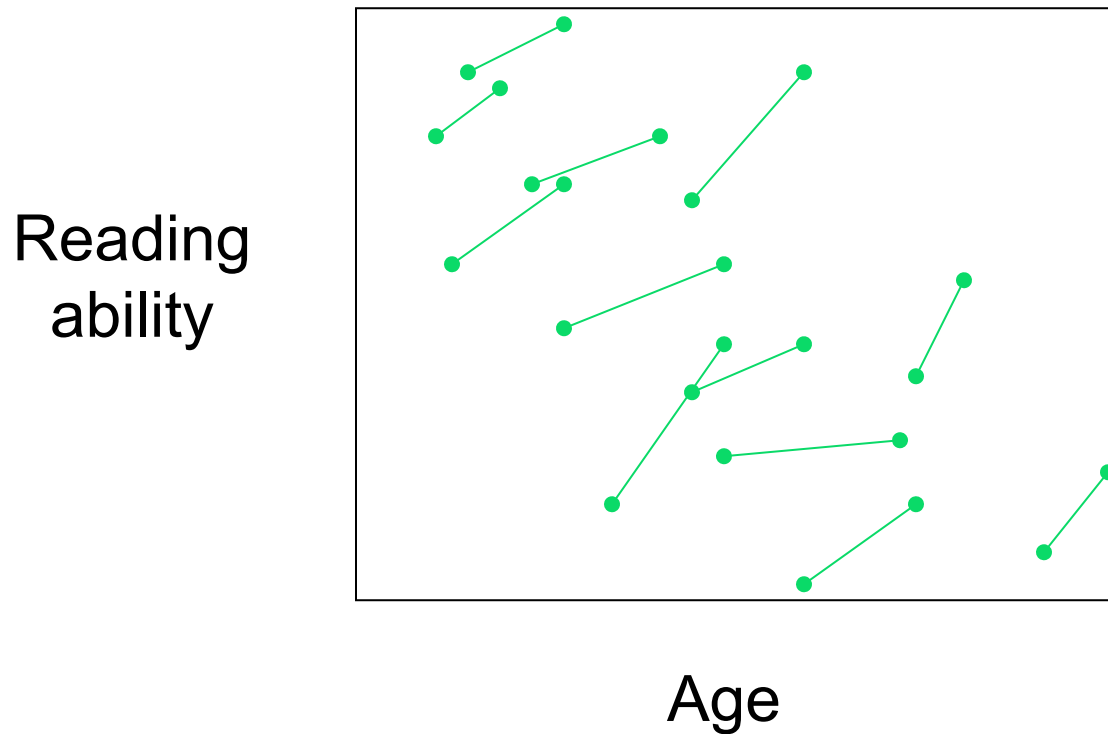
What is longitudinal data?

- Sometimes called *serial* or *repeated* measurements
- Individuals are measured repeatedly through time
- At least 2 measurements per person
- Can be collected prospectively or retrospectively
- Can distinguish changes over time within individuals from differences among people in their baseline levels

Example: cross-sectional survey



Example: longitudinal data



Collecting longitudinal data

Id	T1	T2	T3	T4	T5	T6	T7
1	2.83	1.15	2.34	2.84	0.36	3.37	3.01
2	8.70					0.55	
3	1.98	3.36	2.30	1.86	1.18	0.00	7.44
4	1.82	0.98	1.83	1.04	0.90	0.67	0.17
5	5.20						
6	3.93	0.39	1.05	1.91	0.17	0.13	0.67
7	2.81	1.06	2.10	1.60		1.73	0.09
8	1.80	0.95	2.45	2.00	1.32	8.86	2.45
9	3.59	1.02	0.78	0.52	na	0.00	0.15
10	3.20	4.30		4.10	0.80	0.00	2.30
11	1.20	1.26	1.25	0.78	1.62	0.17	6.03
12	10.29	1.62	1.53	0.00	0.59	8.12	2.21
13	2.70	4.56	2.13	3.19	2.43	1.60	1.16
14	3.40	2.00					
15	13.90	3.40	0.68	0.73	0.76	1.80	1.10
16	5.70	4.50	3.91	6.01	1.36	3.20	0.87
17	2.02	1.12	0.46	0.34	0.32	0.3	0
18	2.40	0.60	0.30	0.59	0.20	0.35	0.33
19	4.30	2.74		1.20	1.11	2.70	3.89
20	6.80	1.76	1.60	1.30	0.57	1.70	4.48
21	1.71			0.43	0.18	0.39	0.93

Defining features of longitudinal data

- Repeated observations on a set of individuals
- Measurements between individuals are independent
- Measurements within individuals are correlated
- Repeated observations on the same individual will be more similar to each other than to observations on other individuals

Defining features of longitudinal data

- Decision on how frequently to measure the outcome depends on clinical area
 - E.g. in a study of cardiovascular disease and obesity, measurements of blood pressure, BMI and cholesterol levels may be **repeated every 3 months** - balanced design
- Balanced and unbalanced designs
- Methods of analysis need to account for non independence of measurements within an individual

What not to do!

- Do not put all the data together as if they were one sample
 - Multiple counting of individual patients
 - Artificially inflates sample size which may lead to spurious statistical significance
- RCT of ketoprofen and aspirin in the treatment of rheumatoid arthritis
 - P -value=0.00000001 was obtained from an analysis of 3944 observations
 - Data were obtained from only 58 patients
 - Average of 68 measurements per patient

What not to do!

- Do not perform separate analyses at each time point
 - Multiple testing
 - Series of non independent results difficult to interpret
 - Focus on those providing statistically significant results
- RCT comparing two active treatments for onchocerciasis and a placebo
 - Outcome: clinical reaction score
 - Data collected daily for 8 days and then at 10 days, 4 weeks, 3 months, and 6 months
 - Pairwise significance tests were performed: gave 36 *P*-values

Types of data

- Choice of analysis must depend on the type of outcome data
- Continuous
 - Data that can take any value in a specified range, e.g. systolic blood pressure, weight
- Binary
 - Presence/absence of an event, e.g. presence of pain, respiratory disease
- Count
 - Number of times an event occurs, e.g. number of seizures per week

Back to basics

- *t*-test

- Used to test null hypothesis $\bar{\mu}_1 = \bar{\mu}_2$
- Means of the two groups are the same

- ANOVA

- Extension of the *t*-test
- Used to test the null hypothesis

$$\bar{\mu}_1 = \bar{\mu}_2 = \cdots = \bar{\mu}_n$$

- Looks to see how much of the overall variability in the data can be explained by variation between individuals *within* each group, and that due to any systematic difference *between* the groups

Example: two-sample t -test



- During the 1970s a psychological test designed to measure extroversion was applied to a group of admitted male streakers and a group of male non-streakers
- Null hypothesis, $H_0: m_1 = m_2$, i.e. there is no difference between population means
- H_0 : there is no difference in mean extroversion scores between male streakers and non-streakers

Example: two-sample *t*-test

- Summary statistics

$$\bar{X}_1 = 15.26 \quad \bar{X}_2 = 13.90$$

$$n_1 = 19 \quad n_2 = 19$$

$$s_1^2 = 2.62 \quad s_2^2 = 4.11$$

- Calculate estimate of **pooled variance**

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{18 \times 2.62 + 18 \times 4.11}{19 + 19 - 2} = 3.365 \end{aligned}$$

Example: two-sample t -test

- Calculate **standard error** of $\bar{x}_1 - \bar{x}_2$

$$SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.834 \sqrt{\frac{1}{19} + \frac{1}{19}} = 0.595$$

- Test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{SE} = \frac{15.26 - 13.90}{0.595} = 2.286$$

- Compare $T = 2.29$ to t -tables with $(n_1 + n_2 - 2)$ degrees of freedom
- P -value = 0.03 \rightarrow reject H_0 at 5% significance level

Example: ANOVA

- 4 groups, 10 people in each group
- Each group receives a different diet supplementation
- Outcome is satisfaction rating
- If we used t -tests then we would have 6 possible pairwise comparisons to look at
 - Multiple testing!
- Use ANOVA and we have 1 test of the hypothesis

$$\mu_A = \mu_B = \mu_C = \mu_D$$

Example: ANOVA

Group			
A	B	C	D
4	5	7	2
4	5	8	1
5	6	7	2
5	6	9	3
6	7	6	3
3	6	3	4
4	4	2	5
4	5	2	4
3	6	2	4
4	3	3	3
4.2	5.3	4.9	3.1

Grand Mean
= 4.375

Example: ANOVA

Calculate the *between* sum of squares

Sum of all the squared differences **between the individual means and the grand mean**

$$\begin{aligned}SS_{\text{between}} &= 10[(4.2 - 4.375)^2 + (5.3 - 4.375)^2 + \\&\quad (4.9 - 4.375)^2 + (3.1 - 4.375)^2] \\&= 27.875\end{aligned}$$

$$\begin{aligned}\text{Degrees of freedom (df)} &= \text{number of groups} - 1 \\&= 3\end{aligned}$$

Example: ANOVA

Calculate the *within* sum of squares

Sum of all the squared differences **between individual data and the group mean within each group**

$$\begin{aligned} SS_{within} &= (4 - 4.2)^2 + \dots + (4 - 4.2)^2 + \\ &\quad (5 - 5.3)^2 + \dots + (3 - 5.3)^2 + \\ &\quad (7 - 4.9)^2 + \dots + (3 - 4.9)^2 + \\ &\quad (2 - 3.1)^2 + \dots + (3 - 3.1)^2 \\ &= 101.500 \end{aligned}$$

$$\begin{aligned} df &= \text{number of observations} - \text{number of groups} \\ &= 40 - 4 = 36 \end{aligned}$$

Example: ANOVA

Calculate the *total* sum of squares

Sum of the squared differences **between all the individual data and the grand mean**

Also equal to $SS_{between} + SS_{within} = 27.875 + 101.500$

df = number of observations – 1
 = 40 - 1

Example: ANOVA

Source	SS	df	Mean square (MS)	<i>F</i>
Between	27.875	3	9.292	3.296
Within	101.500	36	2.819	
Total	129.375	39		

Example: ANOVA

- Compare $F = 3.296$ to F -tables on 3 and 36 degrees of freedom $\rightarrow P < 0.05$
- F is the ratio of the differences between groups to the variation within groups

Conclusion

There is sufficient evidence to suggest that not all diet supplements are equal in terms of satisfaction. We know at least two are different from each other, but we do not know which two.

Reduced data

Reduced data

- Multiple measurements recorded per person
- Interest may be in the **response at one period** in time; usually the last
 - Ignore the others
- Interest may be in the **change between first and last**
 - Ignore the intermediate values

Reduced data

1. Single time point (last value)
2. Change
3. Change relative to baseline
4. Analysis of covariance

Single time point

- Reduces data to one measurement per person
- RCT in obesity: Diet A vs Diet B
 - Participants followed up for 5 months, weigh ins every month

Baseline



Final measurement



Patient	Group	Month 0	Month 1	Month 2	Month 3	Month 4	Month 5
1	A	252	235	225	212	206	202
2	A	168	160	155	148	137	135
3	A	172	167	160	150	137	134
4	A	155	155	149	145	143	140
...
39	B	204	192	183	172	171	168
40	B	168	163	159	155	154	149

Single time point analysis

- Interested in final measurement only
- Methods of analysis
 - Analysis of variance (P -value)
 - t -test (P -value)
 - Calculate the difference in means and 95% confidence interval

Caution

- Single time point of interest has to be specified in advance
- Do not apply these methods separately to each time point
 - Multiple testing
 - Does not account for measurements at different time points being from the same subjects
- Doesn't account for the impact of baseline measurements

Change

- Interest is in total amount of change between first and last observations for each individual
- Calculate the difference between these observations for each individual
 - Have a single value for each individual
- Analyse as for single time point
- Amount of change may be correlated to the baseline value
- Does not correct for baseline imbalance if this is true

Patient	Group	Month 0	Month 1	Month 2	Month 3	Month 4	Month 5	Change
1	A	252	235	225	212	206	202	-50
2	A	168	160	155	148	137	135	-33
3	A	172	167	160	150	137	134	-38
4	A	155	155	149	145	143	140	-15
...
39	B	204	192	183	172	171	168	-33
40	B	168	163	159	155	154	149	-19

Change relative to baseline

- Expresses the amount of change relative to baseline
- Calculated as $\frac{post - pre}{pre}$
- Can be expressed as a percentage
- Calculate this for each individual
- Analyse as for single time point

Patient	Group	Month 0	Month 1	Month 2	Month 3	Month 4	Month 5	% Change
1	A	252	235	225	212	206	202	-19.9
2	A	168	160	155	148	137	135	-19.6
3	A	172	167	160	150	137	134	-22.1
4	A	155	155	149	145	143	140	-9.7
...
39	B	204	192	183	172	171	168	-16.2
40	B	168	163	159	155	154	149	-11.3

Single time point *versus* change

- Trial measuring shoulder pain
 - One baseline measure, one follow up measure
- Comparison of follow-up scores
 - At the end of the trial, mean pain scores were 15mm lower in the Tx group 95% CI (10mm to 20mm)
- Change score
 - Pain reductions were 20mm greater on Tx than control 95% CI (16mm to 24mm)

Single time point *versus* change

- If Tx is effective the relative statistical significance of the treatment effect by the two methods will depend on the correlation between baseline and follow up scores
- If correlation is low using the change score will add more variation
- If correlation is high using follow up score only will lose information
- Incorrect to choose most significant
 - Approach should be specified in the protocol

Single time point *versus* change

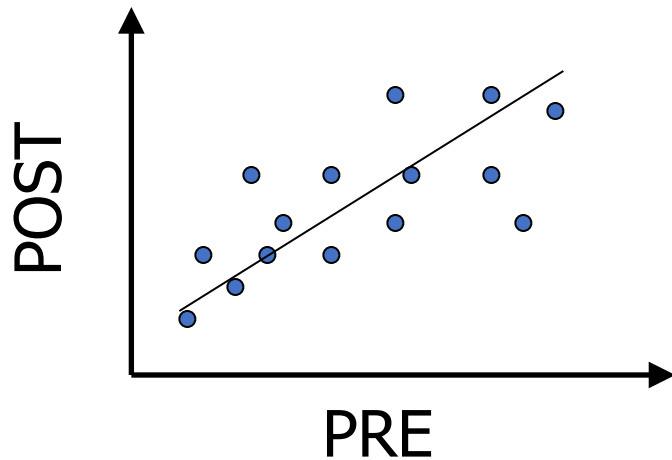
- Analysing change does not always control for baseline imbalance if correlation exists between change and baseline
- % change from baseline is statistically inefficient → low power
- ANCOVA is a better approach

Analysis of covariance (ANCOVA)

- With ANOVA we saw that we assess the effect of an intervention by comparing the amount of variability in the data that the experiment can explain against the variability it can not explain
- If we can explain some of this 'unexplained' variance in terms of other variables (covariates) then we reduce the error variance, allowing us to more accurately assess the effect of the independent variable

Analysis of covariance (ANCOVA)

- Essentially a simple linear regression model



$$\begin{array}{ccccc} \text{post} & & & \text{pre} & \\ \downarrow & & & \downarrow & \\ y_i = \alpha + \beta x_i + \varepsilon_i & & & & \\ \uparrow & \uparrow & & \uparrow & \\ \text{intercept} & \text{slope} & & \text{error} & \end{array}$$

Analysis of covariance (ANCOVA)

- Models final response taking account of baseline measure
 - Line of best fit
- Extend model to include additional covariates e.g. treatment group

$$\begin{array}{ccccccc} \text{post} & & \text{pre} & & \text{Tx indicator} & & \\ \downarrow & & \downarrow & & \downarrow & & \\ y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i & & & & & & \\ \uparrow & \uparrow & & \uparrow & \uparrow & & \uparrow \\ \text{intercept} & \text{slope} & & \text{effect of Tx} & \text{error} & & \end{array}$$

Analysis of covariance (ANCOVA)

- Adjusts each patients follow up score for their baseline score and is unaffected by baseline differences
- If baseline scores are worse in Tx group
 - Tx effect underestimated by follow up score
 - Tx effect overestimated by change score
 - ANCOVA gives the right answer whether or not there is baseline imbalance

Example: acupuncture for shoulder pain*

- 52 patients randomised to true or sham acupuncture
- Pre & post assessment using 100 point rating scale of pain & function
 - Lower score indicates poorer outcome
- Baseline imbalance
 - Better scores in acupuncture group

