POPULATION RESEARCH SEMINAR SERIES

Sponsored by the Statistics and Survey Methods Core of the U54 Partnership

Application of GEE and Mixed Effects Model in Longitudinal Data Analysis

Presented by:

Ling Shi

Umass Boston

DF/HCC

DANA-FARBER/HARVARD CANCER CENTER

A Comprehensive Cancer Center Designated by the National Cancer Institute

UMASS BOSTON

1. FEATURES OF LONGITUDINAL DATA

Example 1

Prospective Longitudinal Evaluation of Quality of Life in Patients With Permanent Colostomy After Curative Resection for Rectal Cancer

(Ito et al, J Wound Ostomy Continence Nurs, 2012).

- Follow a group of rectal cancer patients who were scheduled to undergo curative surgery with a permanent colostomy to evaluate health-related quality of life in patients with a colostomy immediately before and during the first year after surgery.
- Outcome variable: quality of life using the Short Form-36 version 2
- Measurement schedule: One baseline measurement (before surgery) and three follow-up measurements at 2, 6, and 12 months after surgery

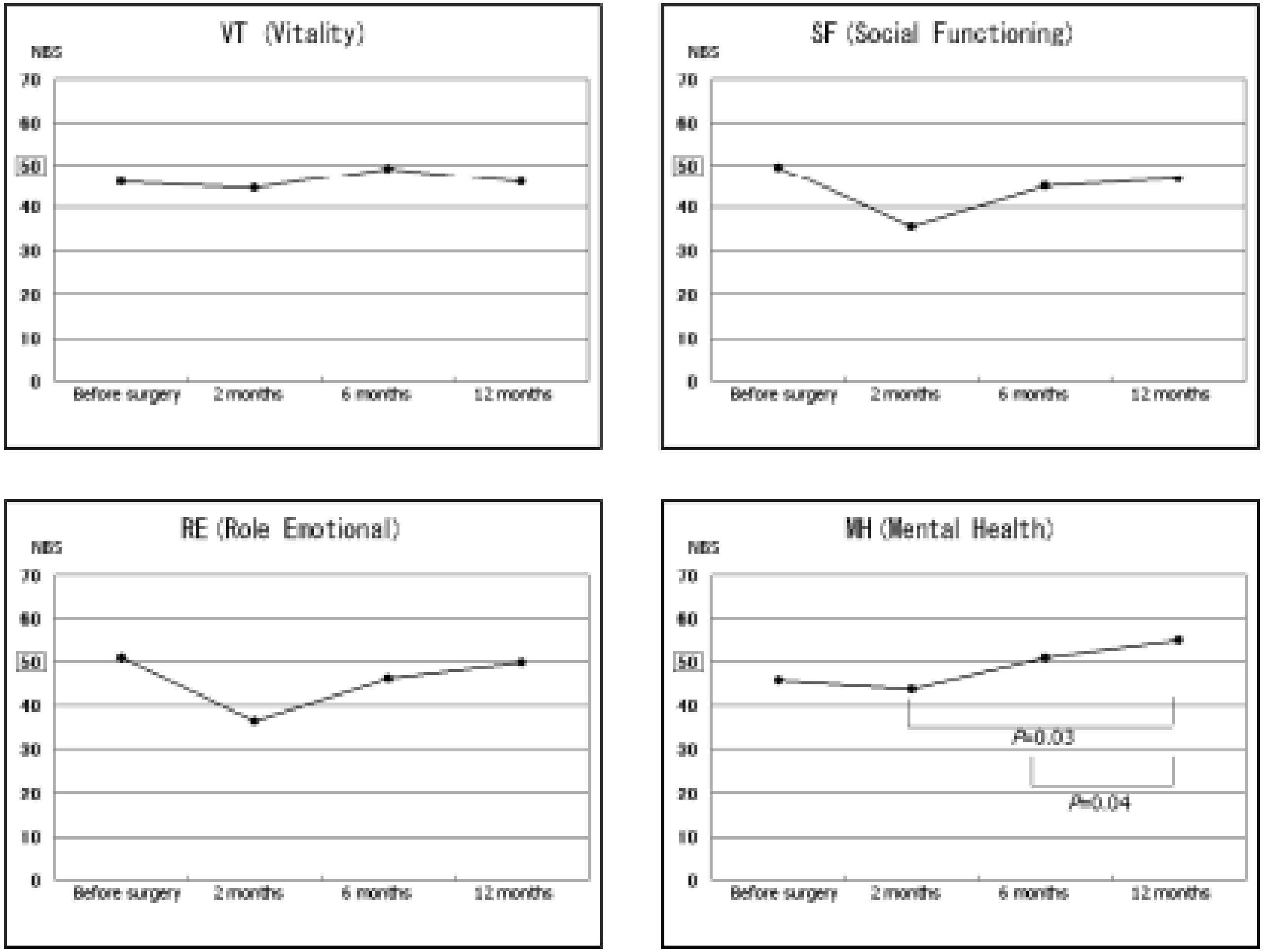
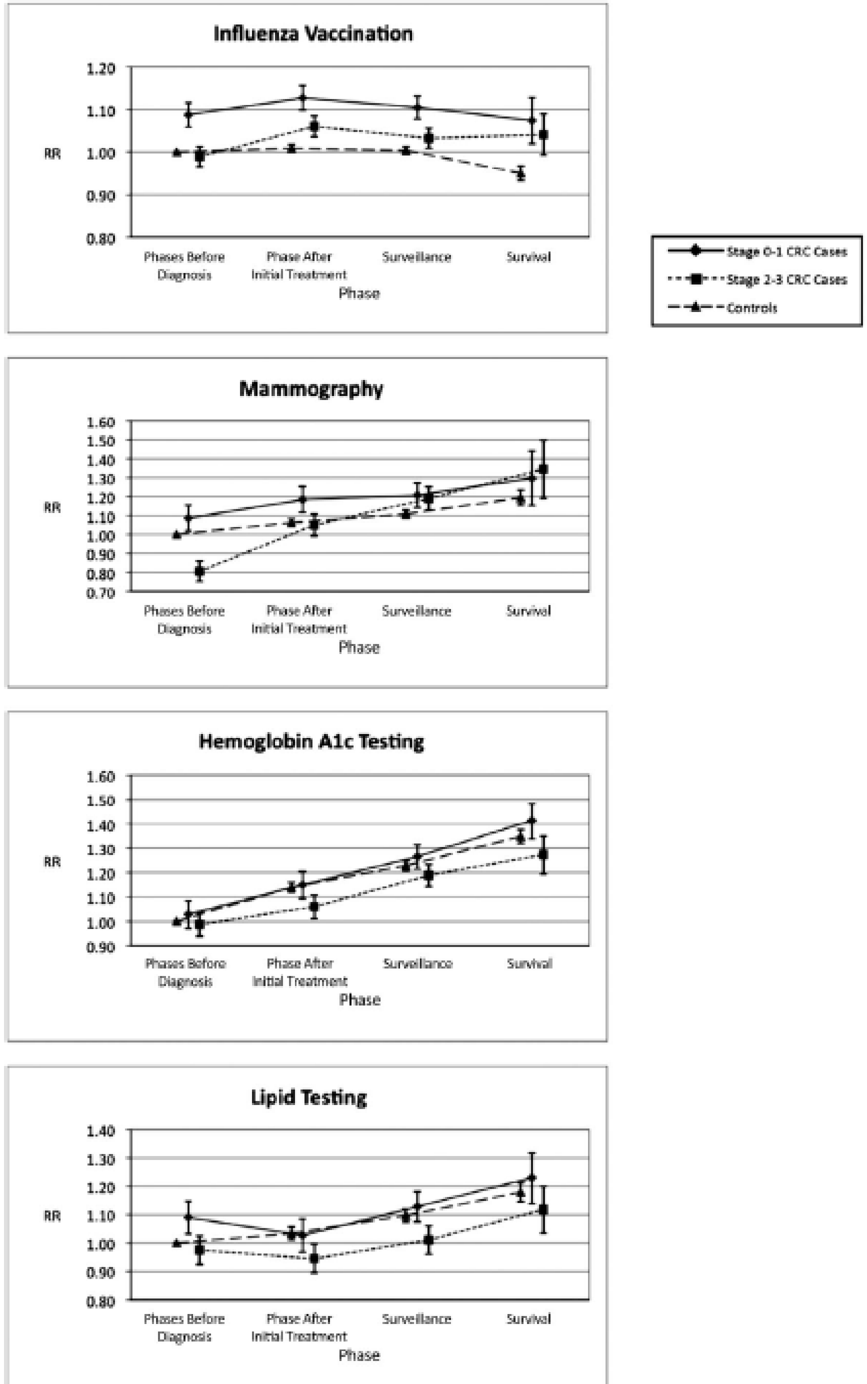


FIGURE 2. Changes in quality-of-life scores of 7 patients completed the questionnaires at all 4 time points (N = 7). Abbreviation: NBS, norm-based scoring.

Example 2

Receipt of general medical care by colorectal cancer patients (Baldwin et al, J Am Board Fam Med, 2011).

