Longitudinal Data Analysis

Samaneh Asgari

August, 2016

Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Sciences (RIES), Shahid Beheshti University of Medical Sciences, Tehran, Islamic Republic of Iran.

Example with time-dependent, continuous predictor…

The level of increase/decrease in wrist circumference (cm) with systolic blood pressure (sbp; mm Hg) was evaluated for 6 participant. At baseline, all 6 participant have similar levels of SBP and wrist circumference (cm)>=14. Researchers measure wrist and SBP levels at three subsequent time points: at 3 years, 6 years, and 9 years post-baseline.

Here are the data in broad form:

id	wrist1	wrist2	wrist3	wrist4	sbp1	sbp2	sbp3	sbp4
1	20	18	15	20	100	110	120	130
2	22	24	18	22	100	100	100	95
3	14	10	24	10	100	199	80	170
4	38	34	32	34	100	110	115	110
5	25	29	25	29	100	100	105	101
6	30	28	26	14	100	110	111	150

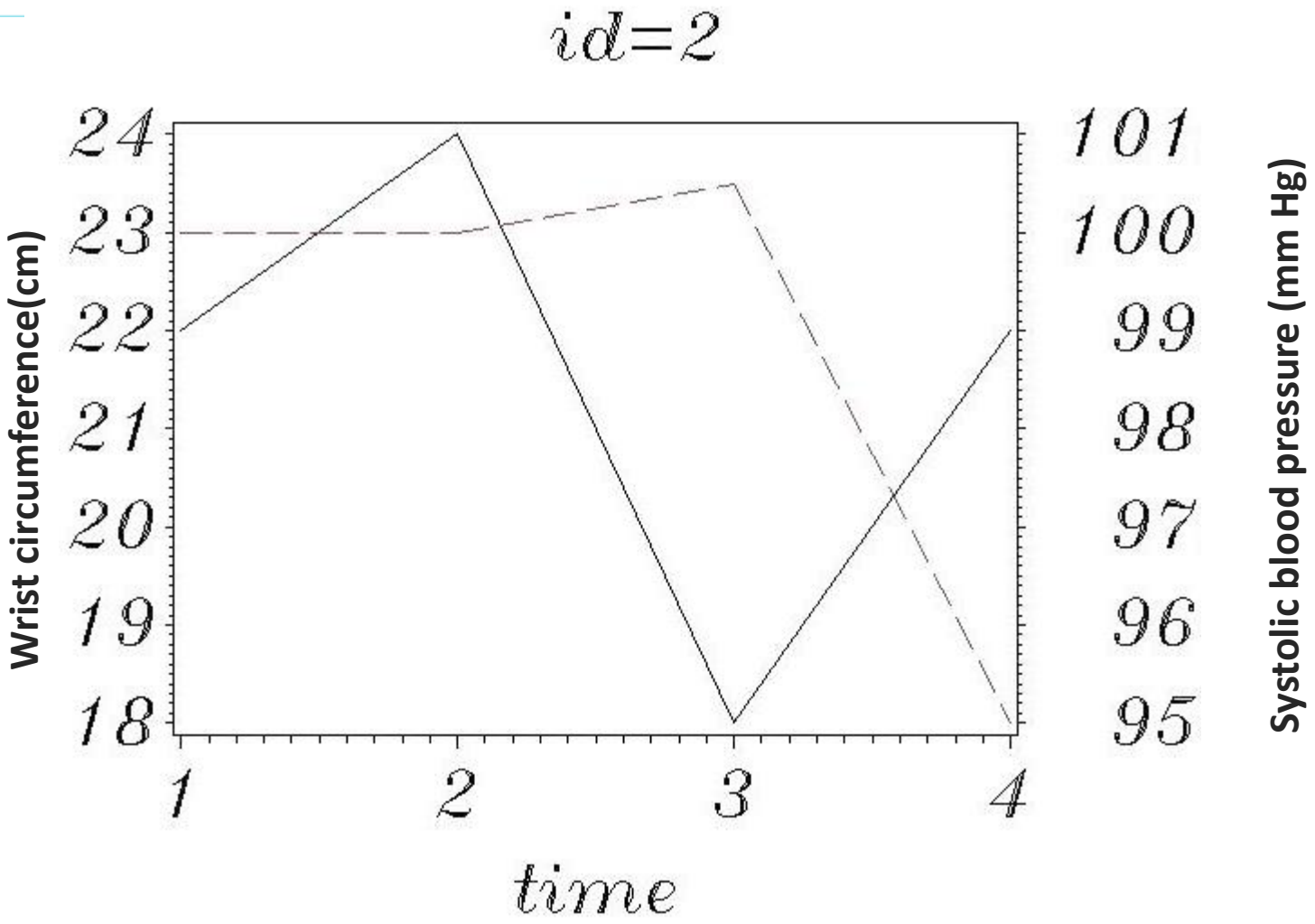
Data in long form:

Wide N=6

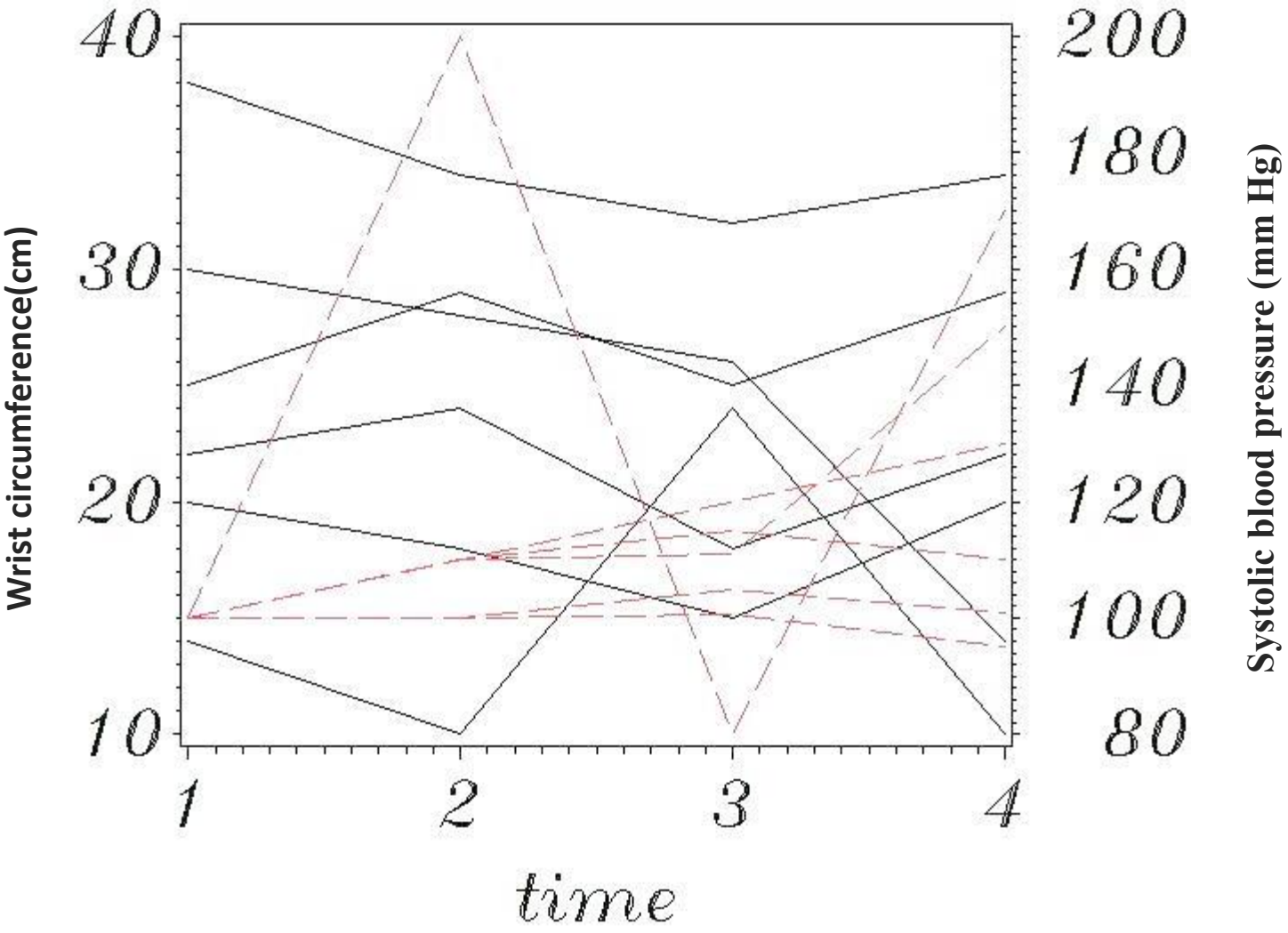
Long N=6*4
=24

id	time	wrist	sbp
1	0	20	100
1	3	18	110
1	6	15	120
1	9	20	130
2	0	22	100
2	3	24	100
2	6	18	100
2	9	22	95
3	0	14	100
3	3	10	199
3	6	24	80
3	9	10	170
4	0	38	100
4	3	34	110
4	6	32	115
4	9	34	110
5	0	25	100
5	3	29	100
5	6	25	105
5	9	29	101
6	0	30	100
6	3	28	110
6	6	26	111
6	9	14	150