Example: acupuncture for shoulder pain

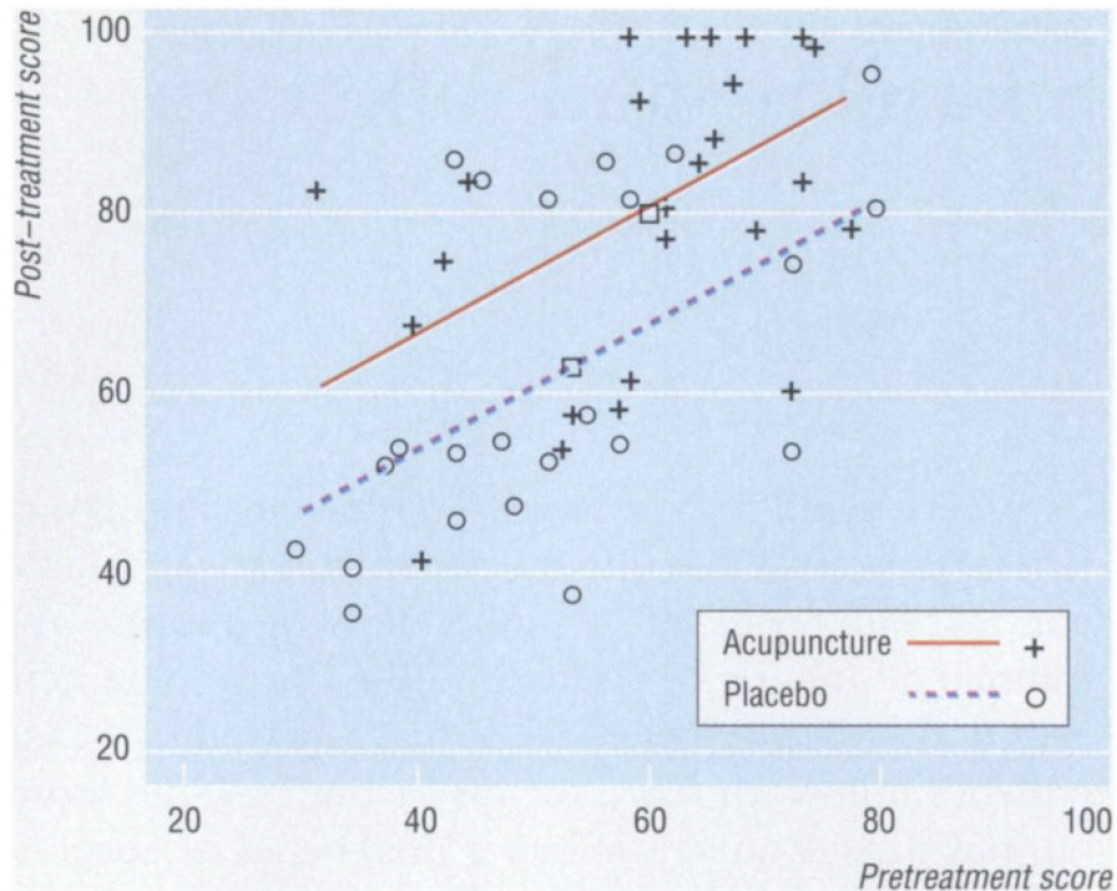
Pain scores, mean (SD)

Analysis		Placebo (<i>n</i> = 27)	Acupuncture (<i>n</i> = 25)	Difference between means	<i>P</i>
	Baseline	53.9 (14.0)	60.4 (12.3)	6.5	
	Follow-up	62.3 (17.9)	79.6 (17.1)	17.3 (7.5 to 27.1)	0.001
	Change score	8.4 (14.6)	19.2 (16.1)	10.8 (2.3 to 19.4)	0.014
	ANCOVA			12.7 (4.1 to 21.3)	0.005

Example: acupuncture for shoulder pain

- Imbalance between groups at baseline
- Baseline and change scores negatively correlated ($r = -0.25$)
- **ANCOVA**
 - $\text{Post} = 24 + 0.71 \times \text{baseline} + 12.7 \times \text{group}$
 - Pain and function score improved by 12.7 points more on average in the acupuncture group than the control
 - Also provides a means of prediction

Example: acupuncture for shoulder pain



Analysis of covariance (ANCOVA)

- **Assumptions**

- For any value of X , Y is normally distributed
- Variances of the groups should be equivalent
- Linear relationship between X and Y
- Direction and strength of this relationship must be similar in each group – parallel regression lines (homogeneity of regression slopes)

- **Benefits**

- Reduces within group error variance
- Elimination of confounders

Statistical power

Correlation	$\rho = 0.2$	$\rho = 0.35$	$\rho = 0.5$	$\rho = 0.65$	$\rho = 0.8$
Post	70.5%	70.5%	70.5%	70.5%	70.5%
% Change	45.1%	56.4%	67.0%	82.7%	97.1%
Change	50.7%	59.2%	70.5%	84.8%	97.7%
ANCOVA	72.3%	76.1%	82.3%	90.8%	98.6%

Statistical power

- ANCOVA has the highest statistical power
- Change from baseline has acceptable power when correlation between baseline and post treatment scores is high ($r > 0.8$)
- When correlation is low, analysing only post-treatment scores reasonable
- % change from baseline has lowest statistical power
 - when range of baseline values is large, variance increases disproportionately and power falls

Summary of reduced data

- Conceptually easy to deal with
- Greatly simplifies the analysis
 - Can use approaches already familiar with
- Valid approach but inefficient
- Why record the other measurements if they are not to be considered in the analysis?

Where to find more information?

- Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Med Res Methodol* 2001;1:6.
- Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ* 2001;323:1123–4.



Next session at
11:00

Summary measures

Summary measures

- Uses all observations collected over time on an individual (profile)
- Summarises each individuals profile with a single value
- Summary measures are then analysed in the same way as if they were the original observations
- Relies on the ability to choose summary measures of clinical relevance

Summary measures for longitudinal data

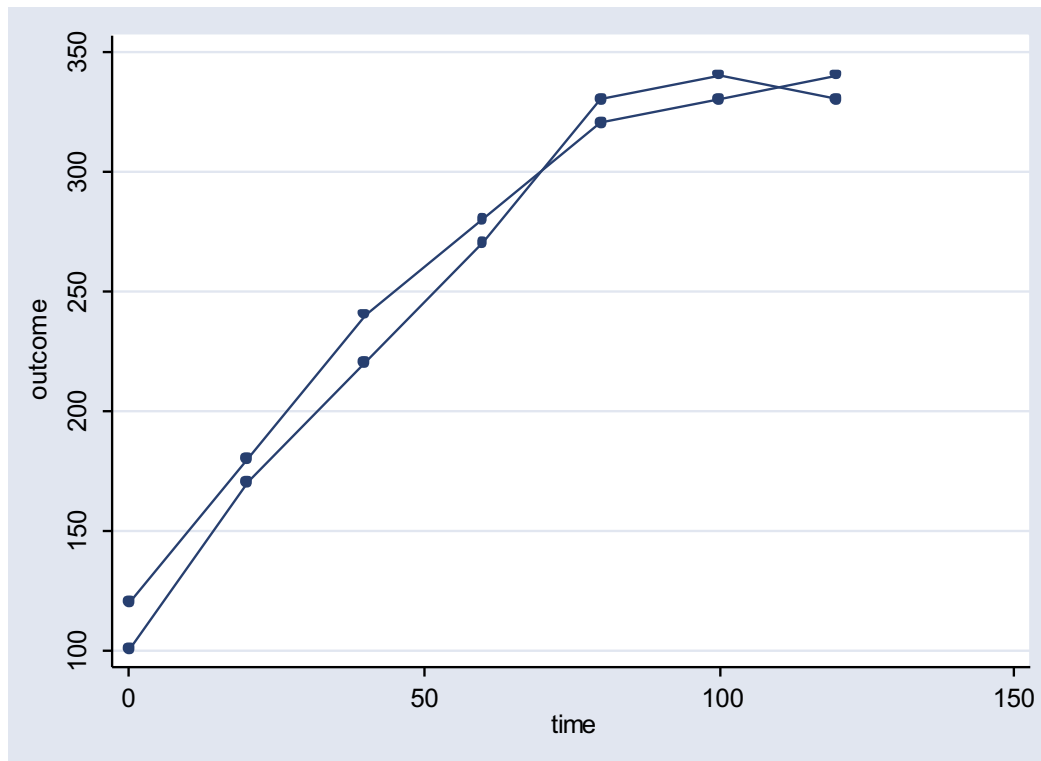
- Post-treatment mean
- Mean change
- Mean change relative to baseline
- Slope
- Area under the curve (AUC)
- Maximum value
- Time to reach a peak/value

Graphical display

- Before performing any form of statistical analysis it is advisable to graph the data in some way
- There are a number of ways of doing this:
 - Plot means by treatment group for each time point
 - Profile plots of individuals
 - Profile plots of individuals by treatment group
 - Individual profile plots

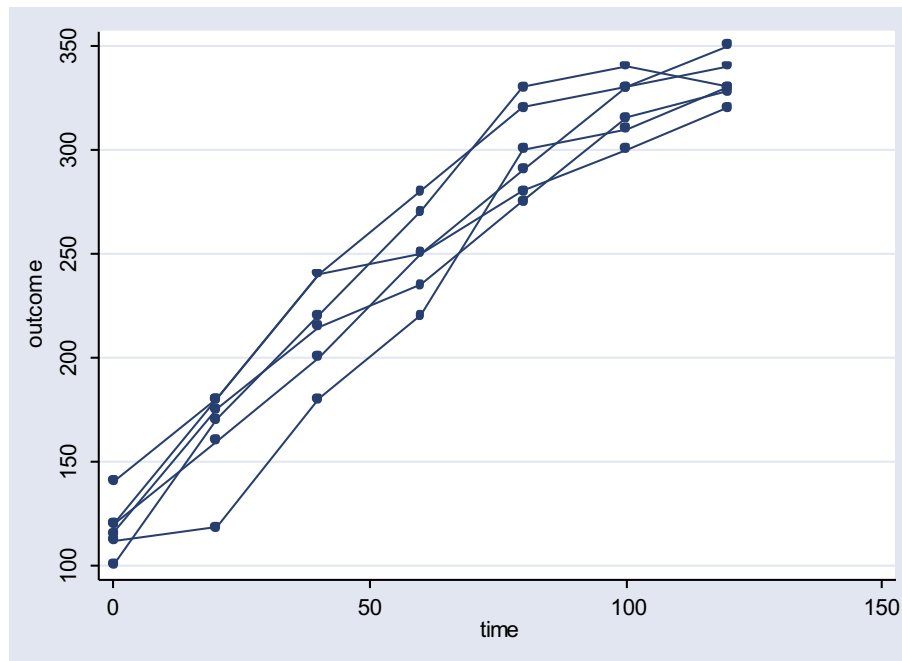
Plot means by treatment group

- Calculate the mean of each treatment group at each time point

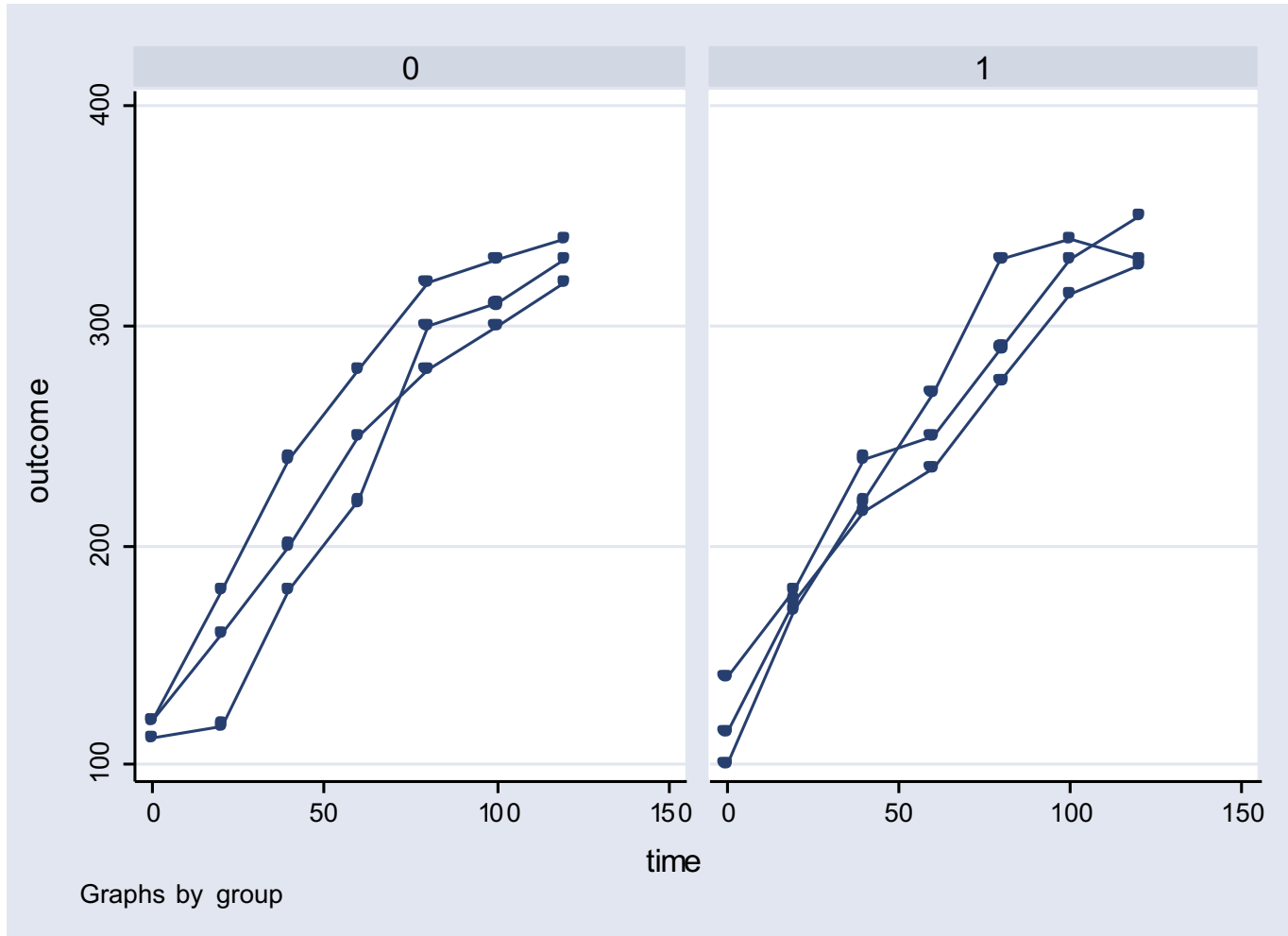


Profile plots of individuals

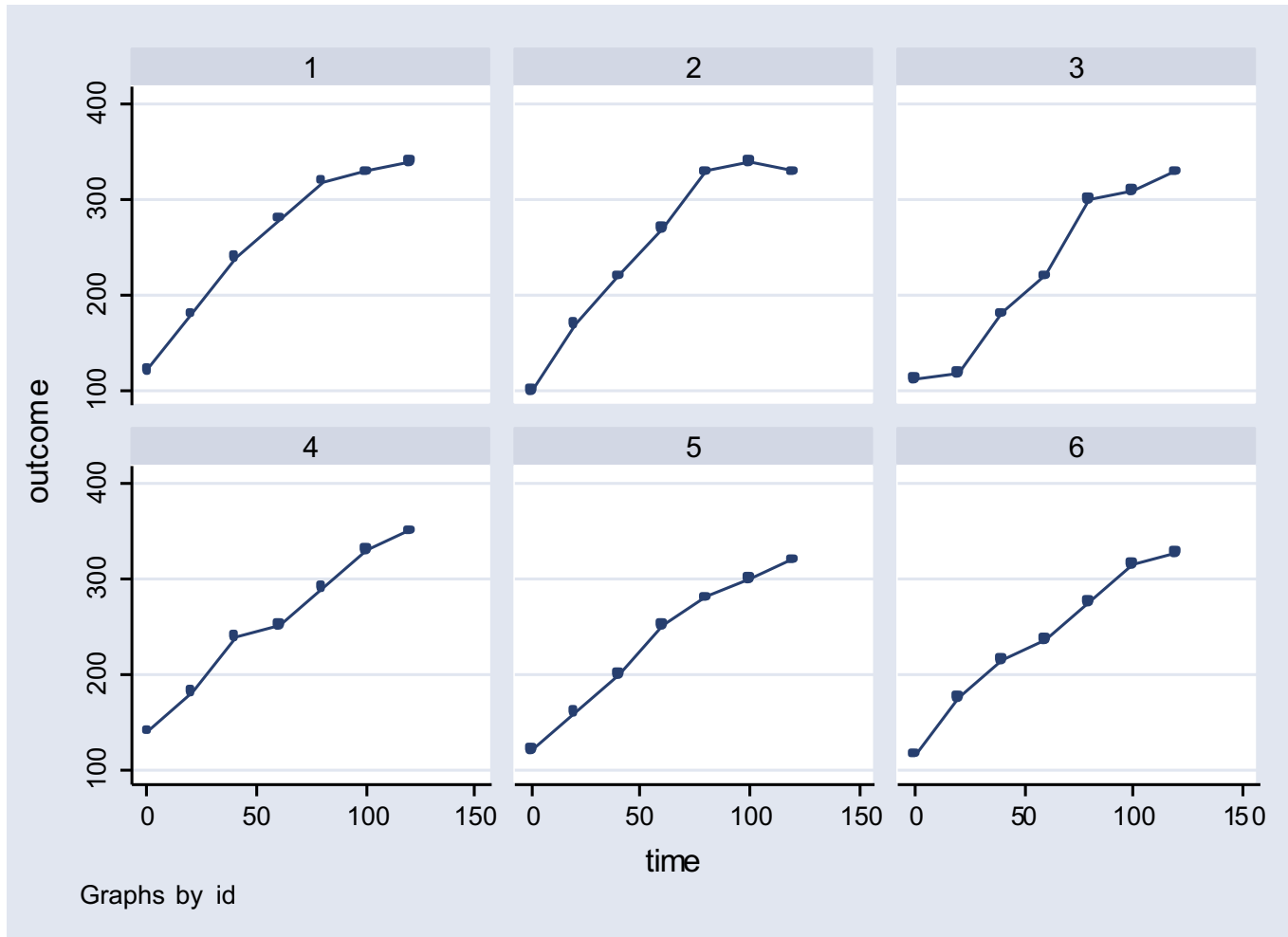
- Trellis plot
- Can be difficult to see a pattern if the graph is too crowded
- Option to select a random sample