- To evaluate changes in general medical care among elderly patients with colorectal cancer (CRC), from before diagnosis through long-term survival.
- Outcome variable: Receipt of preventive services (influenza vaccination, mammography) and, among diabetics, HgbA1c and lipid testing
- Measurement schedule: One baseline measurement (before diagnosis) and three follow-up measurements: after initial treatment, the surveillance phase (2-4 years after washout), and the survival care phase (5-7 years after washout)



Influenza Vaccination
1.20
1.10
1.00
0.90
0.80
RR
Phases Before Diagnosis
Phase After Initial Treatment
Surveillance
Survival
Phase
Stage 0-1 CRC Cases
Stage 2-3 CRC Cases
Controls
Mammography
1.60
1.50
1.40
1.30
1.20
1.10
1.00
0.90
0.80
0.70
RR
Phases Before Diagnosis
Phase After Initial Treatment
Surveillance
Survival
Phase
Hemoglobin A1c Testing
1.60
1.50
1.40
1.30
1.20
1.10
1.00
0.90
RR
Phases Before Diagnosis
Phase After Initial Treatment
Surveillance
Survival
Phase
Lipid Testing
1.40
1.30
1.20
1.10
1.00
0.90
0.80
RR
Phases Before Diagnosis
Phase After Initial Treatment
Surveillance
Survival
Phase

Example 3

Quality-of-life evaluation for advanced non-small-cell lung cancer: a comparison between vinorelbine plus gemcitabine followed by docetaxel versus paclitaxel plus carboplatin regimens in a randomized trial (Kawahara et al, BMC Cancer, 2011).

- A randomized trial of vinorelbine plus gemcitabine followed by docetaxel (VGD) versus paclitaxel plus carboplatin (PC) in patients with advanced non-small-cell lung cancer, to test whether the VGD regimen produced better QOL compared with the PC regimen in patients with advanced NSCLC.
- Outcome variable: Quality of life assessed by the FACT-L, FACT-Taxane and FACIT-Sp QOL instruments
- Measurement schedule: One baseline measurement and three follow-up measurements at 6, 12 and 18 weeks after the treatment

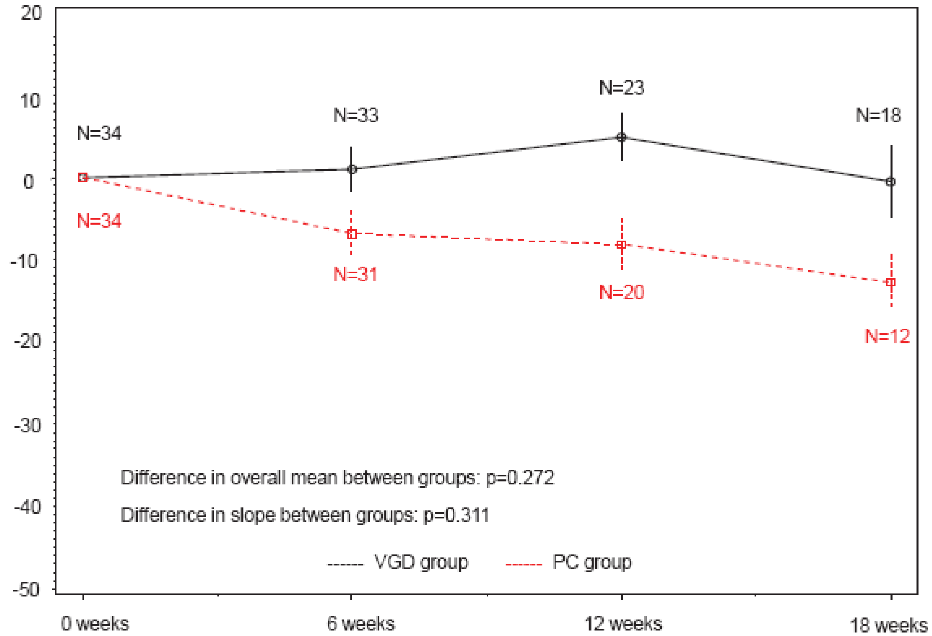


Figure 2 Changes in FACT-L score.

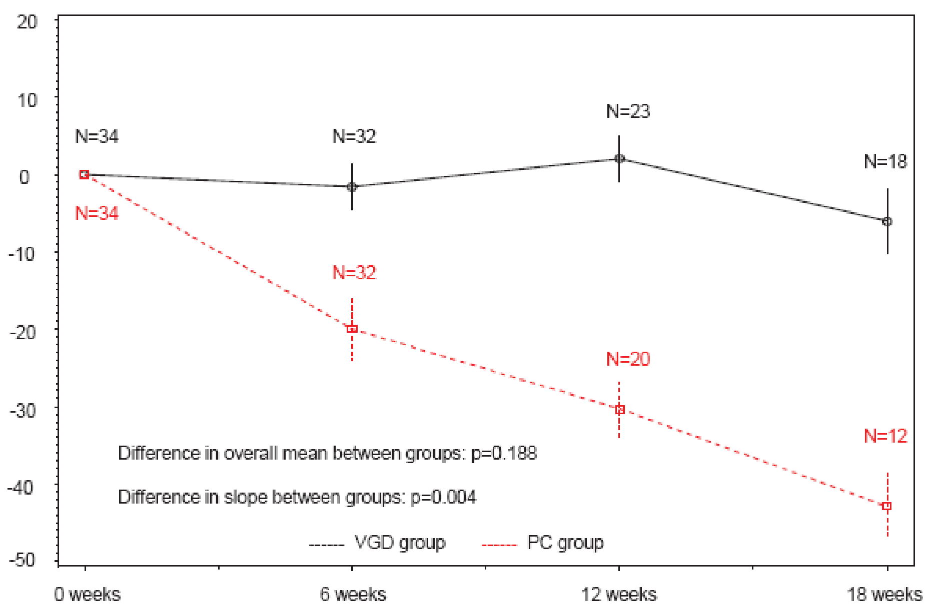
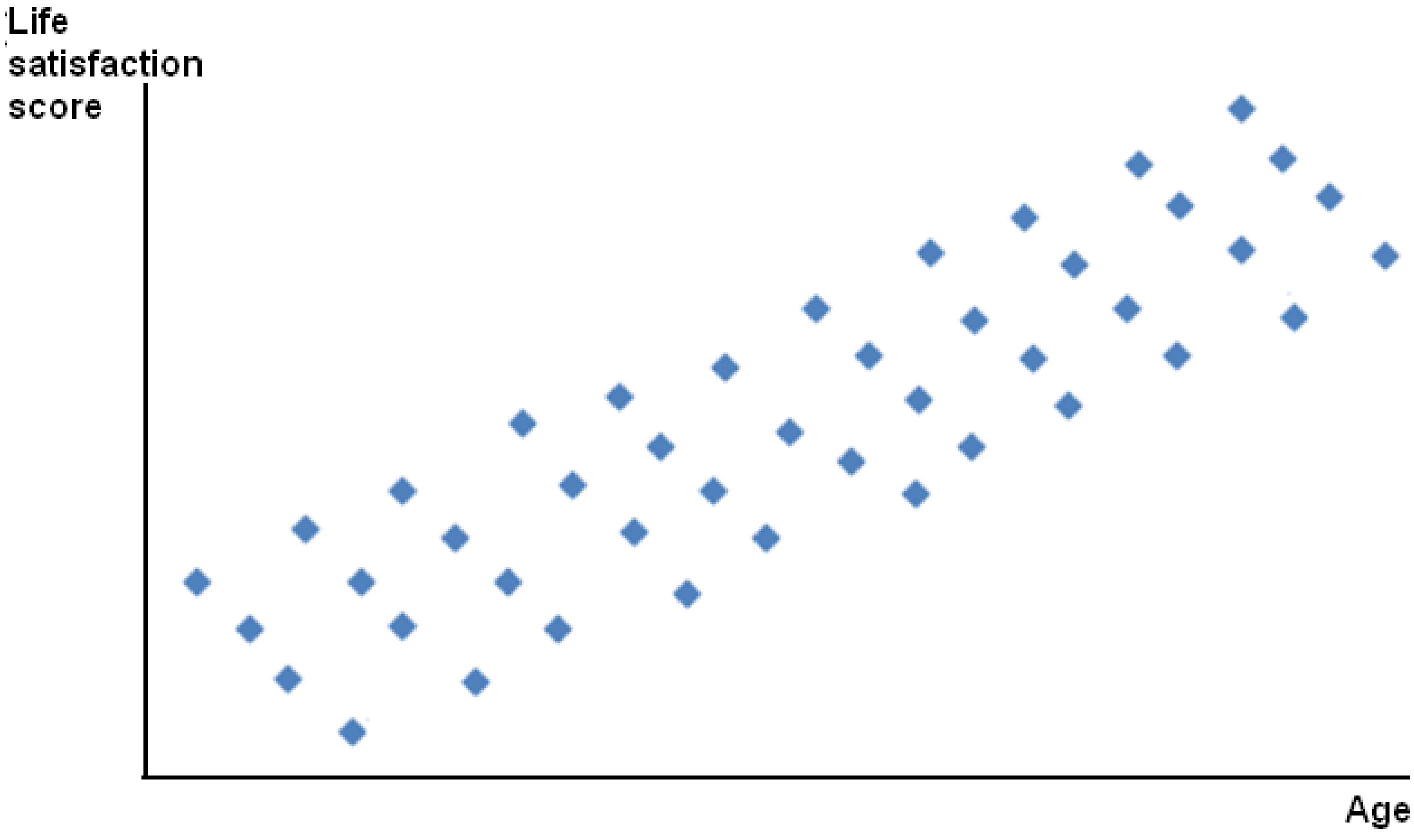


Figure 3 Changes in FACT-Taxane score.