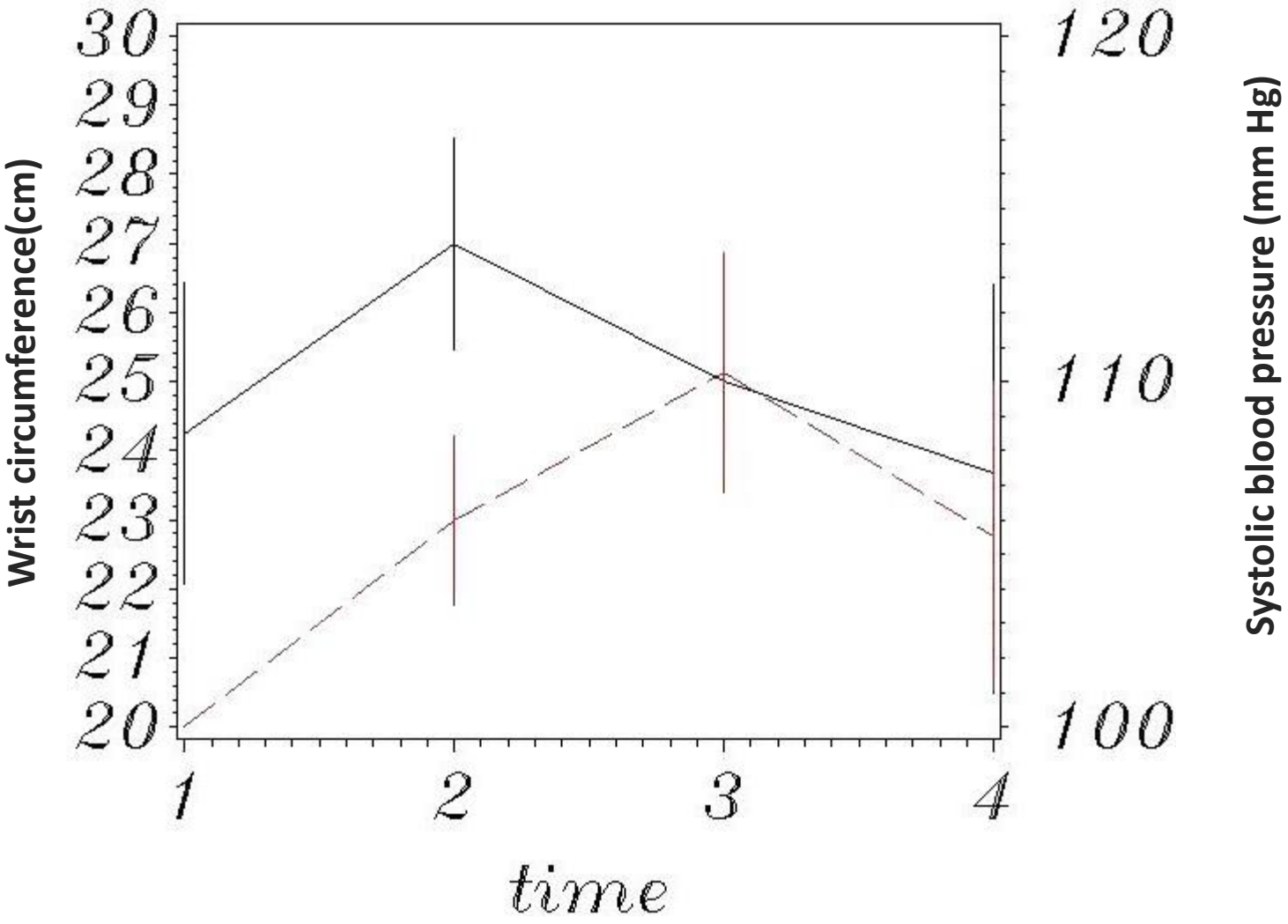
Trajectory plot



Trajectory plot of all 6 subjects at once:



Mean SBP levels compared with mean wrist:



How do you analyze these data?