Profile plots of individuals by treatment group



Individual profile plots



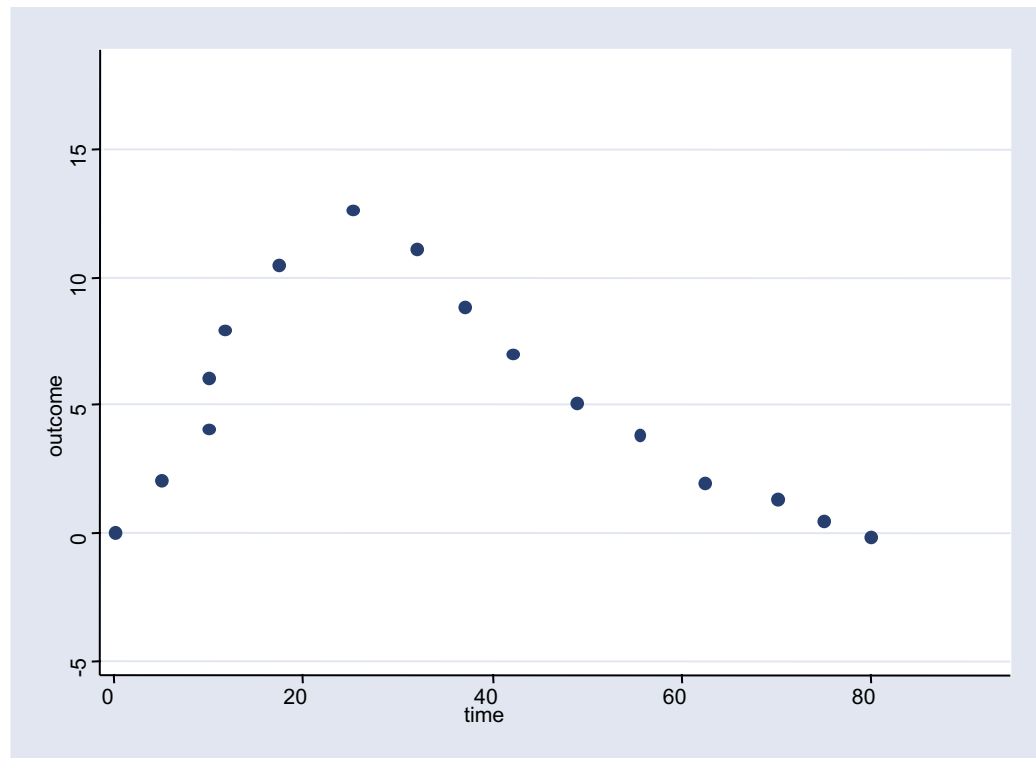
Types of time dependency

- Two main ways in which the outcome variable changes with time
 - Peaked
 - Growth
- Influence choice of summary measure

Types of time dependency

Peaked

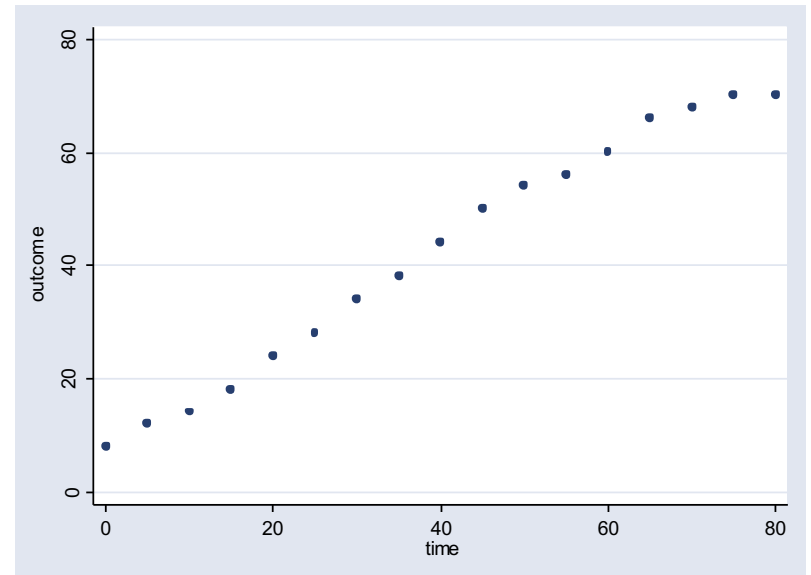
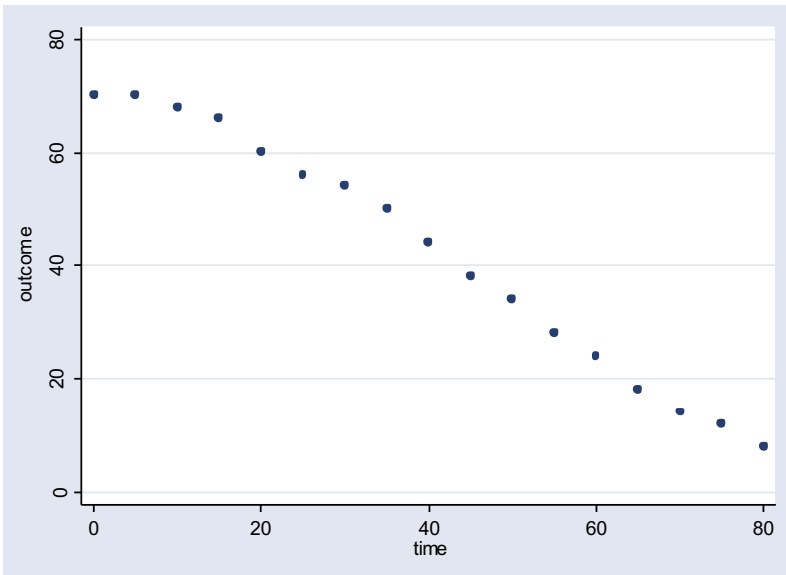
The outcome variable starts from baseline, rises to a peak and then returns to baseline



Types of time dependency

Growth

The outcome variable steadily increases or decreases with time and does not return to baseline over the period of the study



Post-treatment means

- Calculate the mean of each person's post treatment observations
- Single value per person → analyse as before

ID	Tx	T1	T2	T3	T4	Mean
1	A	10	9	12	10	$(10 + 9 + 12 + 10) / 4 = 10.25$
2	A	13	12	9	11	$(13 + 12 + 9 + 11) / 4 = 11.25$
...
40	B	8	8	9	8	$(8 + 8 + 9 + 8) / 4 = 8.25$

Example: depression

- RCT of mood stabilising drugs
- Post treatment assessment of HRQOL (Health-Related Quality Of Life) every month for 4 months
- Initial rise in QOL expected but clinical interest is in the plateau level
 - Calculate the post treatment mean for each person
 - Calculate the difference in post treatment means for group A vs group B

Missing data

- How could missing data effect the results?
- Need to investigate reasons for missing data
- Is missing data informative?

Post-treatment means missing data

- In the previous example missing data could have been caused by a person having a low mood, and therefore you would expect that the QOL measurement to be low
- This could then lead to an artificially high average for that individual

Mean change

- Similar approach to post-treatment means but allow for baseline measure(s)

Pre-treatment						Post-treatment				
ID	Tx	T1	T2	T3	Mean	T4	T5	T6	Mean	Change
1	A	50	55	52	52.3	78	82	82	80.7	28.4
2	A	46	44	44	44.7	62	65	66	64.3	19.6
3	A	47	42	44	44.0	68	67	66	67.0	23.0
...
30	B	39	41	38	39.3	56	58	59	57.6	18.3

Example: depression

- RCT of mood stabilising drugs
- Pre Tx assessment of HRQOL every month for 6 months
- Post Tx assessment of HRQOL every month for 4 months
- Clinical interest is in the change in HRQOL
 1. Calculate the post Tx mean for each person, \bar{x}_{ij}^{post}
 2. Calculate the pre Tx mean for each person \bar{x}_{ij}^{pre}
 3. For each person calculate post treatment mean-pre treatment mean, $\bar{x}_{ij}^{post} - \bar{x}_{ij}^{pre}$
 4. Single value per person
 5. Analysis options as for single time point

Mean change relative to baseline

- This summary statistic is exactly the same as the previous one that we just calculated except that we now divide the mean change by the baseline measurement

$$\frac{\bar{x}_{ij.}^{post} - \bar{x}_{ij.}^{pre}}{\bar{x}_{ij.}^{pre}}$$

- Measurement can be expressed as a percentage
- Used when the degree of change is correlated with baseline measure

Example: depression

- Expected that the amount of change will be related to baseline measure
- Clinical interest is in the change in HRQOL relative to baseline
 - Calculate the post Tx mean for each person
 - Calculate the pre Tx mean for each person
 - For each person calculate post-treatment mean – pre-treatment mean and divide by the pre-treatment mean
 - Single value per person
 - Analysis options as for single time point

Missing data: mean change

- How could missing data effect the results of mean change and change relative to baseline?
- Similar results to that of missing data with post treatment means
- Additional concerns over missing data in pre measurements

Slope

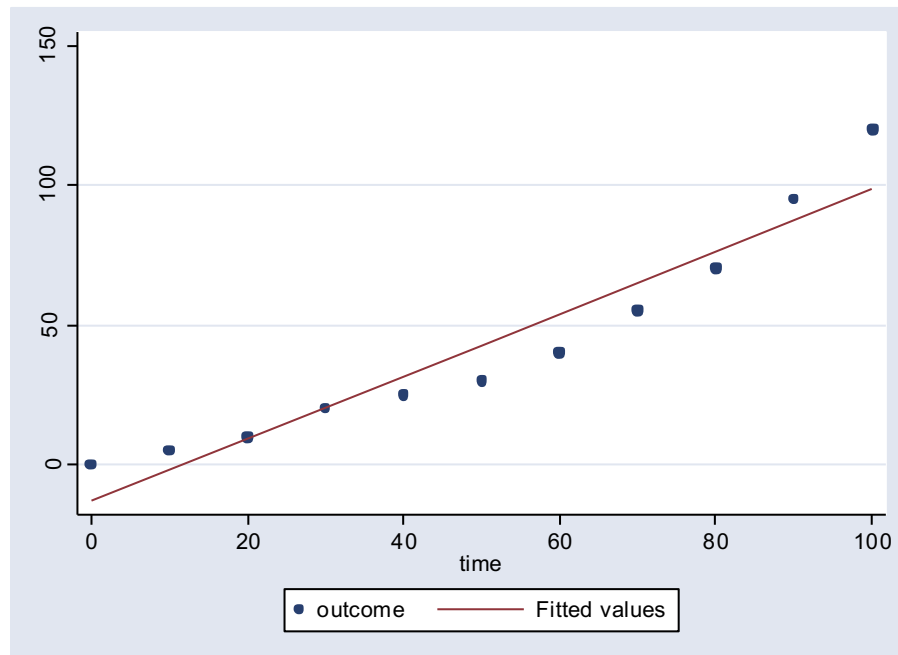
- If we use a separate univariate **linear** regression for each participants set of results:

$$\begin{array}{ccccc} \text{outcome} & & & \text{time} & \\ \downarrow & & & \downarrow & \\ y_j = \alpha + \beta x_j + \varepsilon_j & & & & \\ \uparrow & \uparrow & & \uparrow & \\ \text{intercept} & \text{slope} & & \text{error} & \end{array}$$

- There are as many estimates of α as subjects
- The summary statistic is the coefficient that is obtained from the fitted model, β
 - Represents the rate of change per unit of time

Slope

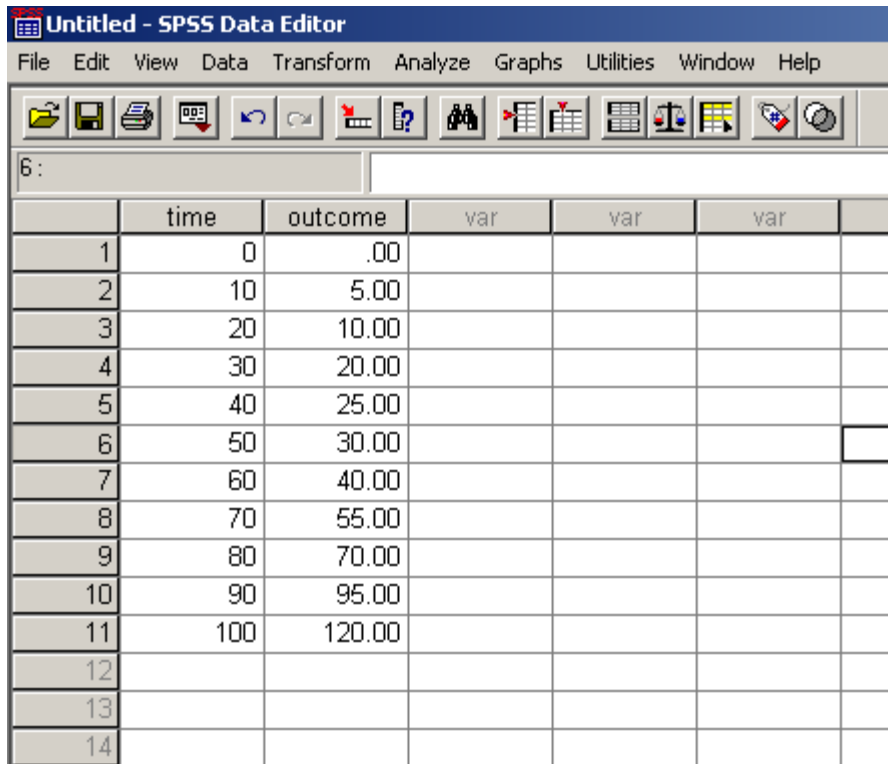
Participants results from a treatment over a period of time are measured, to obtain the slope coefficient we fit a univariate linear model with time as the independent variable



Example: non-small cell lung cancer

- Patients with progressive disease
- Standard therapy is palliative rather than curative
- 4 HRQOL assessments over 6 months using FACT-L
- Steady-rate of change over time expected, reflecting a constant decline in HRQOL

Example



6 :

	time	outcome	var	var	var	
1	0	.00				
2	10	5.00				
3	20	10.00				
4	30	20.00				
5	40	25.00				
6	50	30.00				
7	60	40.00				
8	70	55.00				
9	80	70.00				
10	90	95.00				
11	100	120.00				
12						
13						
14						

The data is entered in columns and a linear regression is fit for each person's set of results

Untitled - SPSS Data Editor

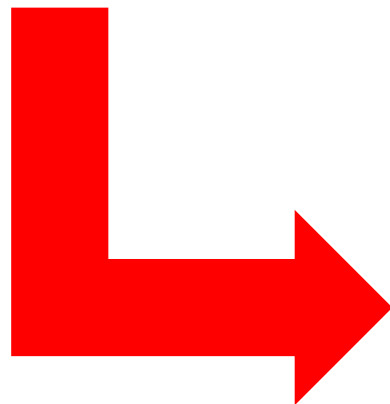
File Edit View Data Transform Analyze Graphs Utilities Window Help

6:

	time	outcome
1	0	.0
2	10	5.0
3	20	10.0
4	30	20.0
5	40	25.0
6	50	30.0
7	60	40.0
8	70	55.0
9	80	70.0
10	90	95.0
11	100	120.0
12		
13		
14		

Analyze

- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Mixed Models
- Correlate
- Regression**
 - Linear...**
 - Curve Estimation...
 - Binary Logistic...
 - Multinomial Logistic...
 - Ordinal...
 - Probit...
 - Nonlinear...
 - Weight Estimation...
 - 2-Stage Least Squares
 - Optimal Scaling...
- Loglinear
- Classify
- Data Reduction
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response
- Missing Value Analysis...
- Complex Samples



Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

6:

Linear Regression

Dependent: outcome

Block 1 of 1

Previous Next

Independent(s): time

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Statistics... Plots... Save... Options...