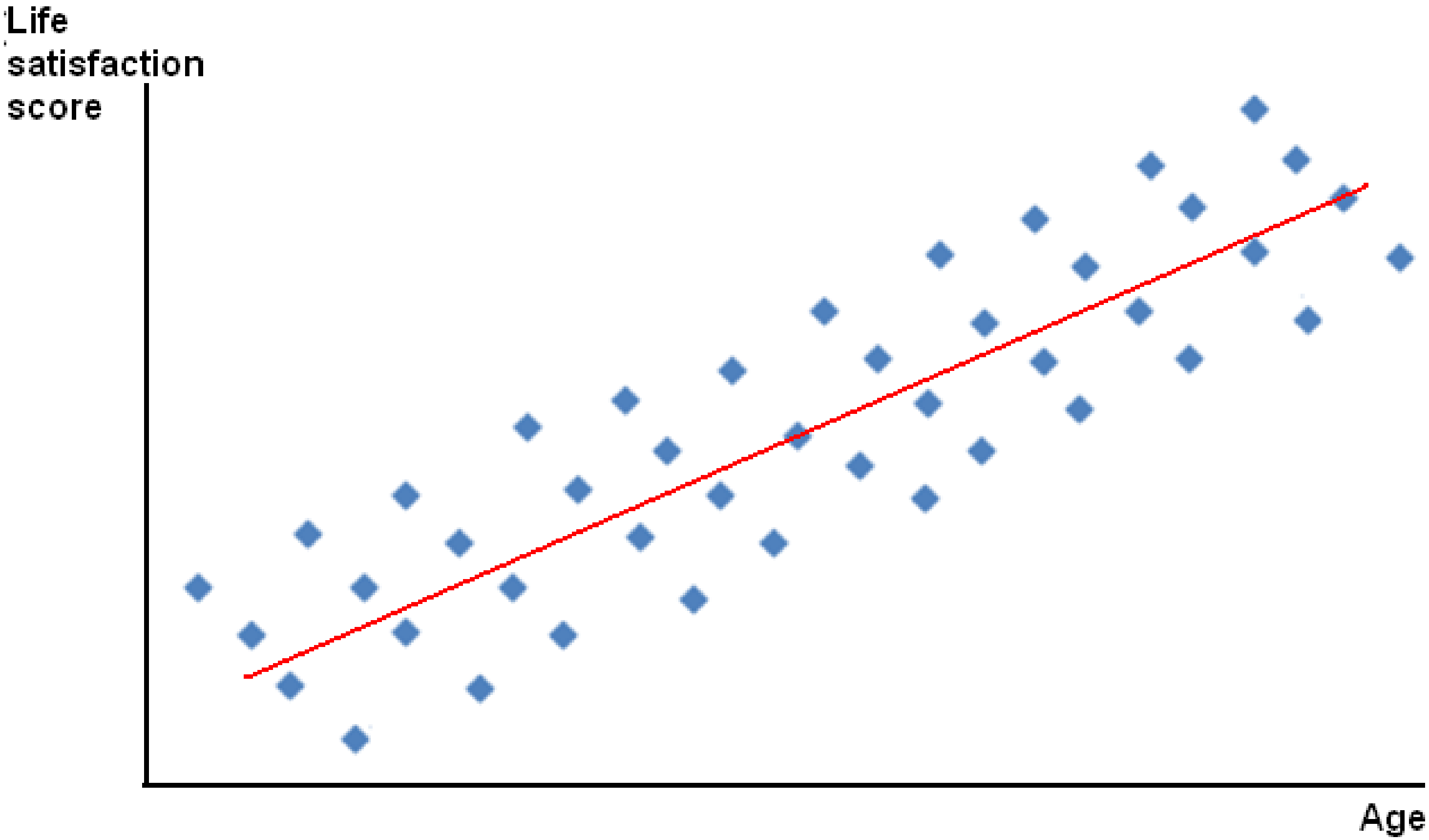
Longitudinal Data

- Repeated measurements on study participants
- The trajectories of outcome variables

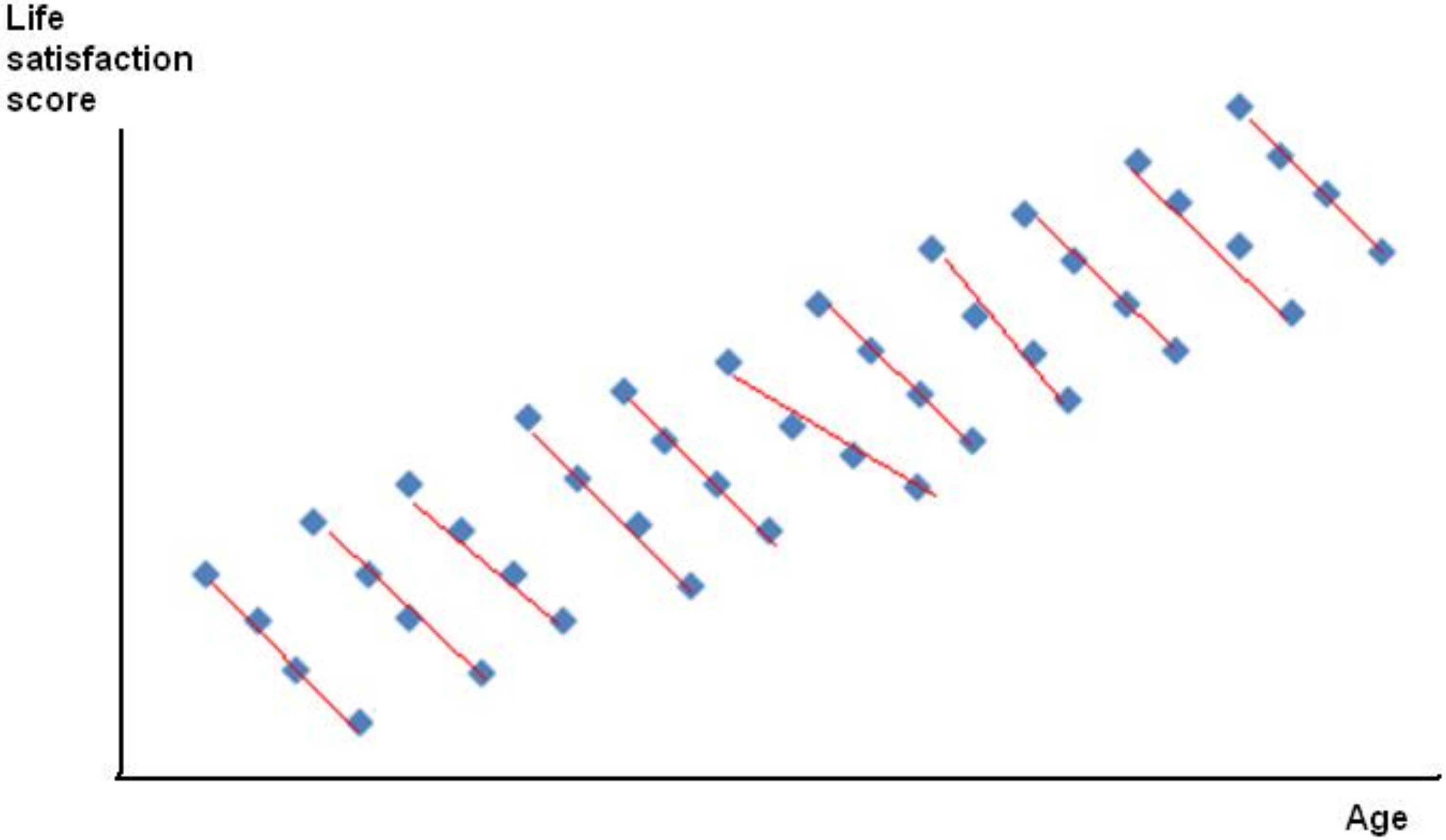
Cross-sectional vs. Longitudinal



Cross-sectional data



Longitudinal data



Features of Longitudinal Data

- Longitudinal data: Repeated measurements usually positively correlated
- OLS regression: Observations are independent of each other
- Ignoring correlation between observations can result in **bias** in SE: significance and CI
- Direction of bias depends on whether the variable is time-dependent (varying) or time-independent (unvarying)

When ignoring correlation….

- Time-dependent variable, e.g. “age”: change in Y by age (slope) →overestimated SE → spuriously large p value → less likely to reject H0

- Time-independent variable, e.g. “treatment”: difference in Y between intervention and control → underestimated SE → spuriously small p value → more likely to reject H0

Special techniques for longitudinal data analysis to account for correlation

2. POPULATION AVERAGE MODEL

Example: Subjects are randomized to receive intervention (Trt=1) and control (Trt=0).

Subject	**Y_{i1}**	**Y_{i2}**	**…**	**Y_{ij}**	**Trt**	**Sex**	**Race**
1	Y_{11}	Y_{12}	…	Y_{1j}			
2	Y_{21}	Y_{22}	…	Y_{2j}			
…	…	…	…	…			
i	Y_{i1}	Y_{i2}	…	Y_{ij}			

- Y_{ij}: j^{th} response of the i^{th} individual
- Different subjects are independent
- Repeated measures on the same subject are correlated

Fit a population average model to show change in Y over time:

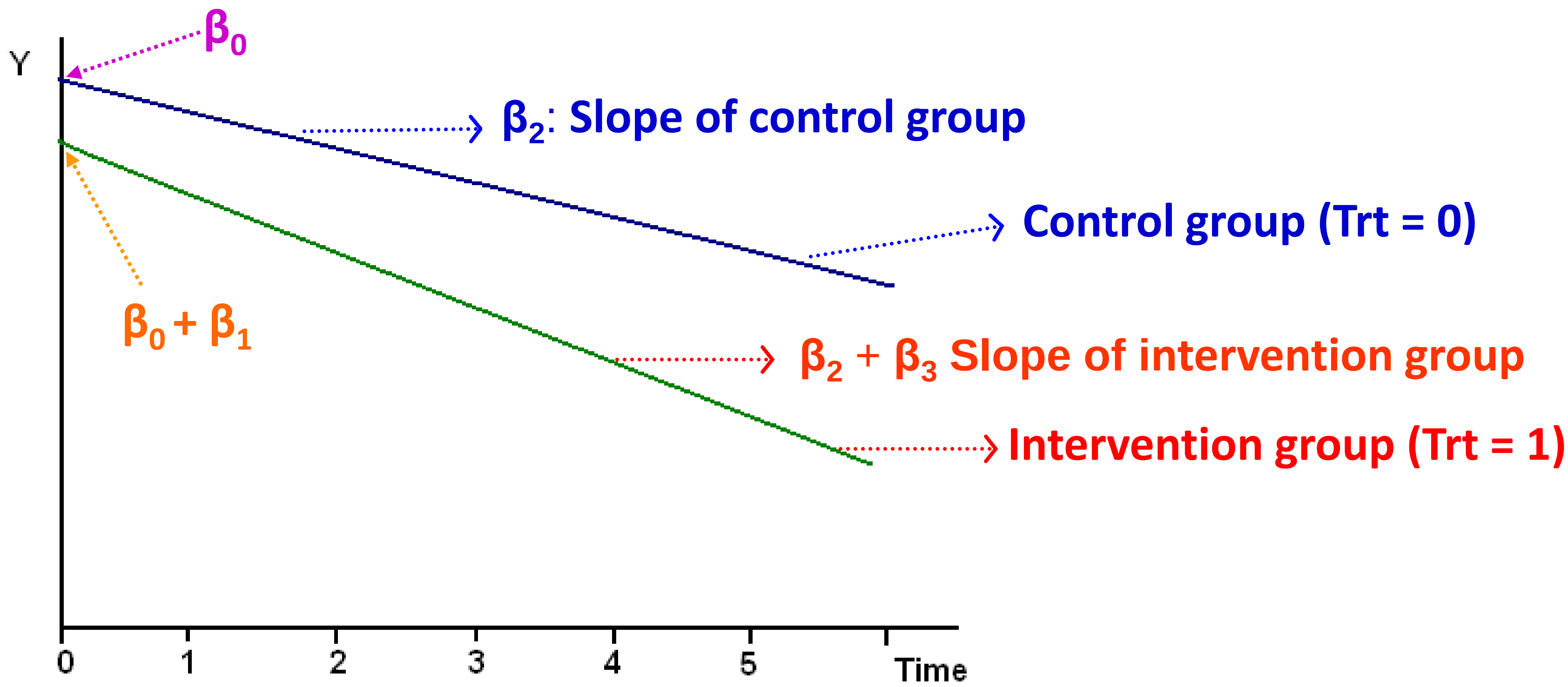
$$Y_{ij} = \beta_0 + \beta_1 \, Trt_i + \beta_2 \, Time_{ij} + \beta_3 \, Trt_i \cdot Time_{ij} + e_{ij}$$

Average difference in Y between Trt = 1 and Trt =0 when Time =0 (baseline)

Rate of change in Y (slope) among Control Group (Trt =0)

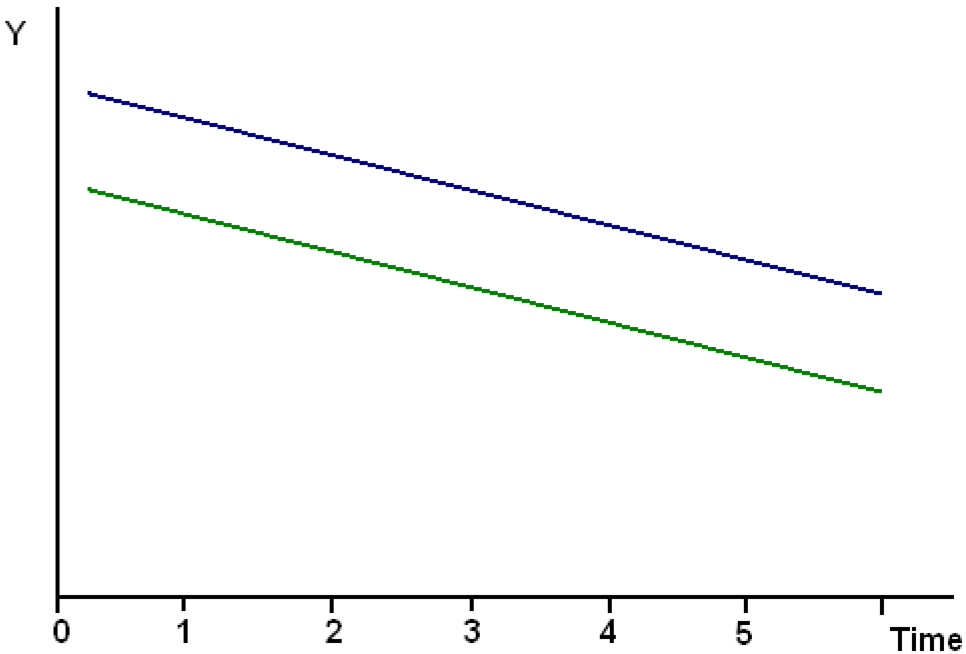
Difference in rate of change (slopes) between Trt = 1 and Trt =0

Population Average Model



$$Y_{ij} = \beta_0 + \beta_1 \, Trt_i + \beta_2 \, Time_{ij} + \beta_3 \, Trt_i \cdot Time_{ij} + e_{ij}$$

Scientific hypotheses

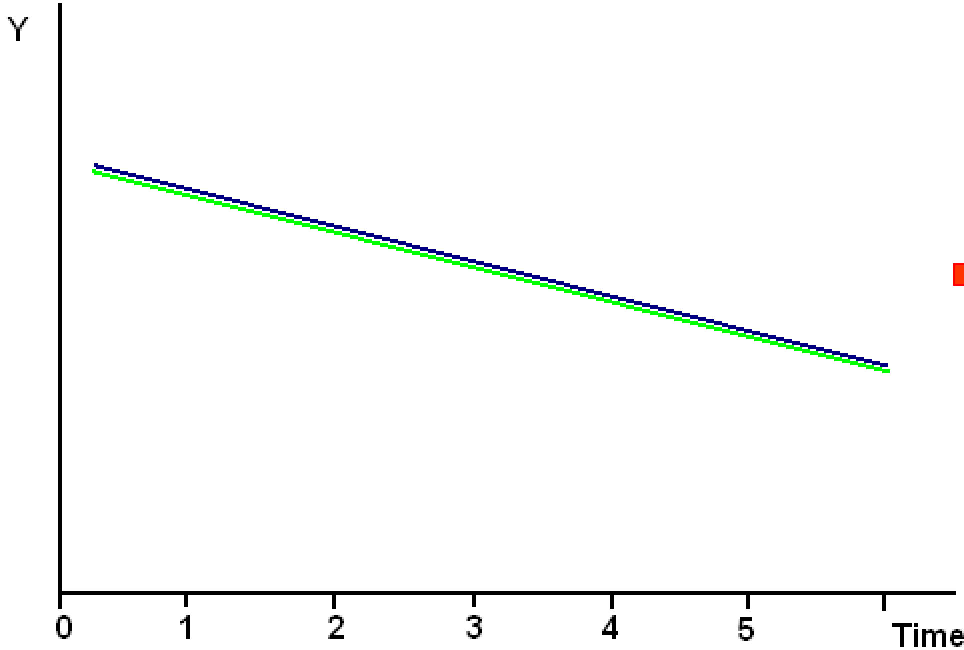


The two lines are parallel: the treatment and control group have the same slope (change rate), i.e. **No group by time interaction**. Note it shows time effect and group effect

Scientific hypotheses

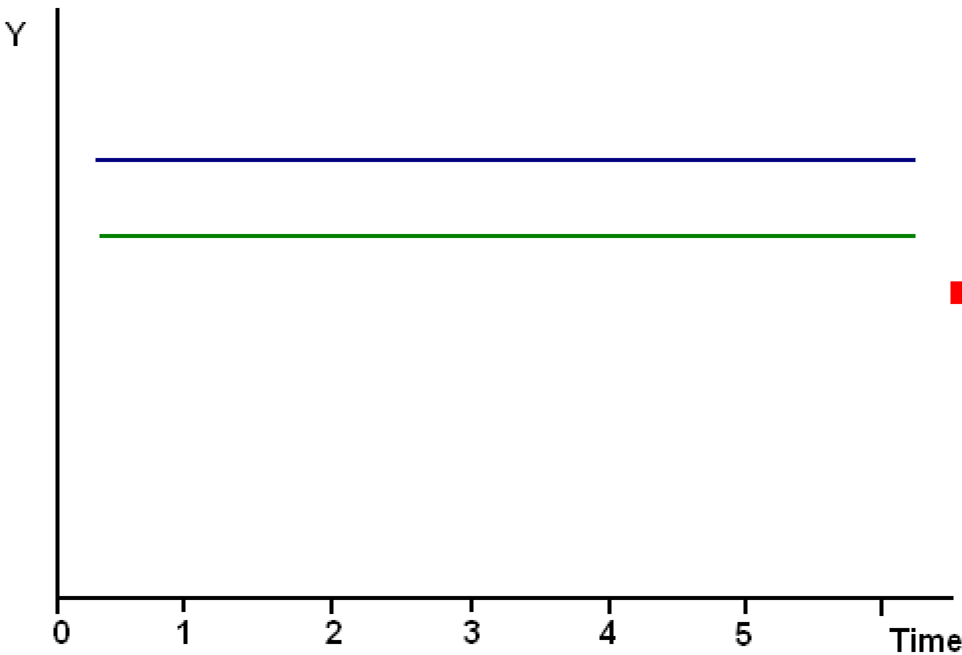
- H01:Are the means similar in the groups, in the sense

is the hypothesis of no



The two lines are overlapping: the treatment and control group have the same response, i.e. **<u>No group effect</u>**