OK Paste Reset Cancel Help

Results

Coefficients^a

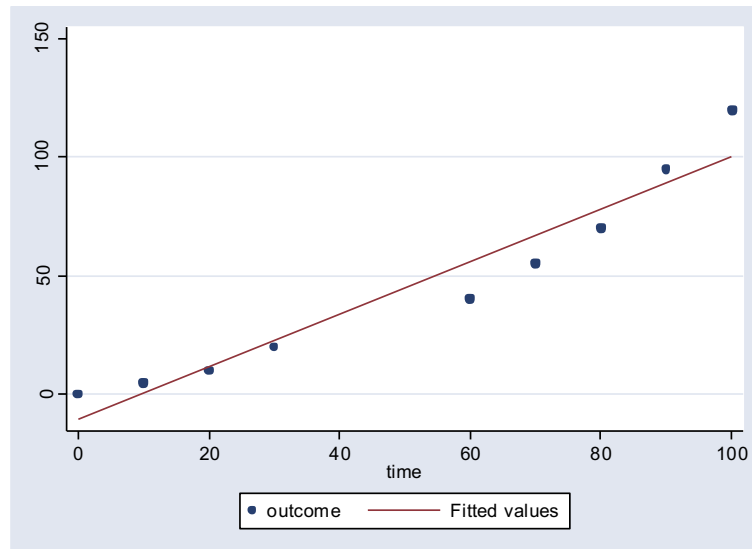
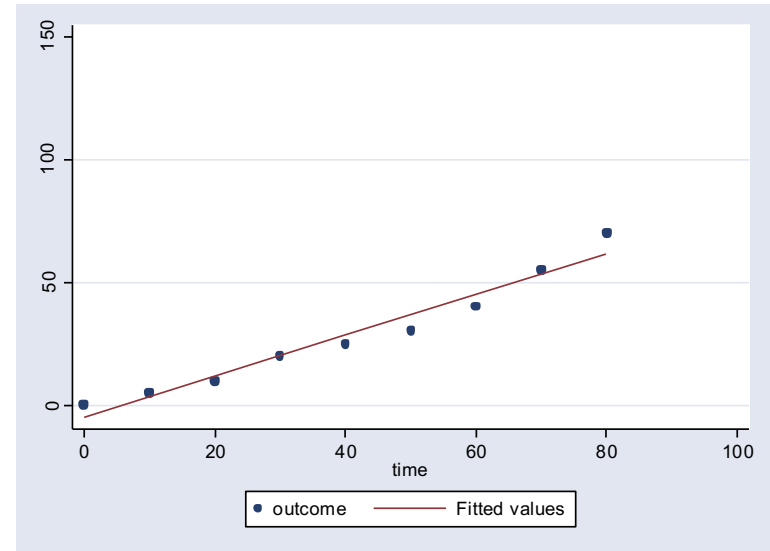
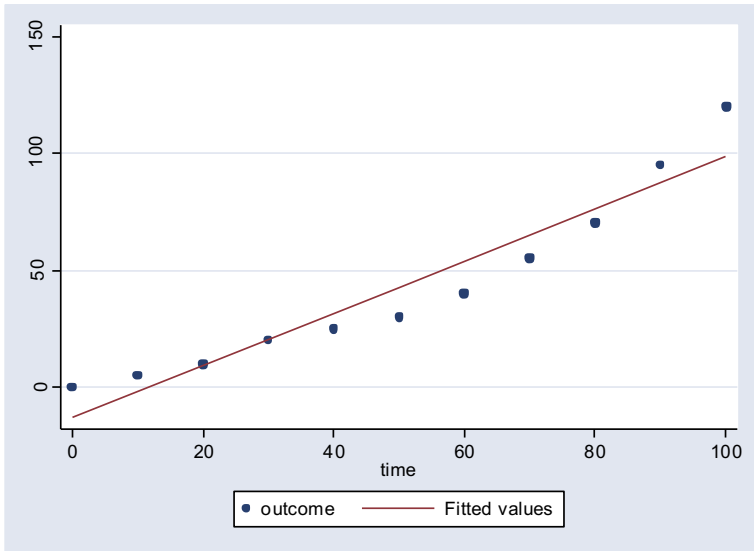
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-12.955	6.717		-1.929	.086
time	1.114	.114	.956	9.809	.000

a. Dependent Variable: outcome

- The coefficient for time (B) is the summary statistic for this individual
- In this example the coefficient for time is 1.11

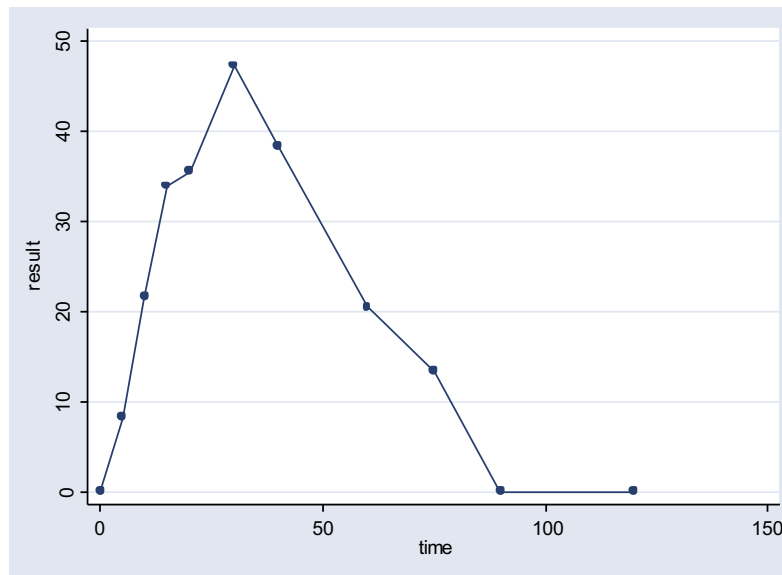
Missing data

- How could missing data effect the estimate?
- The summary statistic for the data shown (the slope in the linear regression model) in the previous graph is 1.11
- What if the data that was missing was the final two time points?
- The summary statistic is now 0.83
- If the two data points were missing from the middle, the the summary statistic is 1.11



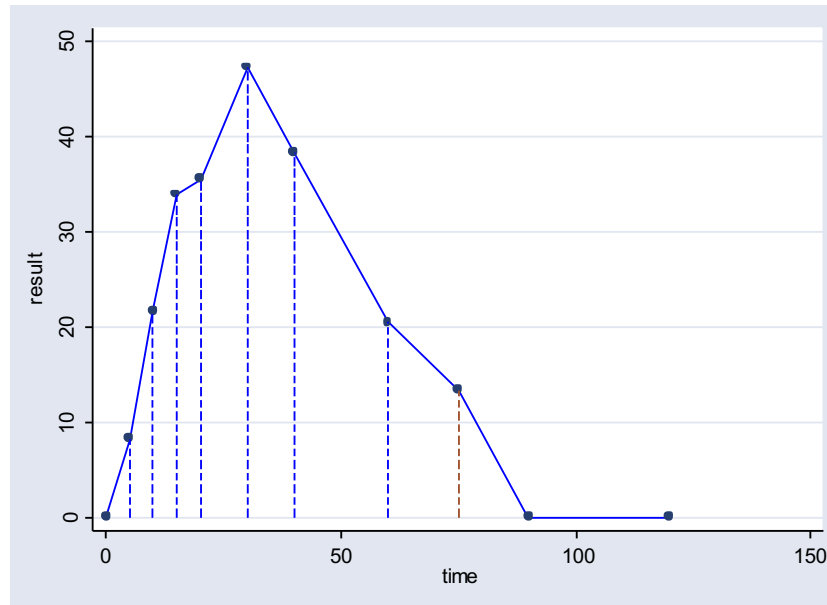
Area under the curve

- Can be thought of as representing a cumulative response to treatment
- We have the following data from a patient in a trial:
 - At times 0, 5, 15, 20, 30, 40, 60, 75, 90, and 120 minutes the result of a test was 0, 8.3, 21.6, 33.9, 35.5, 47.2, 38.3, 20.5, 13.3, and 0, respectively
- A plot of this individuals results looks like this



Area under the curve

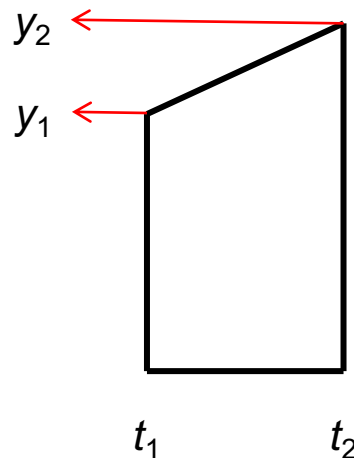
- If we were to draw vertical lines down from each of the time points that the data was measured at we would obtain a graph that looked like the following



- Each segment of the graph is the same shape as a trapezium and if we add up all of the areas of the segments, we obtain the area under the curve

Area under the curve

- AUC is calculated using the trapezium rule
- Each segment looks like



- Where we have measurements y_1 and y_2 at times t_1 and t_2

Area under the curve

- The area under the curve between these two times is the difference between the time points multiplied by the average of the two measurements

$$(t_2 - t_1) \times (y_1 + y_2) / 2$$

- If we have $n + 1$ measurements of y_i at times t_i ($i = 0, \dots, n$), then the area under the curve (AUC) is calculated as:

$$\text{AUC} = \frac{1}{2} \sum_{i=0}^{n-1} (t_{i+1} - t_i)(y_i + y_{i+1})$$

Area under the curve

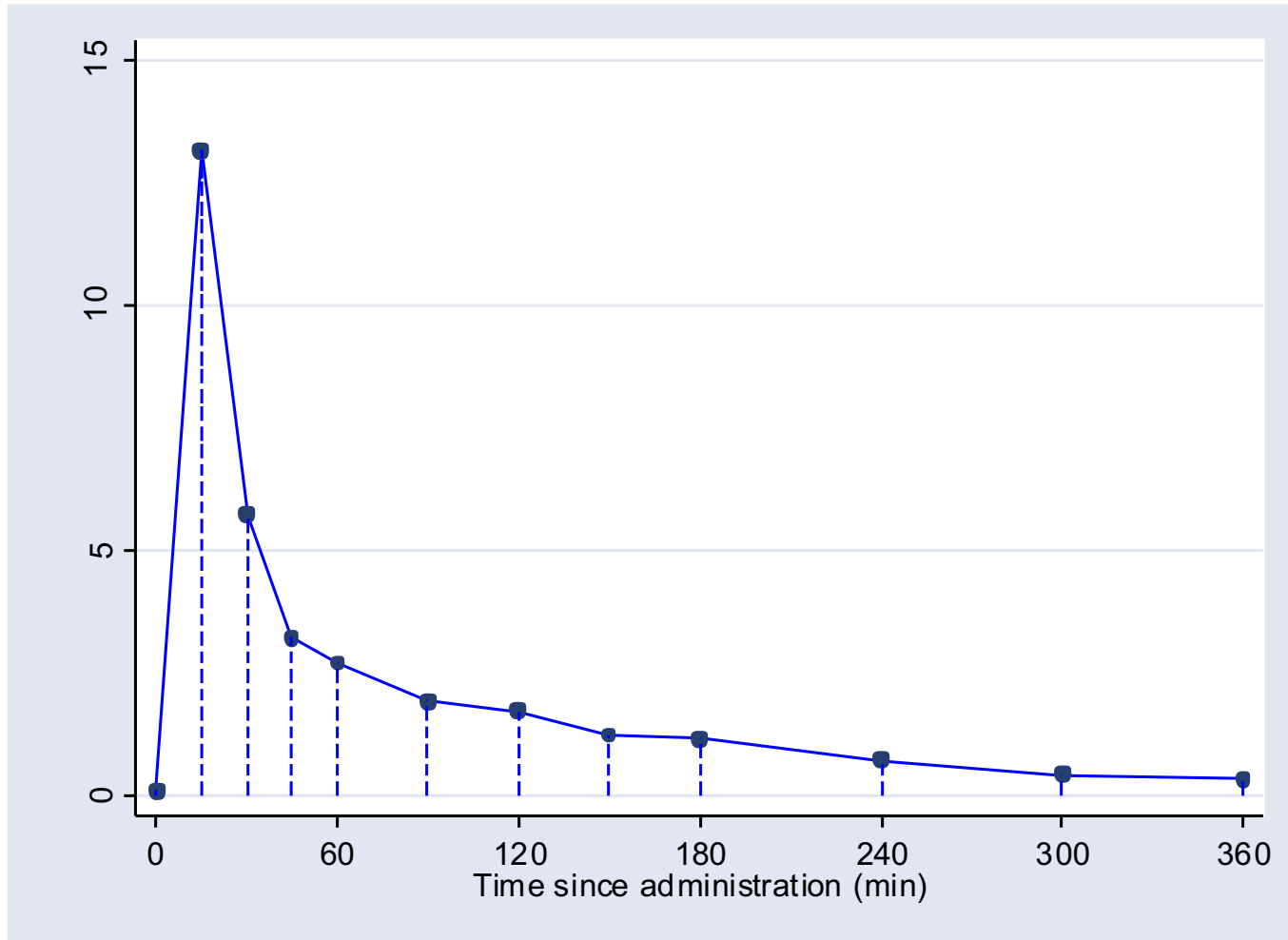
From the data that we looked at earlier in this example we have:

$$\begin{aligned} AUC = & \left\{ (5 - 0) \times \frac{(8.3 + 0)}{2} \right\} + \left\{ (10 - 5) \times \frac{(8.3 + 21.6)}{2} \right\} + \dots\dots\dots \\ & \dots\dots\dots + \left\{ (90 - 75) \times \frac{13.3}{2} \right\} + \left\{ (120 - 90) \times \frac{0 + 0}{2} \right\} = 2190 \end{aligned}$$

If we standardise by the length of the study, 120 minutes, we get $2190 / 120 = 18.25$

Example: AUC

- Investigating levels of AZT in the blood of AIDS patients at several time points after administration of a drug
- Is there a difference between patients with normal fat absorption and those with fat malabsorption?
- We are able to use AUC to answer this question as we expect the levels of AZT to rise initially after administration of drug then return to baseline value



Calculation of the area under the curve for one subject

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help Acrobat

10 Arial

SUM X ✓ = (A2-A1)*(B1+B2)/2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0	0.08	141.375												
2	15	13.15	141.375												
3	30	5.7	66.9												
4	45	3.22	44.325												
5	60	2.69	69												
6	90	1.91	54.45												
7	120	1.72	44.1												
8	150	1.22	35.55												
9	180	1.15	56.8												
10	240	0.71	34.2												
11	300	0.43	22.5												
12	360	0.32	667.425												

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

F6 =

	A	B	C	D	E
1	0	0.08	99.225		
2	15	13.15	141.375		
3	30	5.7	66.9		
4	45	3.22	44.325		
5	60	2.69	69		
6	90	1.91	54.45		
7	120	1.72	44.1		
8	150	1.22	35.55		
9	180	1.15	56.8		
10	240	0.71	34.2		
11	300	0.43	22.5		
12	360	0.32	667.425		

Sheet1 / Sheet2 / Sheet3

W2KS Poster Viewer Mulberry (Connected) INBOX

The screenshot displays the Microsoft Excel interface. The title bar reads "Microsoft Excel - Book1". The menu bar includes File, Edit, View, Insert, Format, Tools, Data, Window, Help, and Acrobat. The toolbar contains various icons for file operations, editing, and formatting. The formula bar at the top shows "F6 =".

The spreadsheet has columns labeled A through O and rows numbered 1 through 34. The following table represents the data visible in the spreadsheet:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		0	0.08	99.225											
2		15	13.15	141.375											
3		30	5.7	66.9											
4		45	3.22	44.325											
5		60	2.69	69											
6		90	1.91	54.45											
7		120	1.72	44.1											
8		150	1.22	35.55											
9		180	1.15	55.8											
10		240	0.71	34.2											
11		300	0.43	22.5											
12		360	0.32												
13		AUC=		667.425											
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															
28															
29															
30															
31															
32															
33															
34															

The status bar at the bottom indicates "Ready" and shows the active sheet as "Sheet1 / Sheet2 / Sheet3". The taskbar at the very bottom shows several open applications, including WZKS Post..., Mulberry (...), INBOX, Microsoft ..., and Microsoft P... along with system icons and the time 09:30.

- The AUC can be easily calculated using a spreadsheet
- The AUC for the first individual is 667.425
- For each individual the AUC should be calculated for their results

Example: AUC

Area under the curve for malabsorption data:

Malabsorption patients		Normal patients
667.425	256.275	919.875
569.625	527.475	599.850
306.000	388.800	499.500
298.200	505.875	472.875
617.850		1377.975

Example: AUC

- Data can then be analysed using two-sample *t*-test

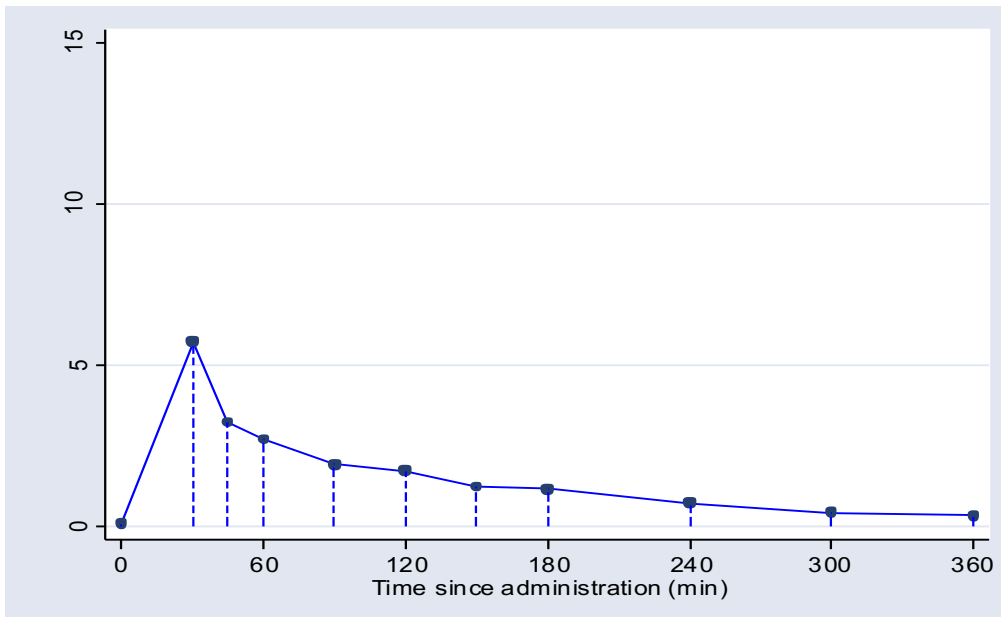
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
AUC	Equal variances assumed	6.826	.023	-2.231	12	.046	-314.29000	140.87732	-621.235	-7.344694
	Equal variances not assumed			-1.766	4.711	.141	-314.29000	177.94338	-780.275	151.6952

- The difference in the Areas under the curves is 314.29 with 95% confidence interval (-780.275 to 151.69)

Missing data

- How would missing data effect the results of the analysis using AUC?
- Need to look at where the missing data is
- If data is missing from what would appear to be the peak, then this could seriously under-estimate the AUC



The AUC is now
513.525 as apposed to
667.425 when we
included all the data

Maximum or minimum value

- Self explanatory
 - For each individual use the max/min observed value as the summary measure
 - Analyse using that value as for single observation per person
- Example: trial objective is to reduce toxicity and maintain acceptable HRQOL
 - Might want to study the worst (min) HRQOL score that occurred for each individual

Time ...

- To reach max value
- To reach a given level*
- To change by a given amount*
- Above a given level*
- To achieve max change from baseline
- To return (near) to baseline level*

* Does the event occur for each person observed?

Choosing a summary measure

Choosing a summary measure

Type of data	Question to be answered	Summary Measure
Peaked	Is the overall outcome variable the same in different groups?	Overall mean (equal time intervals) Area under the curve (unequal time intervals)
Peaked	Is the maximum (minimum) response different between groups?	Maximum (minimum) Value
Peaked	Is the time to maximum (minimum) response different between groups?	Time to maximum (minimum) response
Growth	Is the rate of change of the outcome variable different between groups?	Regression coefficient
Growth	Is the eventual value of the outcome variable the same between groups?	Final value of outcome measure or difference between last and first values, or percentage change between first and last

Summary of summary measures

- Choice of summary measure should be specified in advance
 - Clear clinical relevance
- Analysis uses familiar univariate approaches
- May facilitate interpretation
 - Rate of change and AUC are familiar concepts in medicine
- Consideration must be given to missing data and its potential impact on your choice of summary measure

Where to find more information?

- Omar RZ, Wright EM, Turner RM, Thompson SG. Analysing repeated measurements data: a practical comparison of methods. *Stat Med* 1999;18:1587–603.
- Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *BMJ* 1990;300:230–5.



Next session at
13:00

Extending ANOVA

Extending ANOVA

- Repeated measures ANOVA (RM-ANOVA)
- Multivariate ANOVA (MANOVA)

Why?

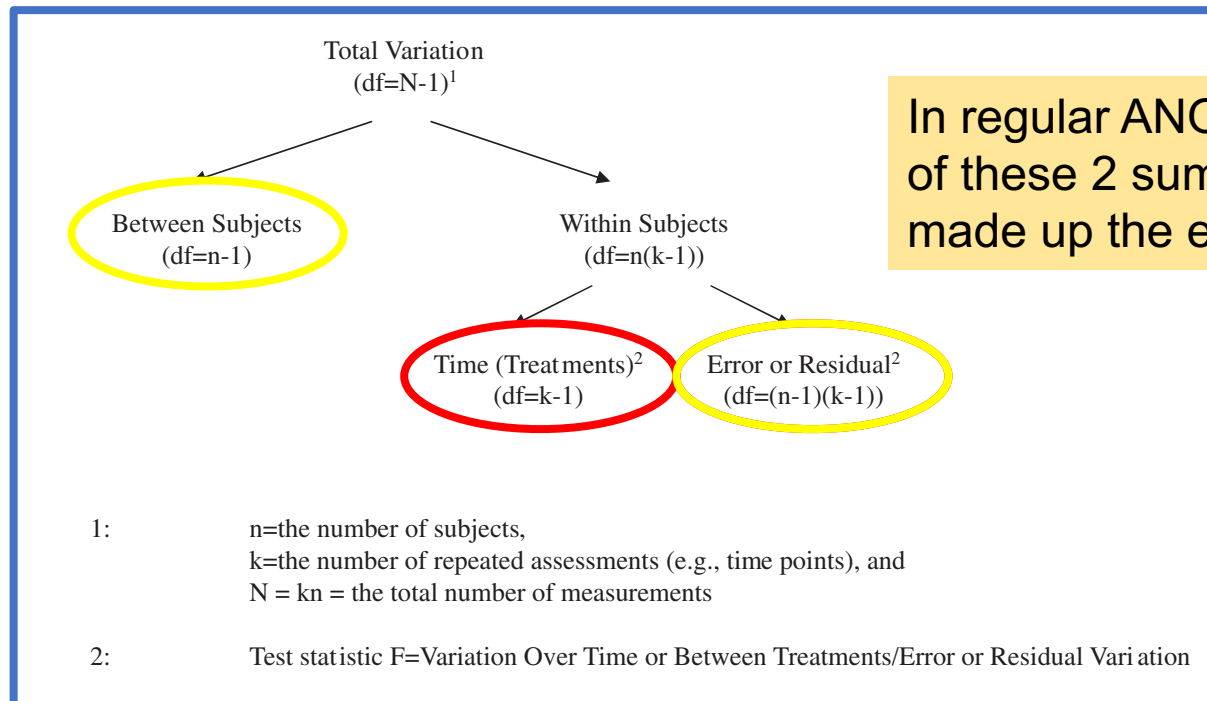
- Include data at all time points
- No data reduction
- Simple to implement

Repeated measures ANOVA

- Approach just like simple ANOVA
- Important distinction is that we have repeated measurements on each individual
- Partition variance across multiple factors in order to
 - Investigate possible effects and interactions
 - Reduce corresponding error terms and thereby increase the power of the test

Repeated measures ANOVA

- One-factor repeated-measures ANOVA when only considering time
- Partitioning of total variance (below)



In regular ANOVA, the sum of these 2 sum-of-squares made up the error term

Example

- Placebo controlled RCT to examine effect of a treatment in anaemic pregnant women
- Outcome: haemoglobin level (g/dl)
- Times of measurement:
 - At recruitment (first antenatal clinic visit)
 - At 4 follow-up visits

Repeated measures ANOVA

Baseline Post treatment follow-up

ID	Group	T0	T1	T2	T3	T4
1	1	11.3	11.4	12.0		
2	1	9.4	7.5	7.7		8.5
3	2	12.1	11.2	11.1		10.0
4	2	11.3	10.3	10.9		
5	2	8.9	9.6	6.9		12.9
6	1	8.3	8.1	9.5	11.5	
7	2	9.4	11.8	12.1	13.1	
8	2	10.9	9.8	10.4	11.3	12.0
9	2	9.8	11.7	12.2	13.4	
10	1	10.0	10.3	10.4	12.3	

Repeated measures ANOVA

- Two-factor repeated-measures ANOVA with repeated measures on 1 factor*
 1. Treatment/control as a between subjects factor
 2. Time as a within-subjects factor
- Baseline measurements should not be treated as another level of the time factor
 - They should be used as a covariate making the analysis a *repeated-measures analysis of covariance*

* Sometimes called a mixed ANOVA

Repeated measures ANOVA

Related to regression: the model is written in the form of “grand mean plus various effects”

The diagram shows the equation $y_{ijk} = \mu + g_i + t_j + (gt)_{ij} + U_{ik} + \varepsilon_{ijk}$ with arrows pointing from descriptive labels to each term. The labels are: 'observation' points to y_{ijk} ; 'group effect' points to g_i ; 'interaction effect' points to $(gt)_{ij}$; 'measurement error' points to ε_{ijk} ; 'mean' points to μ ; 'time effect' points to t_j ; and 'random effect' points to U_{ik} .

$$y_{ijk} = \mu + g_i + t_j + (gt)_{ij} + U_{ik} + \varepsilon_{ijk}$$

Treatment $i = 1, \dots, s$; Follow-up visit $j = 1, \dots, T$, Subject $k = 1, \dots, N$

Repeated measures ANOVA

Main effects

- Group

Is there an overall difference between groups?

- Time

Is there change over time?

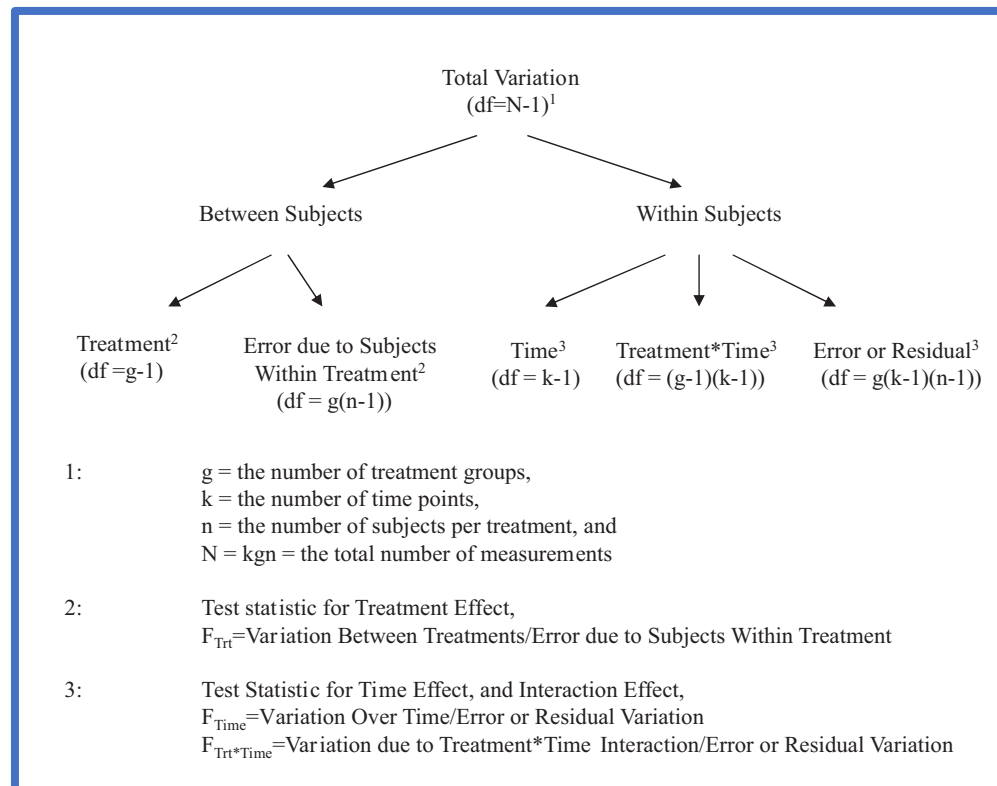
Interactions

- Group * Time

Is there any significant difference in pattern of change over time between the groups?

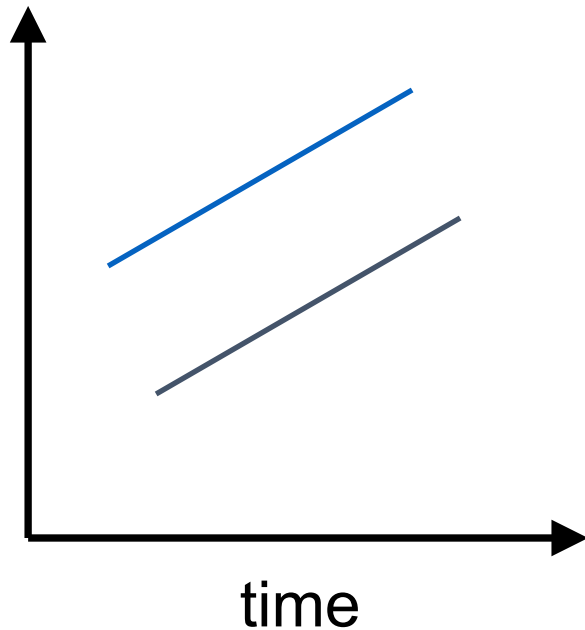
Repeated measures ANOVA

- Partitioning of total variance to evaluate each of the 3 terms

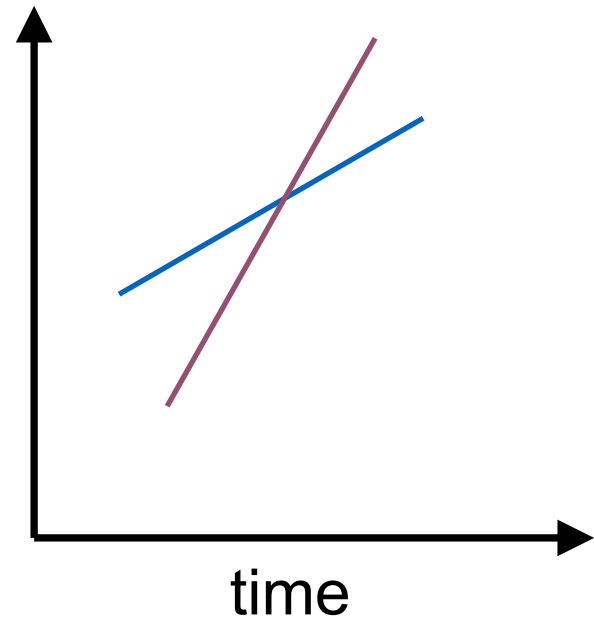


Repeated measures ANOVA

Without the interaction: change over time is the same in the two groups



With the interaction: change over time is different between two groups



Repeated measures ANOVA

- **Assumptions**

- Data are normally distributed at each time point
- Measurements on each subject independent of those on other subjects
- Homogeneity of variance
- Sphericity of the variances
- Correlation of the data within subjects

- Assumptions need to be met for the repeated measures ANOVA to be valid

Sphericity of the variances

- Making an assumption about how the repeated measures within an individual are related
- Refers to the equality of variances of the differences *between* all possible times

$$s_{t_1-t_2}^2 \approx s_{t_1-t_3}^2 \approx s_{t_1-t_4}^2 \approx s_{t_2-t_3}^2 \approx s_{t_2-t_4}^2 \approx s_{t_3-t_4}^2$$

Test sphericity of variance

Mauchly's test

- If Mauchly's test is **significant** we need to be concerned about the unadjusted P -values reported for the analysis
- Test has low power particularly when data are non-normal
- Even if test is **not statistically significant**, doubt sphericity of variance

Measure of sphericity: ε (epsilon)

- Estimates adjustment to degrees of freedom to account for non-sphericity
- The closer ε is to 1, the less the departure from sphericity
- Two common estimates of ε used:
 1. Greenhouse-Geisser
 2. Huynh-Feldt
- Corrects for violation of assumption of sphericity

Epsilon

- Correcting for violation of Sphericity
 - Does not alter the value of the F -statistic
 - Multiply the degrees of freedom by the value of epsilon
 - This alters the location you obtain the P -value from in the F -tables, ultimately correcting for the increased probability of a false positive result

Example

Mauchly's Test of Sphericity^b

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^a		
					Greenhous e-Geisser	Huynh- Feldt	Lower- bound
TIME	.934	24.279	5	.000	.957	.968	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

- a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.
- b.

Design: Intercept+RXGRP
Within Subjects Design: TIME

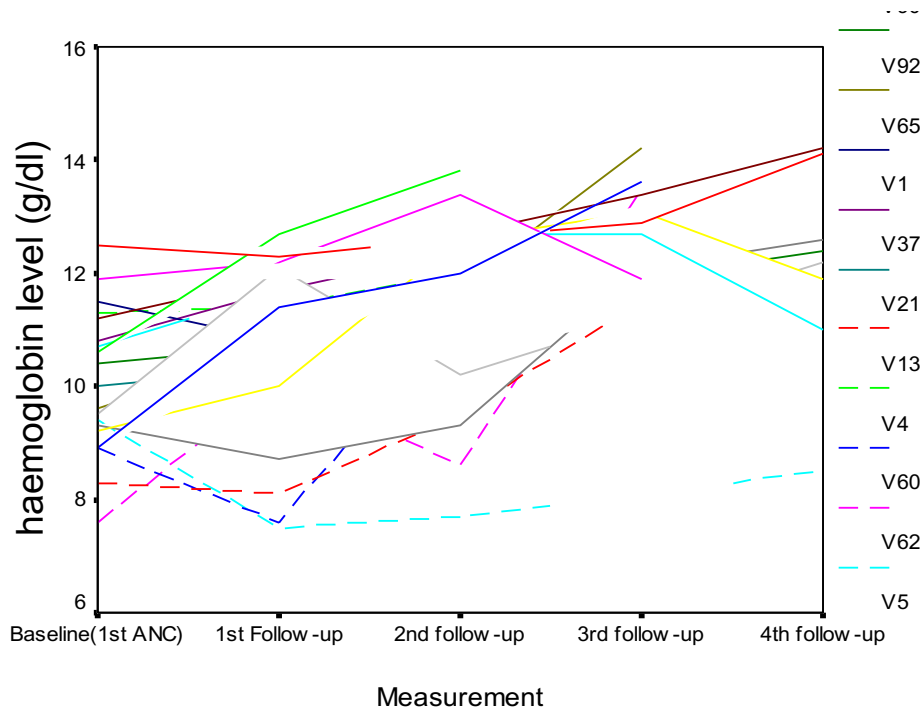
Using the Greenhouse-Geisser correction:

F -test had 3,1074 degrees of freedom, this is corrected as an F -test on $(3*0.957), (1074*0.957)$ degrees of freedom

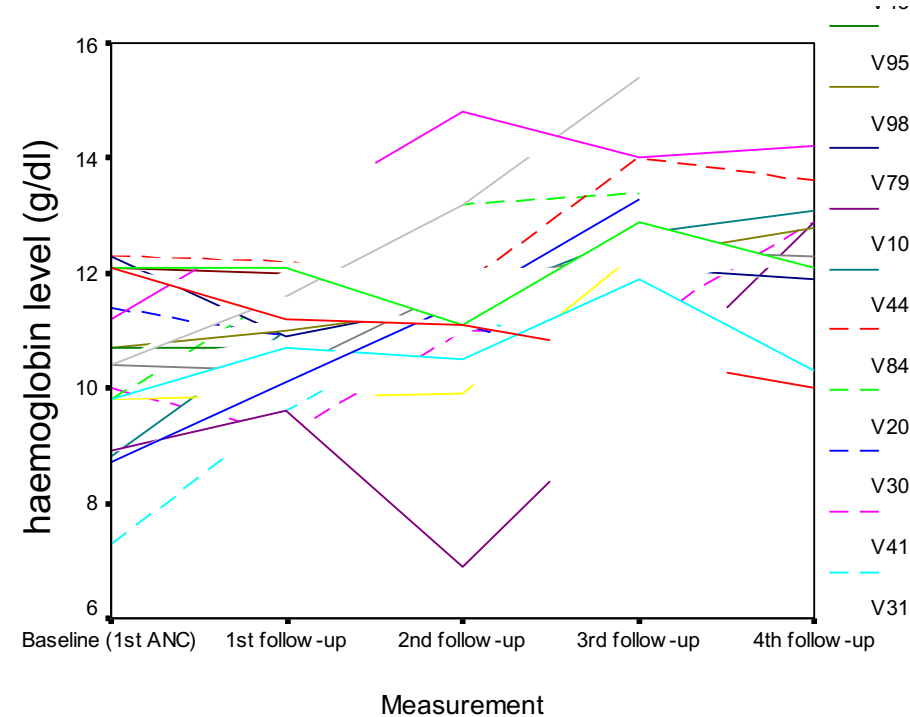
Which ε to use?