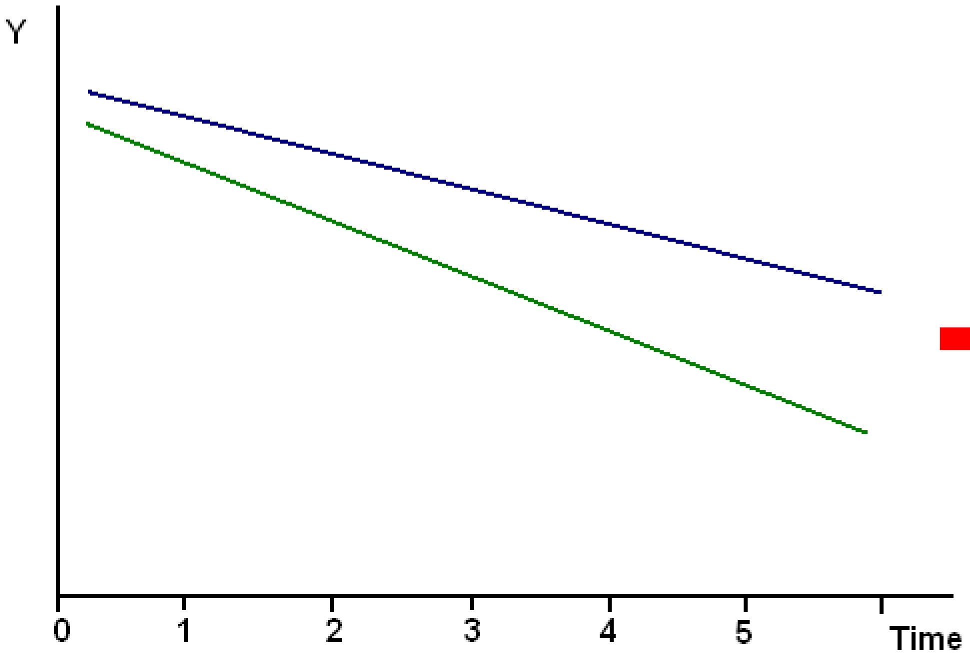
Scientific hypotheses



The two lines are horizontal: the responses of the treatment and control groups do not change over time, i.e. **No time effect**

$\beta 2=0$)

Time, group effects and interaction



Y changes over time, there is **time effect**;
The two lines are not overlapping: **Group effect**;
The two lines are not parallel: **group by time interaction**

Generalized Estimating equations (GEE)

- Developed by Liang and Zeger in 1980s
- Model: use regression model with robust variance estimation, allowing for within individual correlations in response
- Mechanism: Assume a working correlation for the within individual correlation; then estimate the regression coefficients using weighted least squares and the assumed working correlation; then estimate the standard errors robustly

Covariance Structure

- Independent

$$\begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

- Exchangeable (ry)
- Autoregressive
- Unstructured (no specification)

Covariance Structure

- Independent
- Exchangeable (compound symmetry)
- Autoregressive
- Unstructured (no sp

$$\begin{bmatrix} \sigma^2 & a & a \\ a & \sigma^2 & a \\ a & a & \sigma^2 \end{bmatrix}$$

Covariance Structure

- Independent
- Exchangeable (… y)
- Autoregressive

$$\begin{bmatrix} \sigma^2 & \text{ar} & \text{ar}^2 \\ \text{ar} & \sigma^2 & \text{ar} \\ \text{ar}^2 & \text{ar} & \sigma^2 \end{bmatrix}$$

- Unstructured (n…

Covariance Structure

- Independent
- Exchangeable (c)
- Autoregressive
- Unstructured (no specification)

$$\begin{bmatrix} \sigma^2 & a & b \\ a & \sigma^2 & c \\ b & c & \sigma^2 \end{bmatrix}$$

OLS regression

$$\begin{bmatrix} \sigma_{y_{it}}^2 & 0 & 0 \\ 0 & \sigma_{y_{it}}^2 & 0 \\ 0 & 0 & \sigma_{y_{it}}^2 \end{bmatrix}$$

GEE: specify the covariance (correlation structure)

Case Study: change in depressive scores among lung cancer patients

- To examine gender difference in depressive symptoms among lung cancer patients following antidepressant treatment
- 60 lung cancer patients (30 female and 30 male patients) screened for depressive symptoms.
- Depressive score was measured once a month, 6 times over 6 months.

Independent variables include:

- visit time (t)
- gender (gender=1 if female, gender =2 if male)
- baseline age (years)

subject	t1	t2	t3	t4	t5	t6	gender	age
1	22	23	20	22	18	20	1	29
2	32	28	23	22	16	14	2	43
3	17	17	13	11	12	11	1	26
4	18	28	22	17	16	16	1	28
5	16	14	12	7	9	9	2	25
6	24	16	13	12	11	9	1	32

Scientific hypotheses

- H01:Are the trends similar in female and male patients?. (i.e. β3=0: no **group by time interaction**)
- H02:If the trends of female and male are parallel, are they also at the same level? (i.e. β1=0: no **group effec**t)
- H03: If the trends of female and male are parallel, are the means constant over time? (i.e. β2=0: no **time effect**)

$$Y_{ij} = \beta_0 + \beta_1 \, Sex_i + \beta_2 \, Time_{ij} + \beta_3 \, Sex_i \cdot Time_{ij} + \beta_4 \, Age_i + e_{ij}$$

Stata

Wide format

subject	t1	t2	t3	t4	t5	t6	gender	age
1	22	23	20	22	18	20	1	29
2	32	28	23	22	16	14	2	43
3	17	17	13	11	12	11	1	26
4	18	28	22	17	16	16	1	28
5	16	14	12	7	9	9	2	25
6	24	16	13	12	11	9	1	32

- reshape long score, i(subject) j(t)

Long format

subject	t	score	gender	age
1	1	22	1	29
1	2	23	1	29
1	3	20	1	29
1	4	22	1	29
1	5	18	1	29
1	6	20	1	29
2	1	32	2	43
2	2	28	2	43
2	3	23	2	43
2	4	22	2	43
2	5	16	2	43
2	6	14	2	43

Stata

- xtgee

xtgee: GEE
xi: categorical covariates

Gender is a categorical covariates

Interaction of gender & time

```
xi: xtgee score t i.gender age i.gender*t,
fam(gaus) link(iden) i(subject) t(t) corr(exc)
```

Specify family as Gaussian and Link as identity, because Y is continuous & normal

Use subject to identify different individuals

Use t to identify repeated measures within the individual

specify covariance structure as exchangeable

SAS

- proc genmod

Subject ID and Gender are categorical covariates

```
proc genmod data=depress;
class  subject gender;
model score= t gender age gender*t;
repeated subject = subject / type=exch corrw;
run;
```

Use subject to identify repeated measures

Choice of correlation structure

From empirical data:

	t1	t2	t3	t4	t5	t6
t1	1.0000					
t2	0.4982	1.0000				
t3	0.5258	0.8672	1.0000			
t4	0.3933	0.7357	0.7831	1.0000		
t5	0.3674	0.7500	0.8520	0.8449	1.0000	
t6	0.2795	0.6900	0.7967	0.7894	0.9014	1.0000

Independent correlation

Estimated within-subject correlation matrix R:

	c1	c2	c3	c4	c5	c6
r1	1.0000					
r2	0.0000	1.0000				
r3	0.0000	0.0000	1.0000			
r4	0.0000	0.0000	0.0000	1.0000		
r5	0.0000	0.0000	0.0000	0.0000	1.0000	
r6	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Compound Symmetry / Exchangeable

	c1	c2	c3	c4	c5	c6
r1	1.0000					
r2	0.5743	1.0000				
r3	0.5743	0.5743	1.0000			
r4	0.5743	0.5743	0.5743	1.0000		
r5	0.5743	0.5743	0.5743	0.5743	1.0000	
r6	0.5743	0.5743	0.5743	0.5743	0.5743	1.0000

Auto regressive

	c1	c2	c3	c4	c5	c6
r1	1.0000					
r2	0.6862	1.0000				
r3	0.4708	0.6862	1.0000			
r4	0.3231	0.4708	0.6862	1.0000		
r5	0.2217	0.3231	0.4708	0.6862	1.0000	
r6	0.1521	0.2217	0.3231	0.4708	0.6862	1.0000

0.6862^2

0.6862^3

0.6862^4

0.6862^5

Unstructured

	c1	c2	c3	c4	c5	c6
r1	1.0000					
r2	0.5206	1.0000				
r3	0.3927	0.8512	1.0000			
r4	0.3483	0.7289	0.6795	1.0000		
r5	0.2834	0.7379	0.7517	0.7804	1.0000	
r6	0.1593	0.5772	0.5901	0.6405	0.7408	1.0000

Choice of correlation structure

- Compared to the empirical correlation matrix
- Test by comparing the goodness of fit:

    -2 Res Log Likelihood (likelihood ratio test)

    AIC (smaller is better)

    BIC (smaller is better)

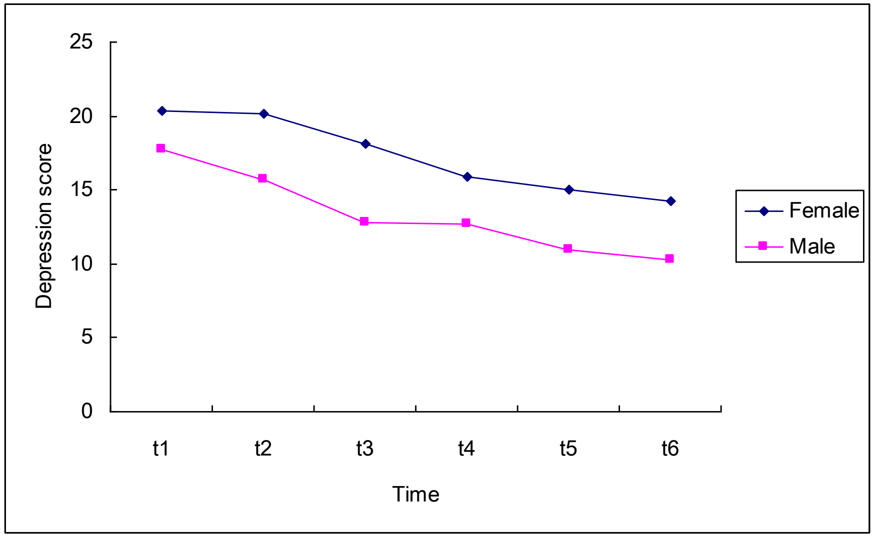
Decision: unstructured

score	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
t	-1.320369	.2473101	-5.34	0.000	-1.805087	-.8356498
_gender_1	-3.613827	1.557062	-2.32	0.020	-6.665611	-.5620418
age	.2351761	.1031182	2.28	0.023	.0330682	.4372841
_genderXt_1	-.00971	.3357138	-0.03	0.977	-.667697	.6482771
_cons	14.34012	3.604429	3.98	0.000	7.275564	21.40467