- Average of the Greenhouse-Geisser and Huynh-Feldt ε
- Alternative is we use what is known about ε in certain conditions
 - If Greenhouse-Geisser $\varepsilon < 0.75$, use this ε to correct the analysis
 - If Greenhouse-Geisser $\varepsilon \geq 0.75$, use the Huynh-Feldt ε

Longitudinal profiles

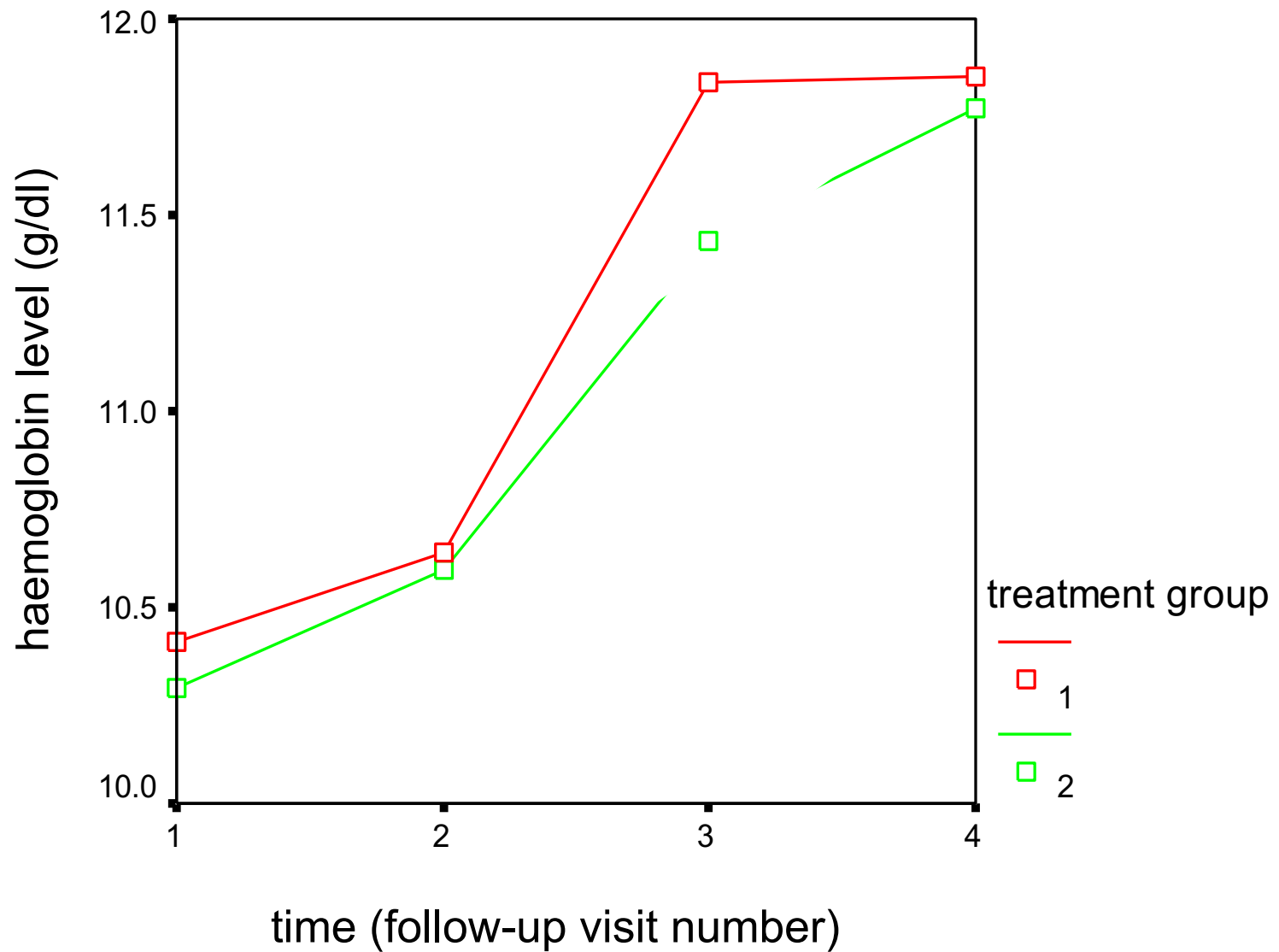


Treatment group (1)



Placebo group (2)

Estimated Marginal Means of MEASURE_1



Time effect and time x treatment interaction

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	497.519	3	165.840	156.5	.000
	Greenhouse-Geisser	497.519	2.870	173.350	156.5	.000
	Huynh-Feldt	497.519	2.904	171.336	156.5	.000
	Lower-bound	497.519	1.000	497.519	156.5	.000
TIME * RXGRP	Sphericity Assumed	6.274	3	2.091	1.973	.116
	Greenhouse-Geisser	6.274	2.870	2.186	1.973	.119
	Huynh-Feldt	6.274	2.904	2.161	1.973	.118
	Lower-bound	6.274	1.000	6.274	1.973	.161
Error (TIME)	Sphericity Assumed	1138.215	1074	1.060		
	Greenhouse-Geisser	1138.215	1027.5	1.108		
	Huynh-Feldt	1138.215	1039.5	1.095		
	Lower-bound	1138.215	358.00	3.179		

Group effect

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	38787.048	1	38787.0	231.34	.000
RXGRP	1.969	1	1.969	1.175	.279
Error	600.244	358	1.677		

Estimates

Measure: MEASURE_1

treatment group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	11.185	.120	10.948	11.421
2	11.026	.083	10.863	11.189

Missing data

- Repeated measures ANOVA cannot include cases with any missing data
- 700 patients, but only 360 with complete data

Where to find more information?

- Sullivan LM. Repeated measures. *Circulation* 2008;117:1238–43.
- Park E, Cho M, Ki CS. Correct use of repeated measures analysis of variance. *Korean J Lab Med* 2009;29:1–9.
- Maurissen JP, Vidmar TJ. Repeated-measure analyses: which one? A survey of statistical models and recommendations for reporting. *Neurotoxicol Teratol* 2016;59:78–84.
- Fitzmaurice GM, Ravichandran C. A primer in longitudinal data analysis. *Circulation* 2008;118:2005–10.

Missing data

Missing values

- Likely that there will be missing values
 - Must try to minimise occurrence in planning/design stage
- Why does missing data matter?
 - **Bias**

If the proportion of missing data is small then little bias will result
 - **Efficiency**
 - Loss of data will lead to an increase in variability

Missing data

Id	T1	T2	T3	T4	T5	T6	T7
1	2.83	1.15	2.34	2.84	0.36	3.37	3.01
2	8.70					0.55	
3	1.98	3.36	2.30	1.86	1.18	0.00	7.44
4	1.82	0.98	1.83	1.04	0.90	0.67	0.17
5	5.20						
6	3.93	0.39	1.05	1.91	0.17	0.13	0.67
7	2.81	1.06	2.10	1.60		1.73	0.09
8	1.80	0.95	2.45	2.00	1.32	8.86	2.45
9	3.59	1.02	0.78	0.52	na	0.00	0.15
10	3.20	4.30		4.10	0.80	0.00	2.30
11	1.20	1.26	1.25	0.78	1.62	0.17	6.03
12	10.29	1.62	1.53	0.00	0.59	8.12	2.21
13	2.70	4.56	2.13	3.19	2.43	1.60	1.16
14	3.40	2.00					
15	13.90	3.40	0.68	0.73	0.76	1.80	1.10
16	5.70	4.50	3.91	6.01	1.36	3.20	0.87
17	2.02	1.12	0.46	0.34	0.32	0.3	0
18	2.40	0.60	0.30	0.59	0.20	0.35	0.33
19	4.30	2.74		1.20	1.11	2.70	3.89
20	6.80	1.76	1.60	1.30	0.57	1.70	4.48
21	1.71			0.43	0.18	0.39	0.93

Patterns of missing data

- Consider a trial measuring FEV (Forced Expiratory Volume) at 3 time points
 - X denotes an observed value
 - ? denotes a missing value
- Summary of the different *patterns* of missing data

1. XXX

2. XX?

3. X?X

4. ?XX

5. X??

6. ?X?

7. ??X

8. ???

Types of missing data

- **Missing Completely at Random (MCAR)**
 - When the probability of response at time t is independent of both the previously observed value and the unobserved at time t
- **Missing at Random (MAR)**
 - When the probability of response at time t depends on the previously observed values but not the unobserved values at time t
- **Not Missing at Random (NMAR)**
 - When the probability of response at time t depends on the unobserved values at time t

Types of missing data: examples

- **Missing Completely at Random (MCAR)**
 - Lab technician accidentally destroys a batch of blood samples
- **Missing at Random (MAR)**
 - Study protocol requires patients be withdrawn if their biomarker exceeds a pre-specified value at a follow-up
- **Not Missing at Random (NMAR)**
 - Pain study in which patients ask for rescue medication when symptoms become too severe, but we do not get opportunity to record their outcome

Other reasons for missing data

- Are the characteristics of patients with missing data different from those for whom complete data are available?
 - Men more likely to dropout than women?
 - Differences between social classes?
 - Differences between treatment groups?

Analysing longitudinal data with missing data

- Analyse those participants with complete data only
 - Loss of power
 - Biased results if reason for missing values related to outcome
- Impute data
 - Simple imputation
 - Multiple imputation
- Use a method that can deal with missing values

Imputations methods

- Motivated by
 - Desire to avoid bias due to missing data
 - Unavailability of methods for incomplete data
 - Methods available now assume missing observations MCAR or MAR
- Useful as part of sensitivity analyses
 - Examine the impact of the results on specific assumptions about individuals with missing observation
- **Warning:** does not provide definitive solution and requires the analyst to make untestable assumptions about the missing data

Simple imputation

- Substituting a single reasonable value for a missing observation
- Examples
 - Mean of observed data
 - Last value carried forward
 - Interpolation
 - Hot decking
 - Regression based model

Mean value substitution

- Computed across all individuals or specific to treatment group
- Calculate the average value of those observed and substitute
- Estimate of the mean of the augmented dataset remains the same as the mean for the original data
- Estimate of the SD will be reduced artificially
 - Distorted significance tests and falsely narrow CI

Mean value substitution

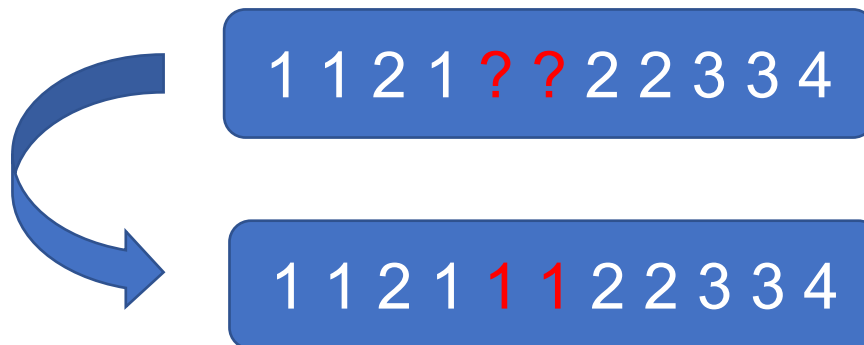
- Assumes MCAR
- Same imputed value at week 6 despite baseline value

ID	Baseline	6 weeks
1	60.9	54.9
2	82.1	54.9
3	57.2	40.6
4	63.5	52.4
5	67.5	71.6

$$(40.6 + 52.4 + 71.6) / 3 = 54.9$$

Last value carried forward

- Assumes no change since last observed assessment
- E.g. if a patient has two missing assessments



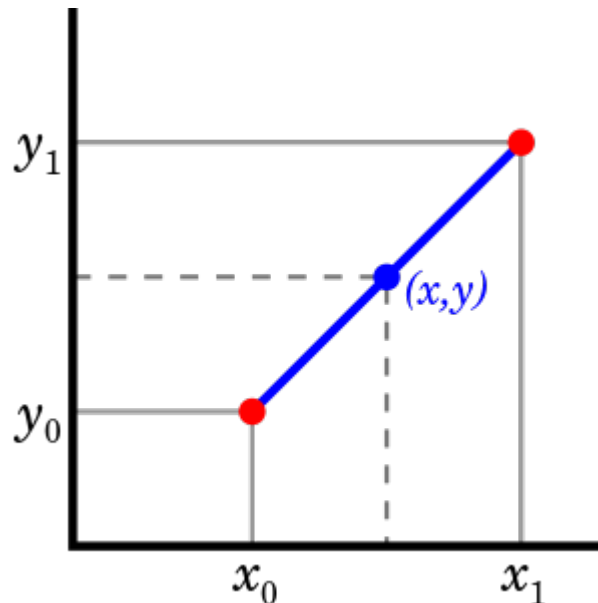
Horizontal mean imputation

- An alternative to simple mean imputation
 - Takes in to account the longitudinal nature of the data
 - Impute the missing value from the mean of the patients own previous scores
 - Reduces to LVCF if there is only one previous assessment available

Interpolation

- Assume constant rate of change between observed time points

$$y = y_0 + \left(\frac{y_1 - y_0}{x_1 - x_0} \right) (x - x_0)$$



Hot-deck imputation

- Selects at random from patients with observed data and substitutes this as the imputed value for the patient with the missing data
- The hot deck refers to the deck of responses of patients with observed data from which the missing value is selected
- The deck chosen may be restricted to those patients that are similar to the patient with the missing value

Regression model

- Identify a regression model to predict the missing observation
- Advantage is that it can include additional information
- Assumes the missingness depends only on the observed data and covariates in the regression

Multiple imputation

- Methods so far select a single value and substitute in to the data as if observed values
 - Results in underestimation of the variance
- Idea of multiple imputation is that many alternative “complete” datasets can be created
- More complex technique incorporating the variability of the outcome but also the uncertainty about the missing observations

Summary of missing values

- Investigators are suspicious about using imputation techniques because of the assumptions involved
- Not imputing missing data makes the assumption that patients failing to respond are similar to those who do
 - Some imputation methods also assume this
- Imputation tries to use the available information to make better allowances for patients with missing data
- No imputation method is a substitute for the real data

Where to find more information?

- Omar RZ, Wright EM, Turner RM, Thompson SG. Analysing repeated measurements data: a practical comparison of methods. *Stat Med* 1999;18:1587–603.



Next session at
14:30

Sample size calculations

Questions

- Why do I need to do a sample size calculation?
- When do I need to do a sample size calculation?
- What information do I need to do a sample size calculation?

Why & when?

- **Scientific**: might miss out on an important discovery (testing too few), or find a clinically irrelevant effect size (testing too many)
- **Ethical**: might sacrifice subjects (testing too many) or unnecessarily expose too few when study success chance low (testing too few)
- **Economical**: might waste money (testing too many) or have to repeat the experiment again (testing too few)
- Also, generally required for study grant proposals + papers, so needs to be done **before** experiment

Necessary information

General case

- Type I error rate
- Power (= 1 - Type II error rate)
- Minimum clinically important difference
- Estimate of the variability

	What is it?	How do you estimate it?
A type I error rate?	False positive rate from hypothesis test	0.05 or 0.01
A type II error rate?	False negative rate from hypothesis test	0.2 or 0.1
A minimum clinically important difference?		
An estimate of variability?		

Example: hypothesis test

- Null hypothesis (H_0):
 - No difference in mean blood pressure of patients on treatment and placebo
- Alternative hypothesis (H_1):
 - Mean blood pressure of patients on treatment and placebo differs

Error rates

		Decision using hypothesis test	
		No evidence of a difference	Evidence of a difference
Truth	No difference	True Negative	False positive Type I error (α)
	Difference	False negative Type II error (β)	True Positive

Convention: $\alpha = 0.05$ or 0.01 , $\beta = 0.1$ or 0.2

NB: Power = $1 - \beta$

	What is it?	How do you estimate it?
A type I error rate?	False positive rate from hypothesis test	0.05 or 0.01
A type II error rate?	False negative rate from hypothesis test	0.2 or 0.1
A minimum clinically important difference?	The smallest difference in the outcome that would lead you to think there is a difference between treatment and control	Clinical judgement
An estimate of variability?		

	What is it?	How do you estimate it?
A type I error rate?	False positive rate from hypothesis test	0.05 or 0.01
A type II error rate?	False negative rate from hypothesis test	0.2 or 0.1
A minimum clinically important difference?	The smallest difference in the outcome that would lead you to think there is a difference between treatment and control	Clinical judgement
An estimate of variability?	Variability in the outcome (same for treatment and control groups)	From other similar studies or do a pilot study

Necessary information

- When doing a longitudinal study, what **extra information** do I need to do a sample size calculation?
- Type I error (α)
- Power ($1 - \beta$)
- Minimum clinically important difference
- Standard deviation
- Estimate of correlation (ρ)
- Number of pre (v) and post (w) intervention measurements

Considerations for longitudinal studies

- How many repeated measurements should be taken?
- When should these measurements be taken?
- Correlation between the measurements from the same subject
- These should be decided on a study-by-study basis

Sample size: ANCOVA

Compound symmetry assumption

- Observations made at time t_1 have a correlation ρ with observations made at time t_2
- This correlation is assumed to be the same for all values of t_1 and t_2 , provided they are not equal

Example: compound symmetry

Blood pressures for one person measured at 5 time points:

bp_1 bp_2 bp_3 bp_4 bp_5

Compound symmetry assumes:

- $\text{Corr}(bp_1, bp_1) = 1$
- $\text{Corr}(bp_2, bp_3) = \text{Corr}(bp_2, bp_4) = \dots = \text{Corr}(bp_3, bp_5) = \rho$

Sample size: ANCOVA

- Two groups of size n
- Δ = standardised difference
= difference / standard deviation
- Use an ANCOVA model with means of pre- and post-measurements

Sample size formula: ANCOVA

$$n = R \left[\frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} \right]$$

where R is a **correction factor**, given by

$$R = \left[\frac{1 + (w - 1)\rho}{w} - \frac{v\rho^2}{1 + (v - 1)\rho} \right]$$

Statistical and mathematical symbols

- Values for $(z_{1-\alpha/2} + z_{1-\beta})^2$ can be obtained from this table:

		β (Type II error)			
α (Type I error)		0.05	0.10	0.20	0.50
	0.10	10.8	8.6	6.2	2.7
	0.05	13.0	10.5	7.9	3.8
	0.02	15.8	13.0	10.0	5.4
	0.01	17.8	14.9	11.7	6.6

Example: blood pressure

- **Example:** study of agent versus placebo for reducing blood pressure
 - 1 baseline measurement
 - 1 follow up measurement
 - Standardized effect size = 0.4
 - Correlation = 0.7
 - Type I error rate = 5%
 - Power = 80%
- How many people do we need?

Example: blood pressure

$$R = \frac{1 + (1 - 1)0.7}{1} - \frac{1(0.7^2)}{1 + (1 - 1)0.7} = 0.51$$

$$n = R \left[\frac{2(7.9)}{0.4^2} \right] = 98.175 \times R$$

$$= 98.175 \times R = 50.07$$

Therefore we need 51 people per group

Should I plan to include another measurement?

- When more measurements are taken, fewer subjects are needed to achieve the same statistical power...
- ...but is the sample size greatly reduced or not?

Example: impact of one more measurement

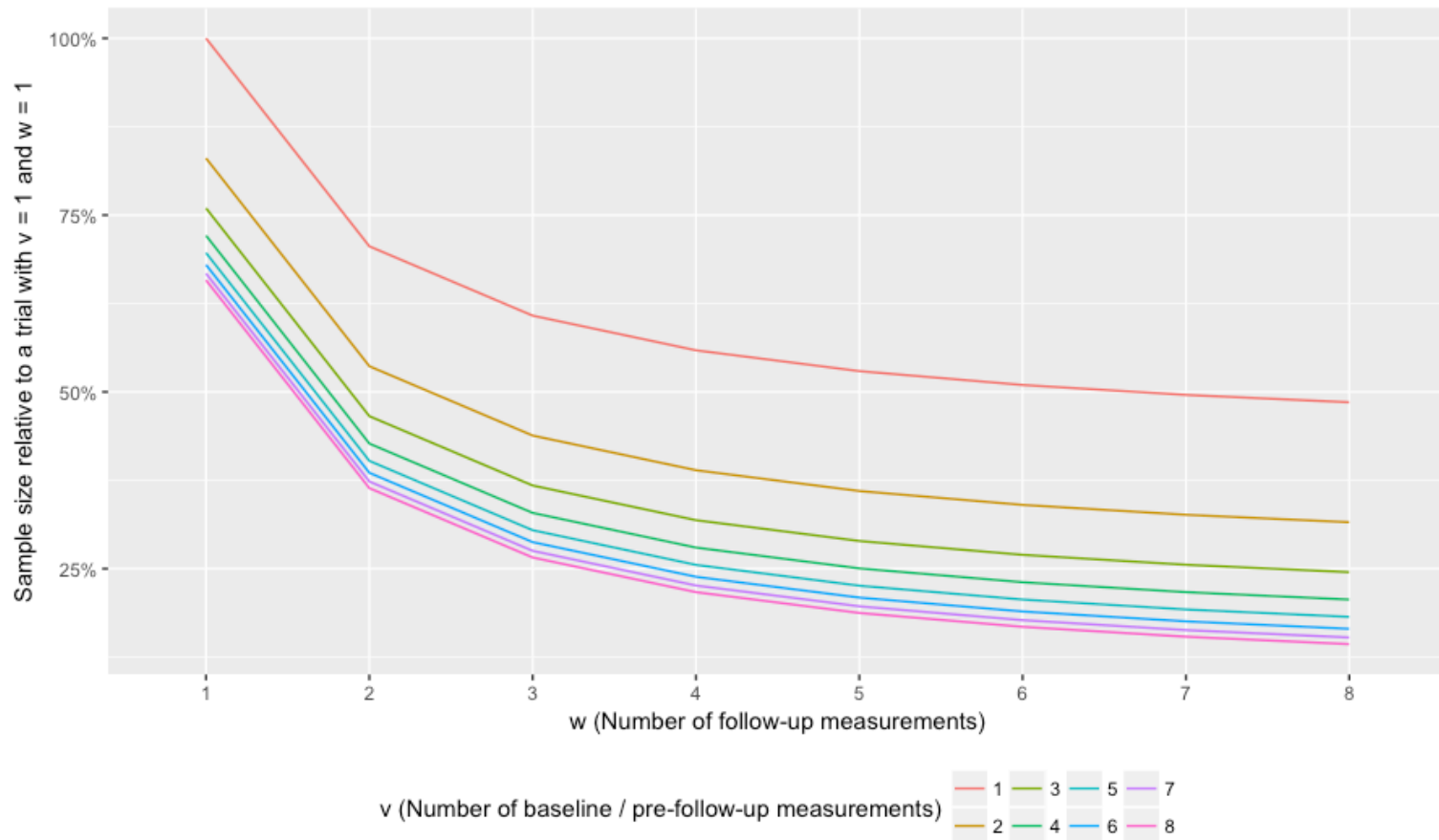
- Using 2 follow up measurements ($w = 2$)
- We need to recalculate R the correction factor
 $R = 0.36$
- We would require 36 ($= 0.36 * 98.175$) patients per group
- Sample size has been reduced by $(51 - 36) / 51 = 0.29$, or approximately 30%

Example: impact of two more measurement

- Using 3 follow up measurements ($w = 3$)
- We need to recalculate R the correction factor
 $R = 0.31$
- We would require 31 ($= 0.31 * 99.71$) patients per group
- Sample size has been reduced by $(51 - 31) / 51 = 0.39$, or approximately 40%

Sample size reductions

- Sample size relative to a trial with one baseline and one follow up measurement



Increasing the number of measurements

- Increasing the number of follow up and/or baseline measurements can dramatically reduce the sample size
- The increases in power for each additional measure rapidly decreases with the increasing number of assessments
- This can be useful if the number of subjects is limited

Other considerations

- When more measurements are taken, fewer patients are needed
 - Typically no benefit from more than 4 (without baseline data) or 7 (with baseline data) repeated measures needed
 - Exception is when correlation is low, e.g. episodic studies
- But you must also consider:
 - Burden on the subject
 - Cost, time, resources
- Number of measurements and their timing should be stated in the study protocol

Where to find more information?

- Vickers AJ. How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Med Res Methodol* 2003;3:22.
- Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistic and its implications for design. *Stat Med* 1992;11:1685–704.
- Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Second Edi. Oxford, UK: Oxford University Press; 2013.

Advanced methods

Advanced methods

- Linear mixed models
- Generalised estimating equations

What is a linear mixed model (LMM)?

- Also referred to as *random effects models*, *hierarchical models*, *multilevel models*, *random growth-curve models*
- Includes subject-specific effects that allows for a separate time trend for each subject
- Accounts for correlation between repeated measures within subjects

What is a LMM?

Fixed effects model: $y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$

Mixed effects model: $y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij}$

- (b_{0i}, b_{1i}) are the **random effects** – one pair per subject
 - b_{0i} is the subject-specific deviation from the average intercept
 - b_{1i} is the subject-specific deviation from the average slope
- (β_0, β_1) are the **fixed effects**
- Mixed effects model = **random effects** + **fixed effects**

Example: sleep deprivation study*

- The average reaction time per day for subjects in a sleep deprivation study
- On day 0 the subjects had their normal amount of sleep
- Starting that night they were restricted to 3 hours of sleep per night
- **Outcome:** the average reaction time (in milliseconds) on a series of tests given each day to each subject