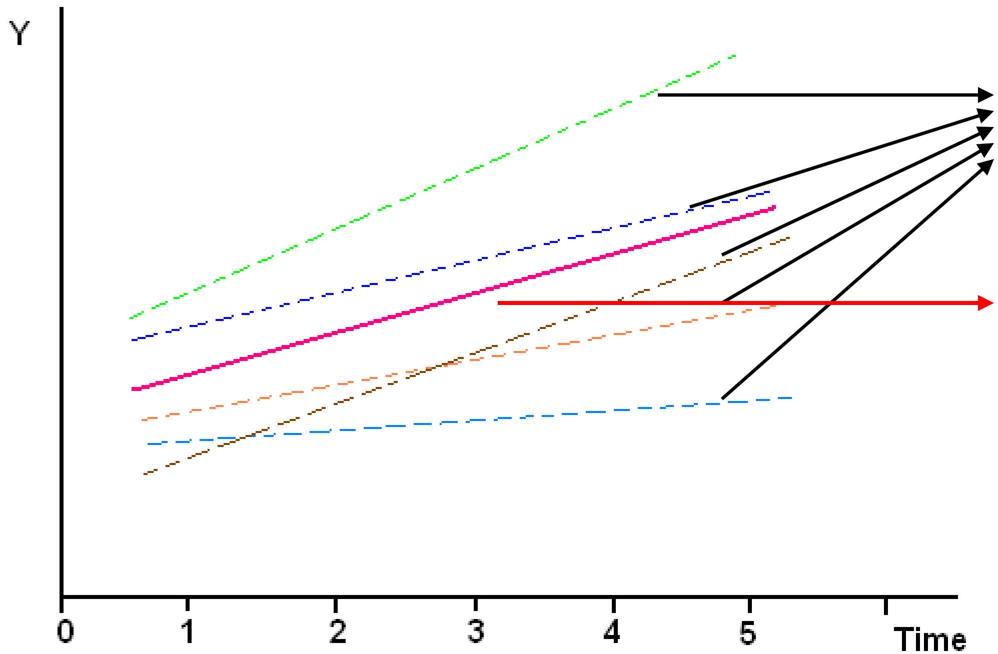
No group by time effect

Time effect: depression score decreases by 1.3 each month.

Group effect: Male < Female



3. SUBJECT SPECIFIC MODEL



These are subject-specific regression lines (random effects)

This is the population average (i.e. fixed effect)

Population Average (fixed effects):

$$Y_{ij} = \beta_0 + \beta_1 \text{Trt}_i + \beta_2 \text{Time}_{ij} + \beta_3 \text{Trt}_i \cdot \text{Time}_{ij} + e_{ij}$$

Subject Specific (random effects):

$$Y_{ij} = \beta_0 + \mathbf{b_{0i}} + \beta_1 \text{Trt}_i + \beta_2 \text{Time}_{ij} + \mathbf{b_{2i} Time_{ij}} + \beta_3 \text{Trt}_i \cdot \text{Time}_{ij} + e_{ij}$$

Random intercept

Subject i's departure from average baseline measurement

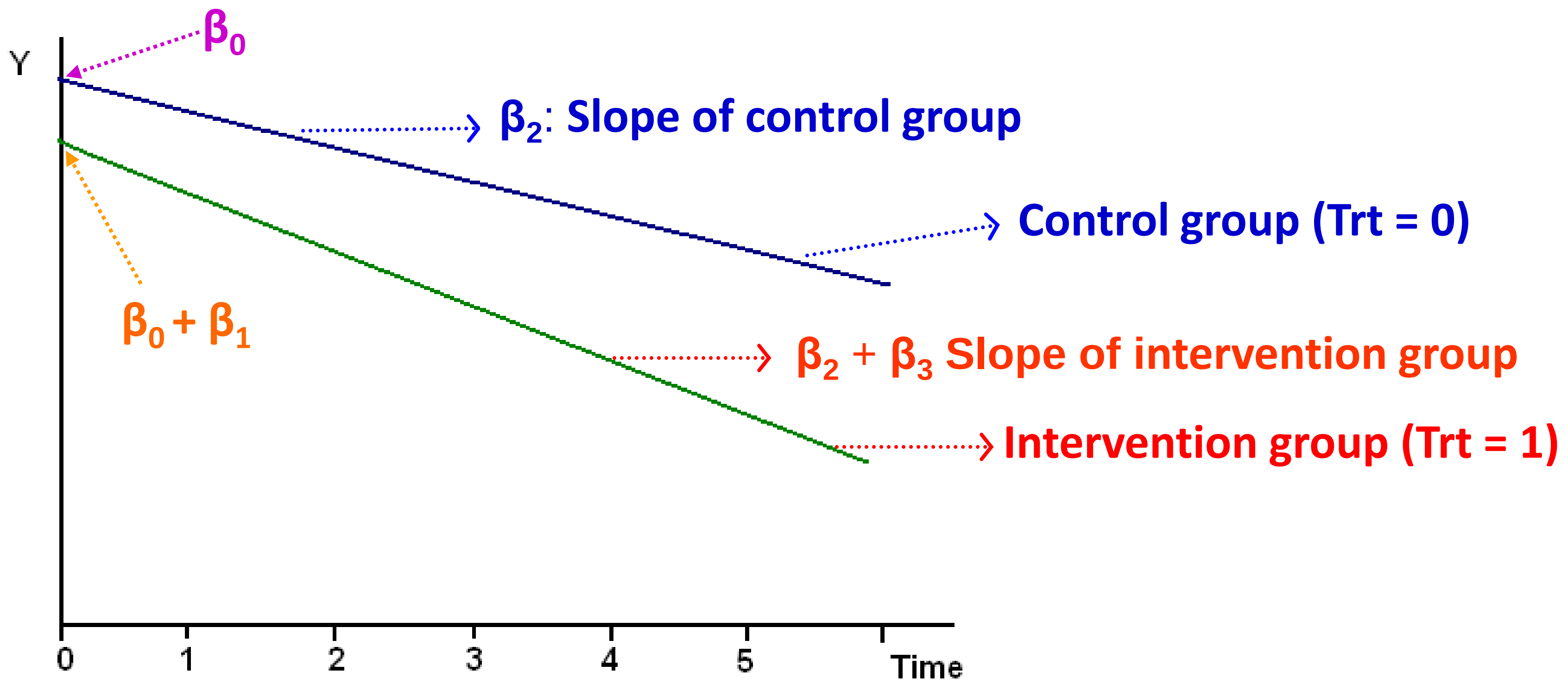
Random slope

Subject i's departure from average slope

Mixed Effects Model

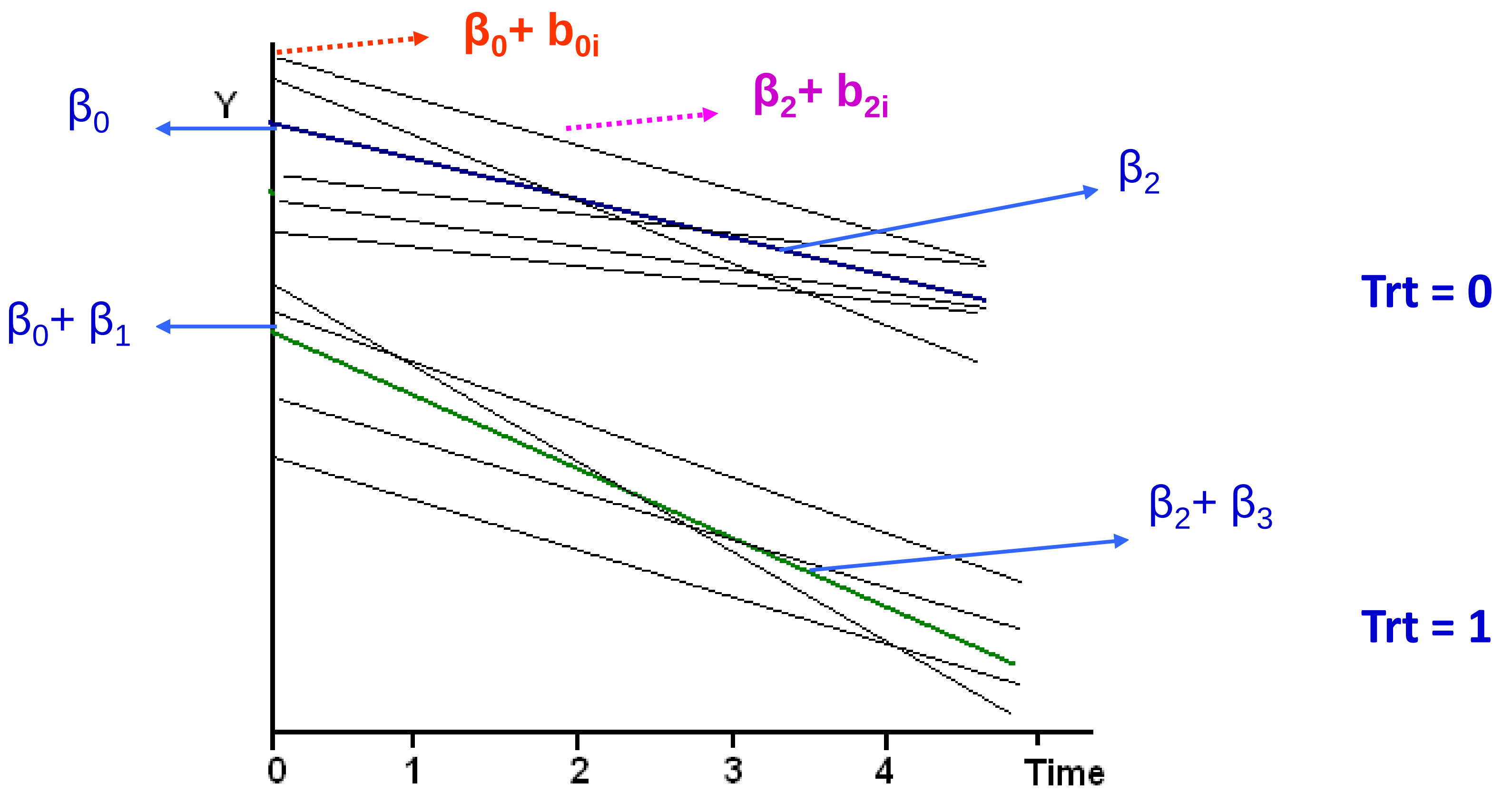
- Model both fixed effects (i.e. population average) and random effects (i.e. subject/individual specific)

Population Average Model



$$Y_{ij} = \beta_0 + \beta_1 \, Trt_i + \beta_2 \, Time_{ij} + \beta_3 \, Trt_i \cdot Time_{ij} + e_{ij}$$

Subject Specific (Mixed Effects) Model



$$Y_{ij} = \beta_0 + \mathbf{b_{0i}} + \beta_1 Trt_i + \beta_2 Time_{ij} + \mathbf{b_{2i} Time_{ij}} + \beta_3 Trt_i \cdot Time_{ij} + e_{ij}$$

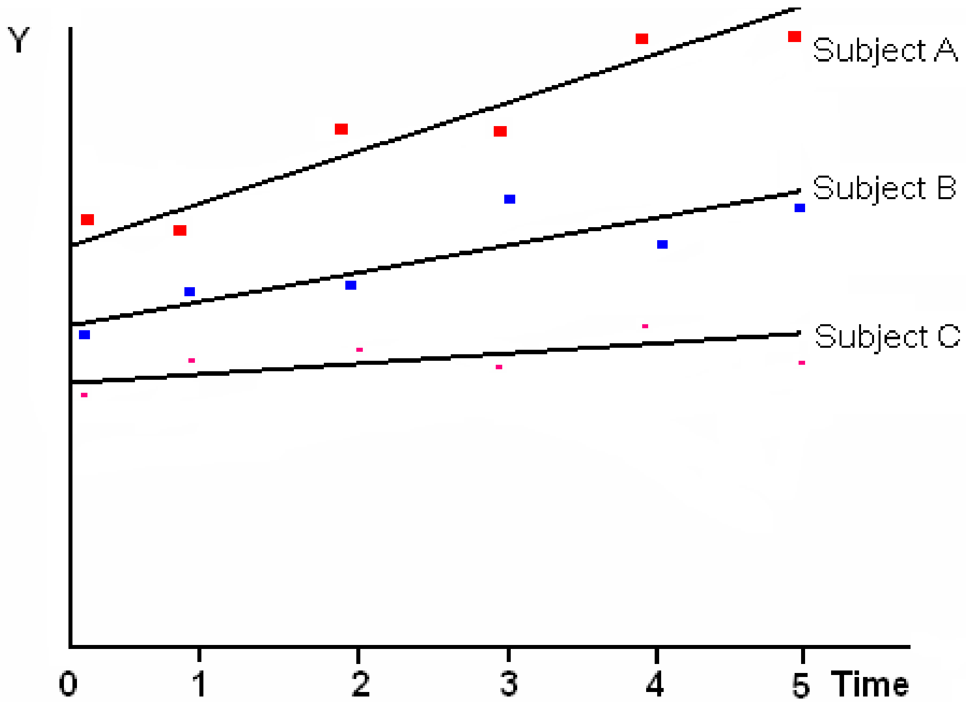
Subject Specific (Mixed Effects) Model

$$Y_{ij} = \beta_0 + \mathbf{b_{0i}} + \beta_1 Trt_i + \beta_2 Time_{ij} + \mathbf{b_{2i} Time_{ij}} + \beta_3 Trt_i \cdot Time_{ij} + e_{ij}$$

$$e_{ij} \sim N(0, \sigma^2_{ij})$$

Error or residual: variation of Y value of subject i from its average at time j (i.e. fitted regression line for subject i)

It describes **within-subject random errors**



Subject Specific (Mixed Effects) Model

$Y_{ij} = \beta_0$ **+ b_{0i}** $+ \beta_1 Trt_i + \beta_2 Time_{ij}$ **+ $b_{2i}Time_{ij}$** $+ \beta_3 Trt_i \cdot Time_{ij} + e_{ij}$

$$b_{0i} \sim N(\beta_0, \sigma^2_{b_0})$$

Variability in intercepts between subjects

$$\mathrm{b}_{2\mathrm{i}} \sim \mathrm{N}(\beta_2, \sigma^2_{\mathrm{b}2})$$

Variability in slopes between subjects

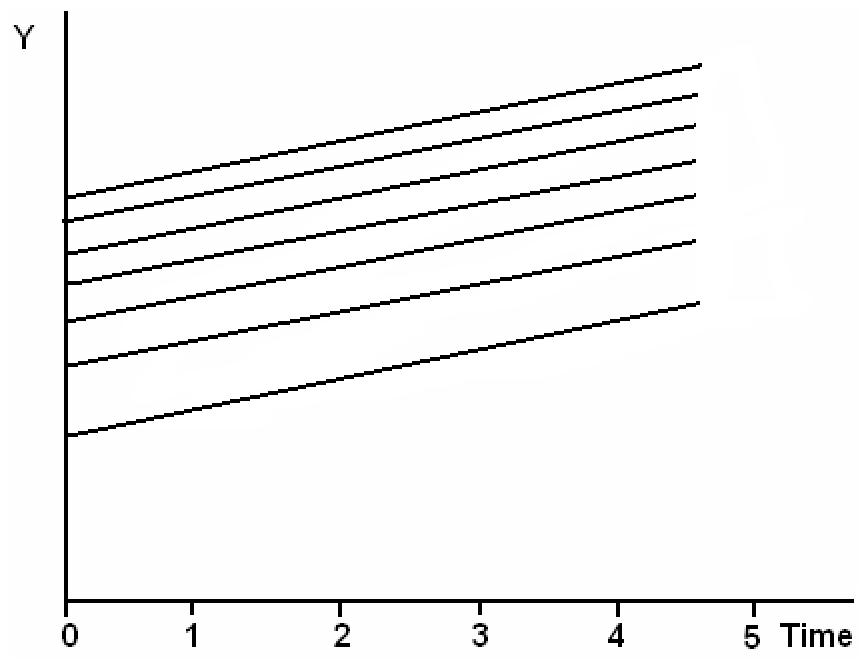
$$G = \begin{bmatrix} \sigma^2_{b_0} & Cov_{b_0 b_2} \\ Cov_{b_0 b_2} & \sigma^2_{b_2} \end{bmatrix}$$

Variance of intercepts: Deviations of individual intercepts from the average intercept

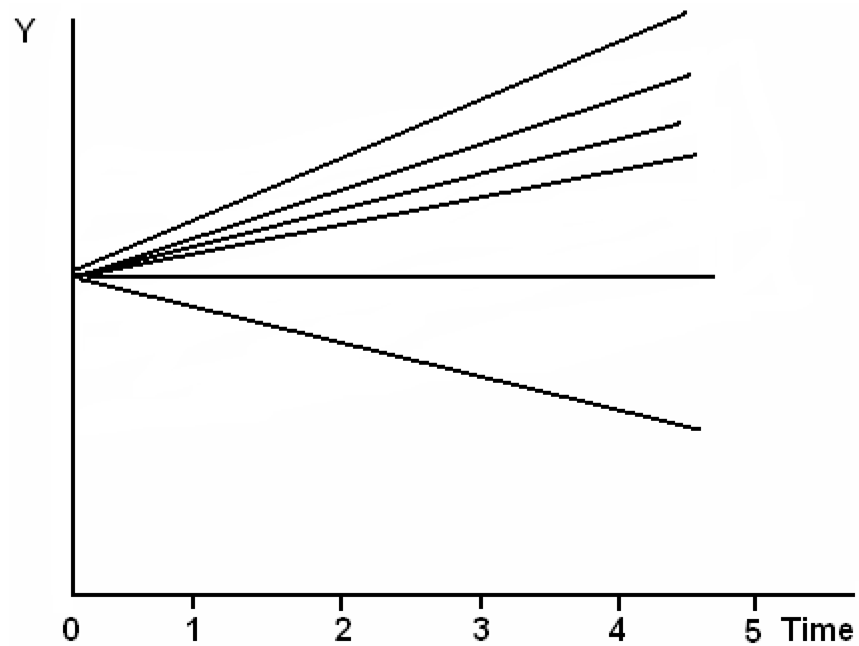
Covariance between intercepts and slopes