Reaction

400

300

200

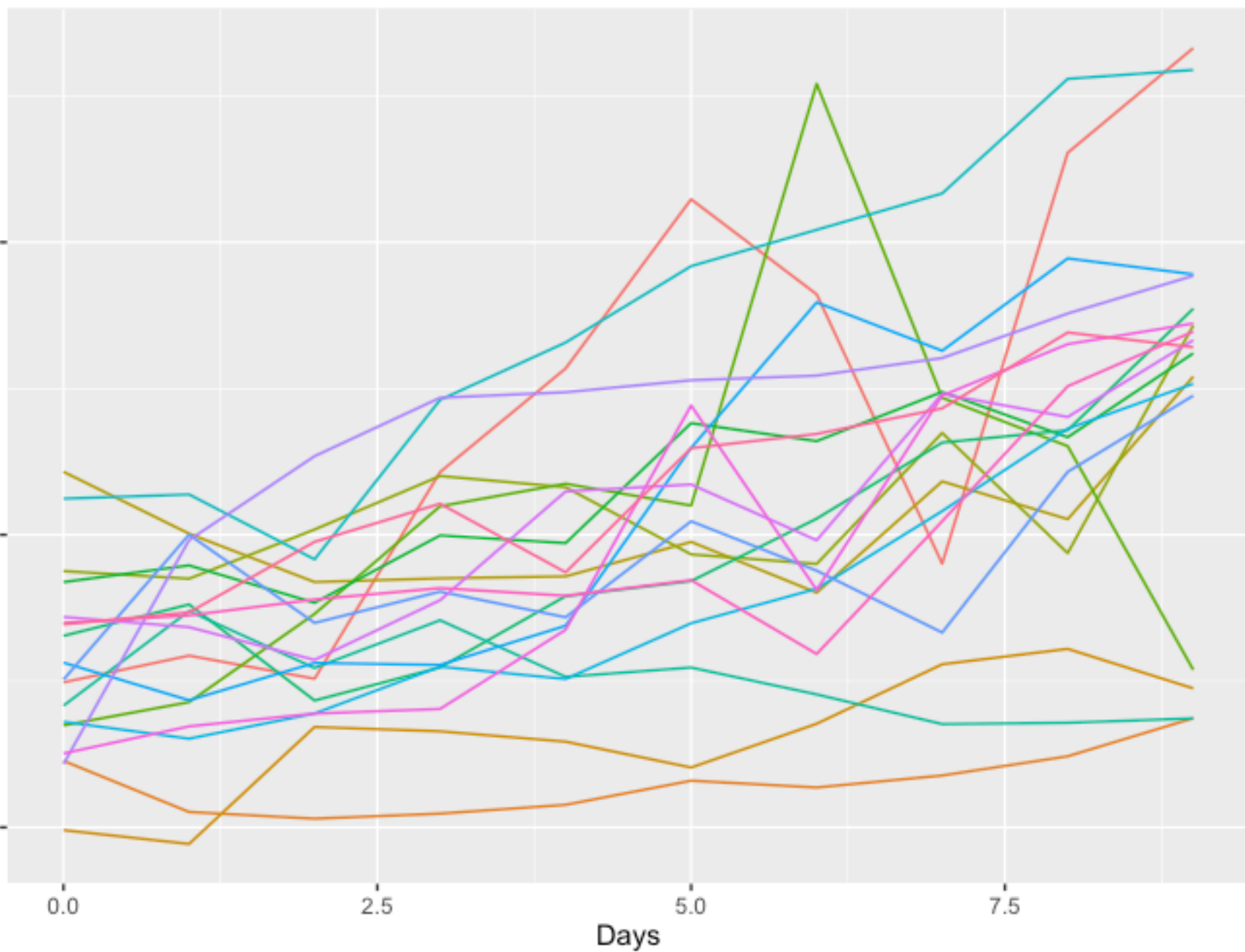
0.0

2.5

5.0

7.5

Days



Fixed effects regression line

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$$

Reaction

400

300

200

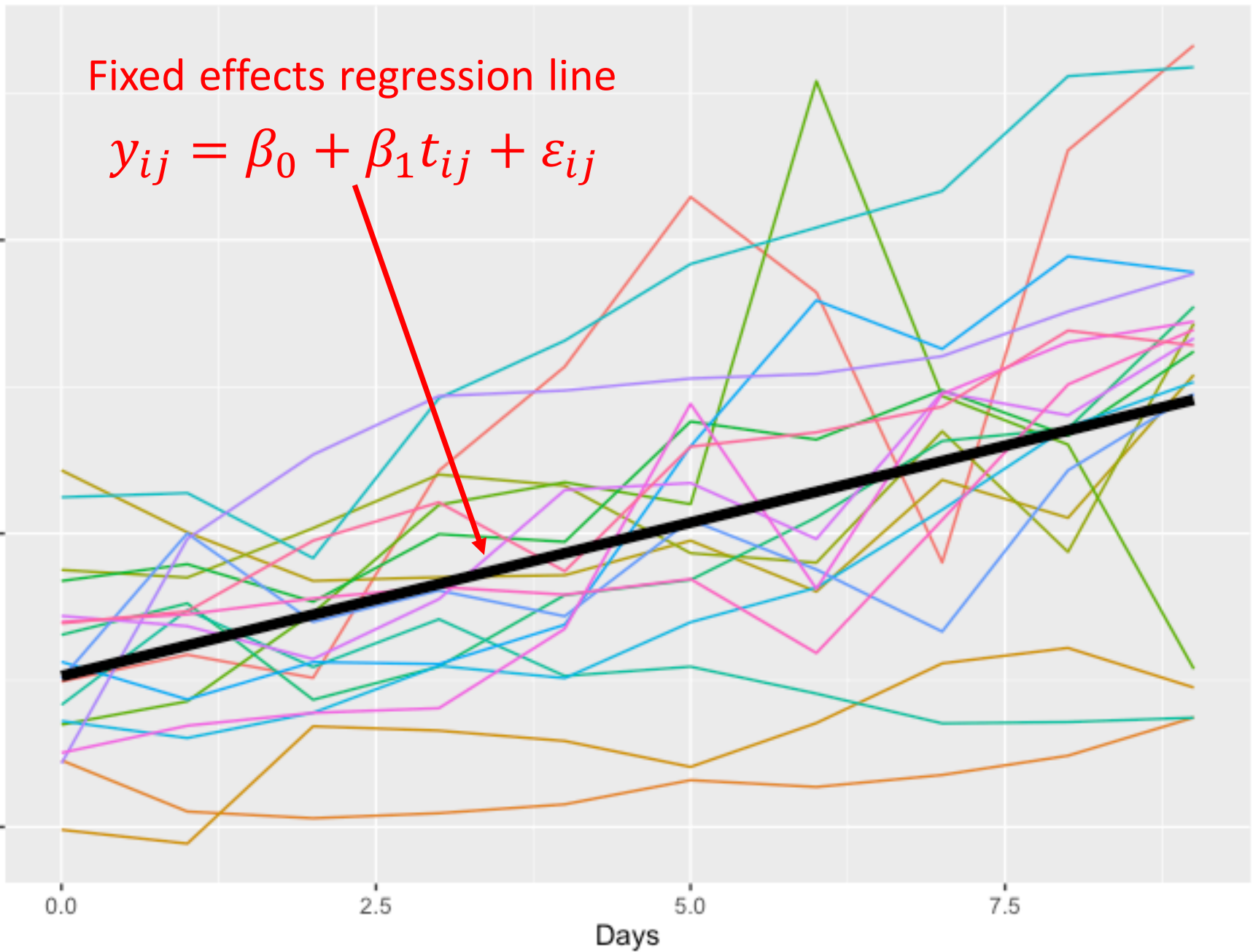
0.0

2.5

5.0

7.5

Days



Fixed effects regression line + within-subject intercepts

$$y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \varepsilon_{ij}$$

Reaction

400

300

200

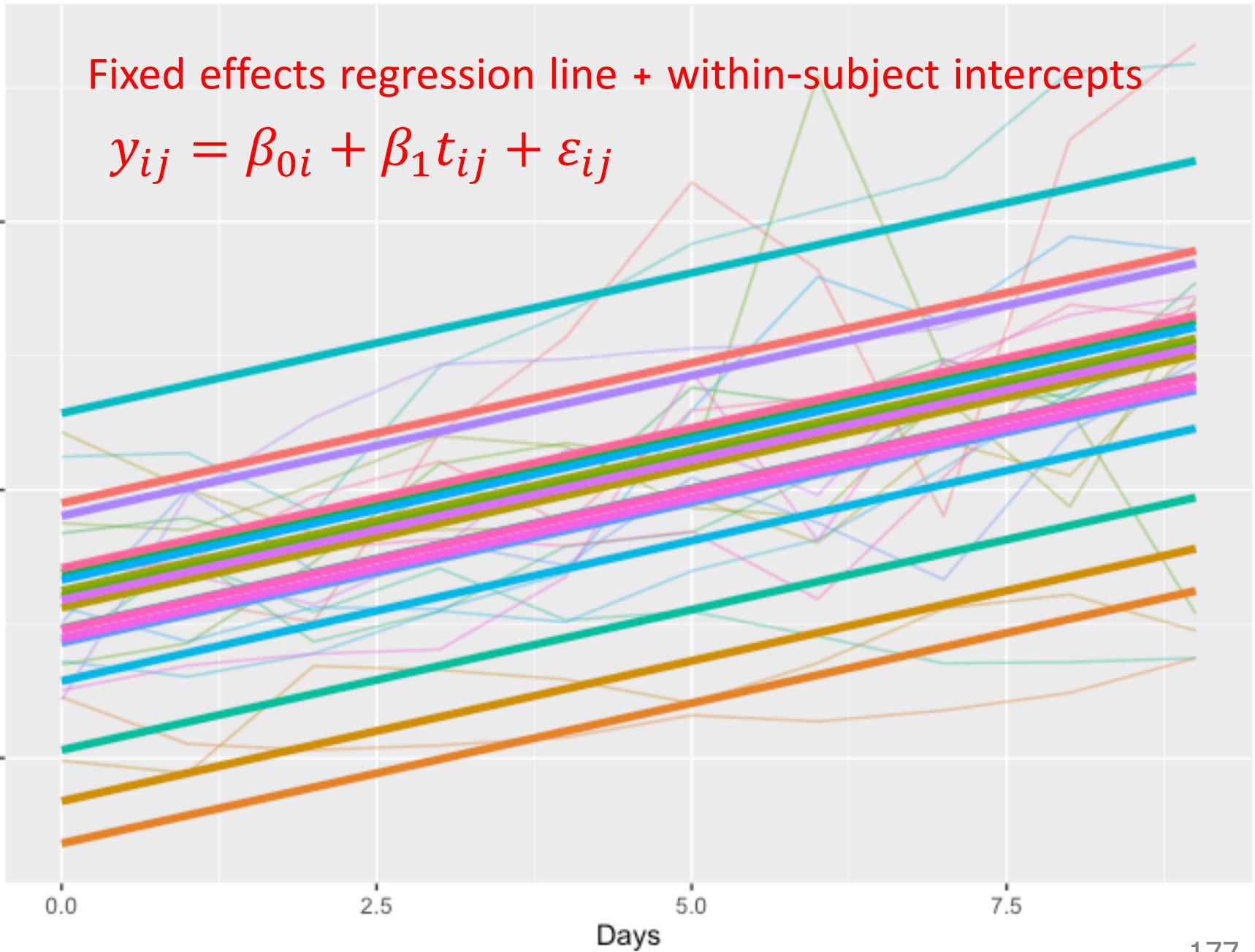
0.0

2.5

5.0

7.5

Days



Within-subjects fixed effects regression lines

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \varepsilon_{ij}$$

Reaction

400

300

200

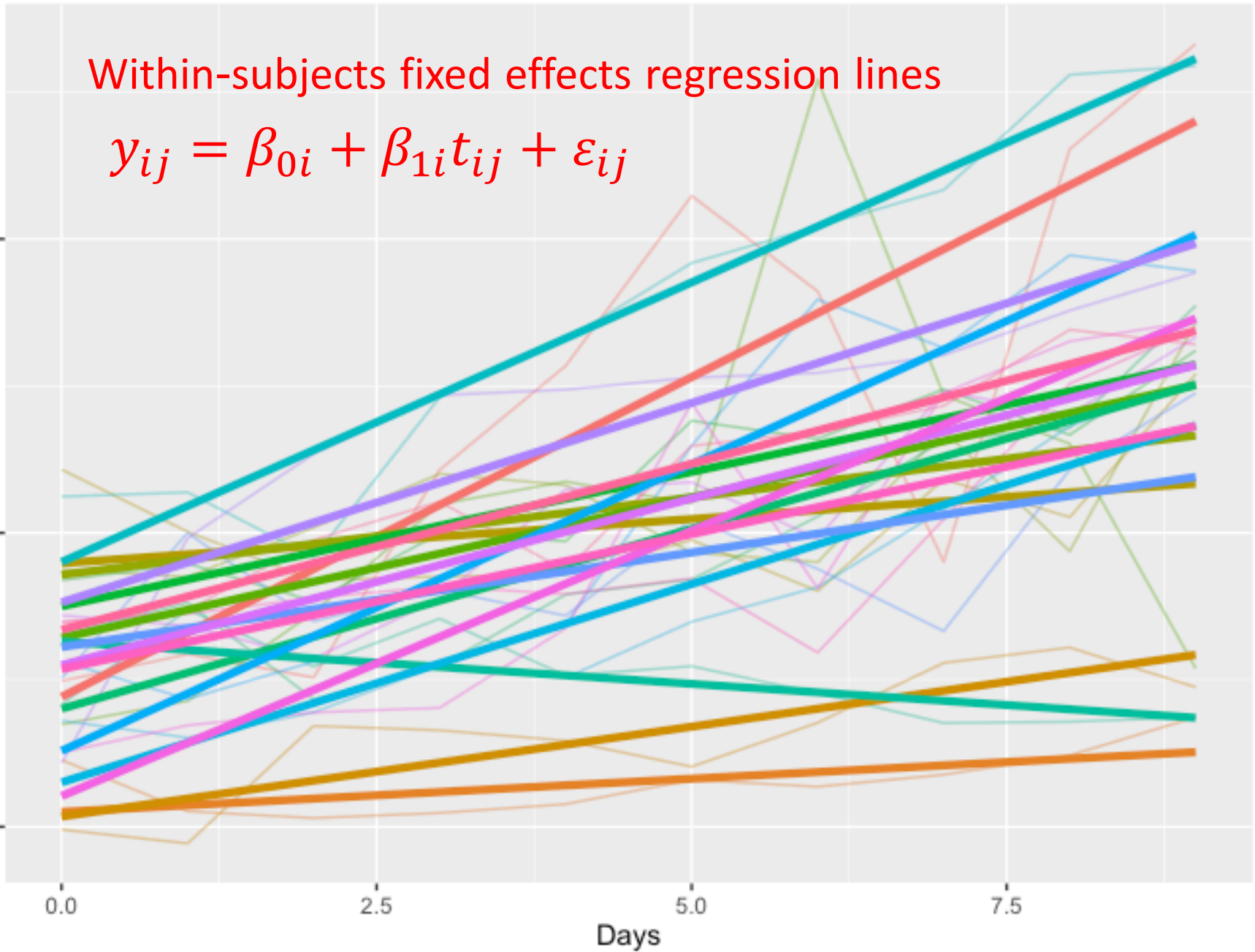
0.0

2.5

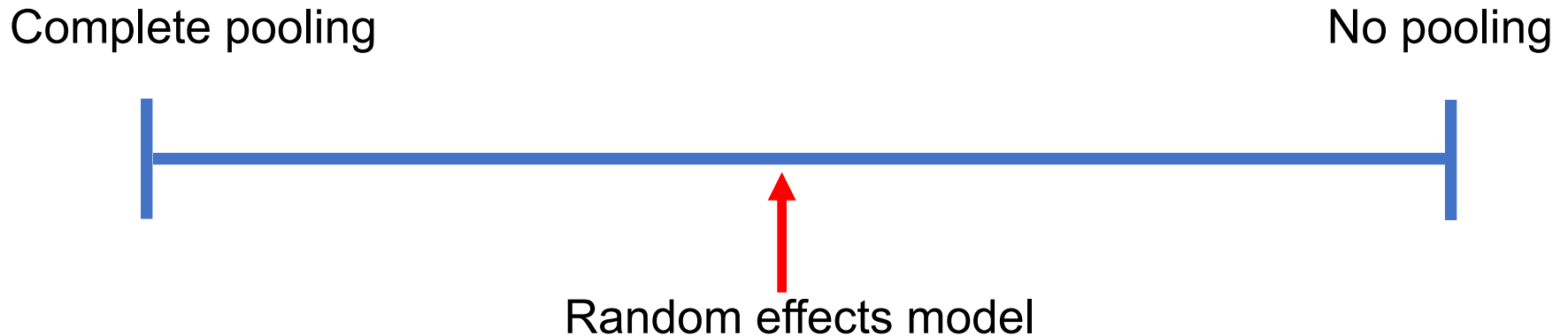
5.0

7.5

Days



What is a LMM?



- A compromise between the 2 extremes of
1. $y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$ (complete pooling)
 2. $y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}$ (no pooling)

What is a LMM?

- Needs additional distributional assumption on the random effects
- Common assumption: (b_{0i}, b_{1i}) from a multivariate normal distribution with mean 0 and common covariance matrix Σ

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- ρ is the correlation between random intercepts and slopes

Example: sleep deprivation study

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	251.405105	6.824557	17	36.838	.000	237.006549	265.803661
Days	10.467286	1.545789	17.000	6.771	.000	7.205956	13.728615

a. Dependent Variable: Reaction.

Interpretation

The average reaction time at baseline is $\hat{\beta}_0 = 251.4$ ms, which significantly ($P < 0.001$) increases by an average of $\hat{\beta}_1 = 10.5$ ms per day of sleep deprivation

Example: sleep deprivation study

Random Effect Covariance Structure (G)^a

	Intercept Subject	Days Subject
Intercept Subject	612.089939	9.604333
Days Subject	9.604333	35.071661

Unstructured

a. Dependent Variable: Reaction.

Residual Covariance (R) Matrix^a

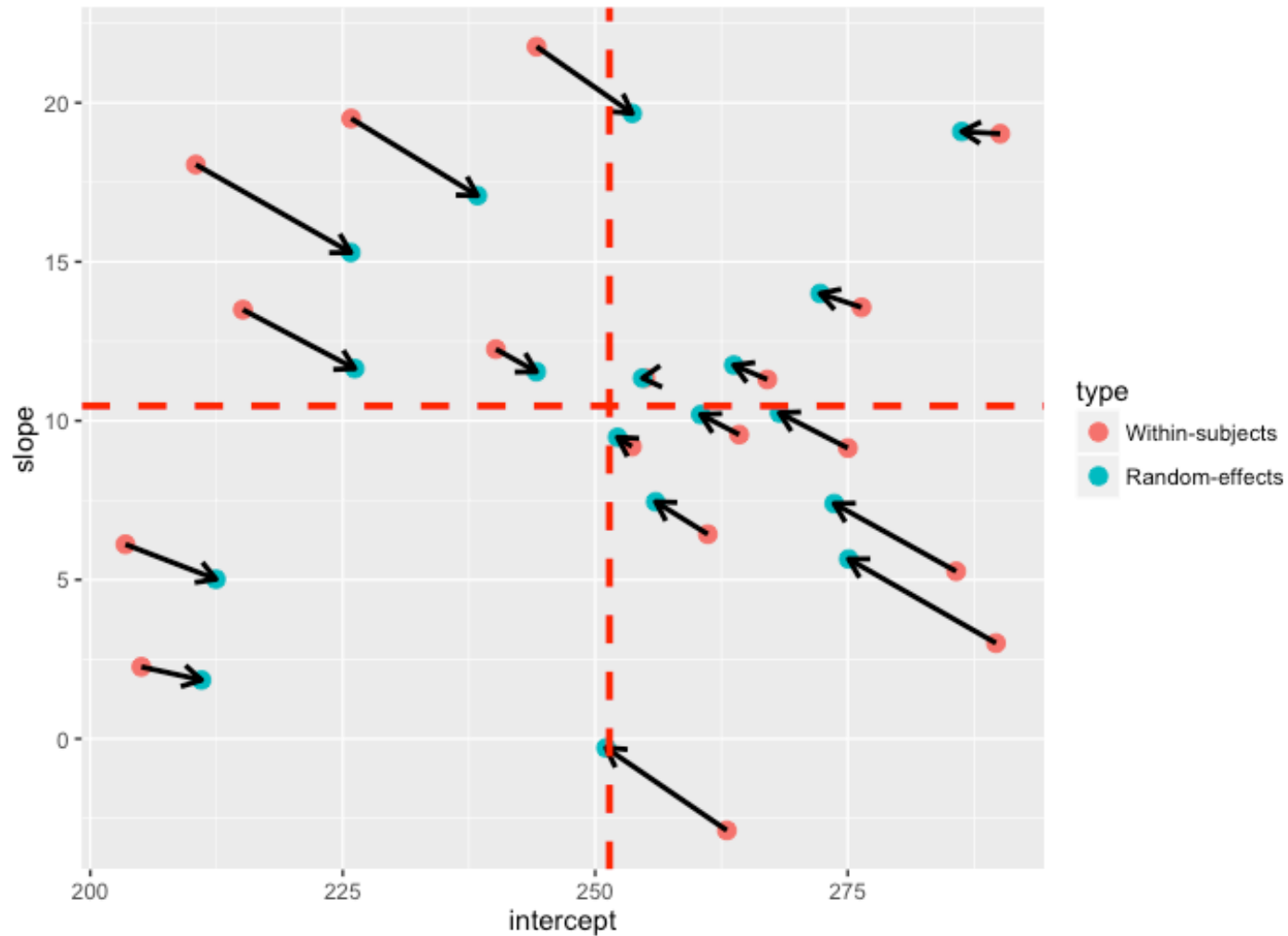
	Residual
Residual	654.941027

a. Dependent Variable: Reaction.

Interpretation

The positive covariance (9.60) between intercepts and slopes suggests that subjects with longer baseline reaction lines to some extent have slightly larger slopes (worse reaction times after sleep deprivation) with correlation $\hat{\rho} = \frac{9.60}{\sqrt{612.09 \times 35.07}} = 0.07$

Shrinkage



Comparing LMMs

- Compute the Akaike Information Criterion (AIC) statistic
- Trades of model fit against model complexity
- **Rule-of-thumb**: smaller AIC is preferred

Random intercepts + random slopes

Information Criteria^a

-2 Restricted Log Likelihood	1743.628
Akaike's Information Criterion (AIC)	1751.628
Hurvich and Tsai's Criterion (AICC)	1751.859
Bozdogan's Criterion (CAIC)	1768.355
Schwarz's Bayesian Criterion (BIC)	1764.355

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable:
Reaction.

Random intercepts only

Information Criteria^a

-2 Restricted Log Likelihood	1786.465
Akaike's Information Criterion (AIC)	1790.465
Hurvich and Tsai's Criterion (AICC)	1790.534
Bozdogan's Criterion (CAIC)	1798.829
Schwarz's Bayesian Criterion (BIC)	1796.829

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable:
Reaction.

Extensions of the LMM

- When data are not continuous, e.g. binary correlated data, we can use generalised linear mixed models (GLMMs)
- When linear trajectories unsuitable
 - Higher order terms, e.g. $+(\beta_1 + b_{2i})t_{ij}^2$
 - Splines + random effects
 - Recommend to consult a statistician

Benefits of using LMMs

- Allows for missing data (MAR or MCAR)
- Can incorporate:
 - Additional multilevel covariate, cf. multi-center effects
 - Time-varying covariates
 - Sophisticated covariance structures
- Does not require sphericity
- Does not require regularly spaced time points or even common measurement times
 - Observational data studies almost never have fixed time points

Where to find more information?

- Andrinopoulou E-R, Rizopoulos D, Jin R, Bogers AJJC, Lesaffre E, Takkenberg JJM. An introduction to mixed models and joint modeling: analysis of valve function over time. *Ann Thorac Surg* 2012;93:1765–72.
- Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. 2nd Ed. 2004.

- Any questions?
- Please can you complete the evaluation form