Variance of slopes: Deviations of individual slopes from the average slope

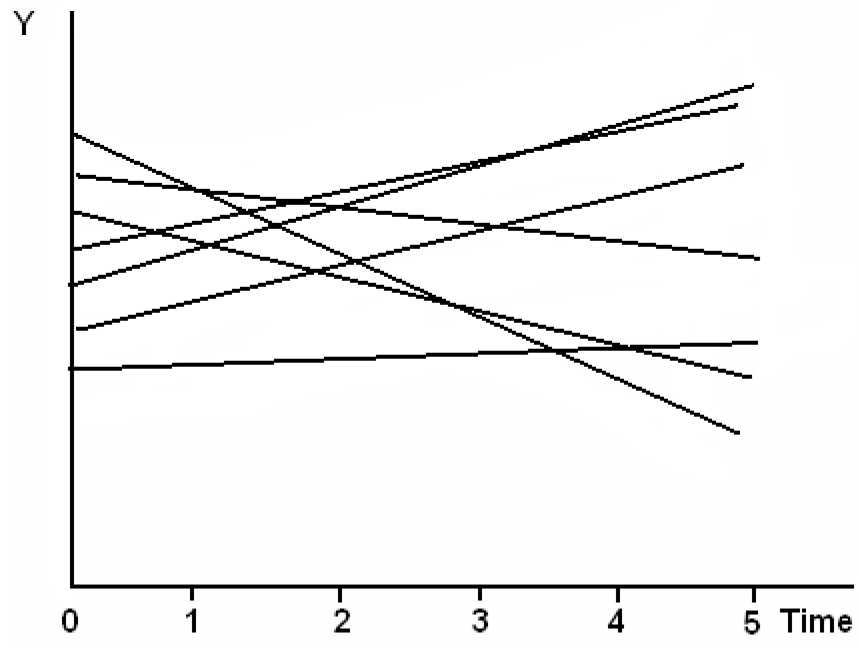
$$G = \begin{bmatrix} \sigma_{b_0}^2 & 0 \\ 0 & 0 \end{bmatrix}$$



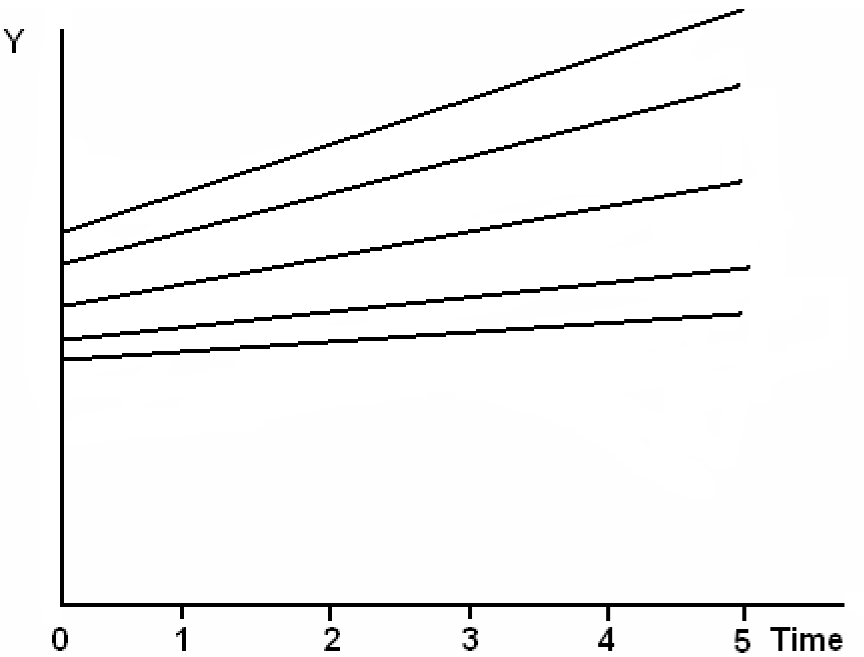
$$G = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{b_2}^2 \end{bmatrix}$$



$$G = \begin{bmatrix} \sigma_{b_0}^2 & 0 \\ 0 & \sigma_{b_2}^2 \end{bmatrix}$$



$$G = \begin{bmatrix} \sigma^2_{b_0} & Cov_{b_0 b_2} \\ Cov_{b_0 b_2} & \sigma^2_{b_2} \end{bmatrix}$$



Case Study: RCT of antidepressant

We are interested to know the effects of antidepressant on cancer patients' depressive score. A total of 20 cancer patients were randomly assigned to receive antidepressant and placebo. They were measured depressive score for four times.

ID	Trt	Y1	Y2	Y3	Y4
1	1	218	206	176	194
2	1	228	228	224	210
3	1	226	216	196	206
4	1	192	188	198	194
5	1	216	220	192	208
6	1	220	212	220	214
7	1	226	220	212	206
8	1	224	216	198	216
9	1	242	222	192	230
10	1	196	206	196	214
11	0	226	238	202	228
12	0	246	234	194	206
13	0	224	218	216	228
14	0	198	196	156	176

Scientific hypotheses

- H01:Are the trends similar in intervention and control groups? (i.e. $\beta 3=0$)
- H02:If the trends of the intervention and control groups are parallel, are they also at the same level? (i.e. $\beta 1=0$)
- H03: If the trends of the intervention and control groups are parallel, are the means constant over time? (i.e. $\beta 2=0$)
- In addition, we are interested in whether there are any differences between subjects on the trend.

$$Y_{ij} = \beta_0 \mathbf{+ b_{0i}} + \beta_1 Trt_i + \beta_2 Time_{ij} \mathbf{+ b_{2i} Time_{ij}} + \beta_3 Trt_i \cdot Time_{ij} + e_{ij}$$

SAS

```
data long;
set wide;
y=T1; t=1; output;
y=T2; t=2; output;
y=T3; t=3; output;
y=T4; t=4; output;
drop T1 - T4;
run;

proc mixed data = long covtest;
class ID Trt;
model y=Trt t t*Trt/ s chisq;
random intercept t/ type=un sub=ID g v vcorr;
run;
```

Change to long format

It gives hypothesis testing on the random effects

It gives regression coefficients and Wald test

This specifies random intercept and slope

G matrix as unstructured

G matrix, e_{ij}, and correlation matrixes for random effects

Estimated G Matrix

Row	Effect	ID	Col1	Col2
1	Intercept	1	107.30	22.4899
2	t	1	22.4899	0.2408

$$G = \begin{bmatrix} \sigma^2_{b_0} & Cov_{b_0 b_2} \\ Cov_{b_0 b_2} & \sigma^2_{b_2} \end{bmatrix}$$

$\sigma^2_{b_0}$ Variability in intercept

$Cov_{b_0 b_2}$ Greater slopes for greater intercept

$\sigma^2_{b_2}$ Variability in slope

σ^2_{ij} variability of a subject's score about his/her own regression line

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	ID	107.30	107.91	0.99	0.1600
UN(2,1)	ID	22.4899	26.7820	0.84	0.4011
UN(2,2)	ID	0.2408	10.2476	0.02	0.4906
Residual		126.83	28.3600	4.47	<.0001

Subjects do not differ in baseline and slope; no correlation between intercept and slopes

Total variability in Y: between subject variability + within subject random error=107.3+0.2408+126.83 = 234.3708

54% of variability in depressive score is explained by within-subject variation. More within-subject variation than between-subject variation.

The Mixed Procedure

Solution for Fixed Effects

Effect	Trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		220.90	5.4548	18	40.50	<.0001
Trt	0	6.9000	7.7142	40	0.89	0.3764
Trt	1	0	.	.	.	.
t		-4.1800	1.6002	18	-2.61	0.0176
t*Trt	0	-3.4000	2.2630	40	-1.50	0.1408
t*Trt	1	0	.	.	.	.

$$Y_{ij} = \beta_0 + \mathbf{b_{0i}} + \beta_1 Trt_i + \beta_2 Time_{ij} + \mathbf{b_{2i} Time_{ij}} + \beta_3 Trt_i \cdot Time_{ij} + e_{ij}$$

No group*time trend
No group difference
Strong effect of time.

GEE

```
proc genmod data=long;
class ID Trt;
model y=Trt t t*Trt;
repeated subject = ID / type=exch;
run;
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		220.9000	5.1739	210.7594	231.0406	42.70	<.0001
Trt	0	6.9000	7.3188	-7.4446	21.2446	0.94	0.3458
Trt	1	0.0000	0.0000	0.0000	0.0000	.	.
t		-4.1800	1.3578	-6.8412	-1.5188	-3.08	0.0021
t*Trt	0	-3.4000	2.1470	-7.6080	0.8080	-1.58	0.1133
t*Trt	1	0.0000	0.0000	0.0000	0.0000	.	.

GEE vs. Mixed Effects model

Parameter	**Population Average** (GEE exchangeable)	**Subject Specific** (Random b_0 & $b_{1)}$
Trt	6.9 (7.32)	6.9 (7.71)
Time	-4.1 (1.36)	-4.2 (1.60)
Trt*time	-3.4 (2.15)	-3.4 (2.26)

Mixed model is close to GEE with exchangeable correlation:
Nearly identical coefficients and slightly different SEs

GEE vs. Mixed Effects Model

- GEE (population average model):
  - On average, is there a trend in score change over time?
  - Robust: even if correlation model is wrong, SE still valid
- Mixed effects (subject specific model):
  - Are there any differences between subjects on the trend in score change over time? (Do all subjects have the same trend over time?)
  - Advantages
    - Characterization of heterogeneity
    - Incomplete unbalanced data

Extension

- Generalized linear model: outcome can be binary, counts, etc.
- Multilevel data analysis:
  - For longitudinal data with repeated measures, data are clustered within the subject
  - Studies by families (data are clustered within family)
  - Studies by school, clinic, school district, etc. (data are clustered within these levels)

correlation or dependencies in the data