- ✓ Marginal or population averaged models
- ✓ Random-effects or subject-specific models

But first…naïve analysis…

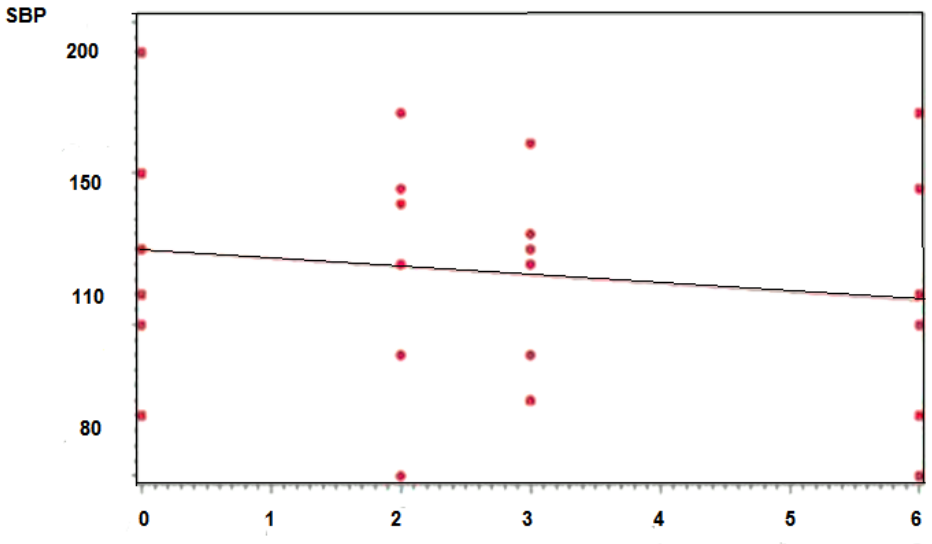
- The data in long form could be naively thrown into an ordinary least squares (OLS) linear regression…
- i.e., look for a linear correlation between SBP and wrist ignoring the correlation between subjects. (the cheating way to get 4-times as much data!)
- Can also look for a linear correlation between SBP and time.

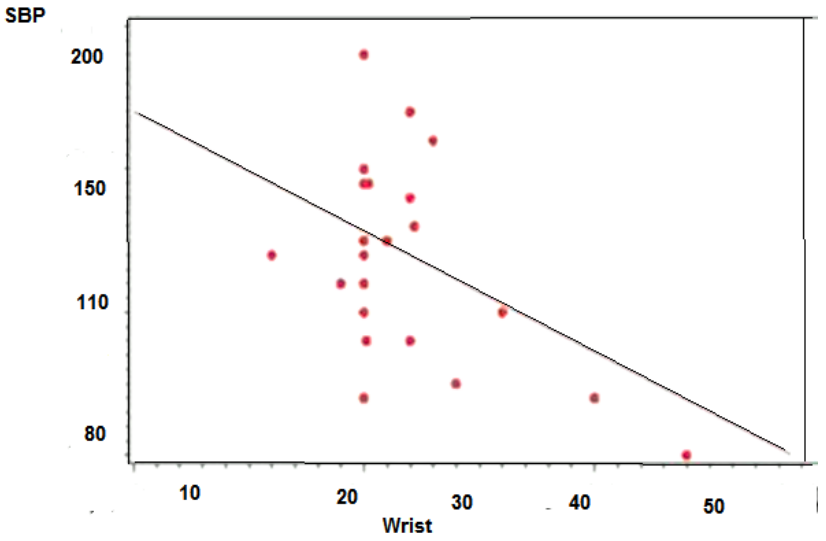
Graphically…

Naïve linear regression here looks for significant slopes (ignoring correlation between individuals):

$\hat{Y}$= 24.90889 - 0.557778*time.

$\hat{Y}$=42.44831-0.01685*wrist





N=24—as if we have 24 independent observations!

The linear regression model:

$$Y_i = \beta_0 + \beta_{wrist}(wrist_i) + \beta_{time}(time_i) + Error_i$$

$$\hat{Y}_i = 42.46803 - .01704(wrist_i) + .07466(time_i)$$

1-unit increase in wrist is associated with a .0174 decrease in SBP(1.7 points per 100 units SBP)

Each year is associated only with a 0.07 increase in SBP, after correcting for wrist changes.

Classical statistical method is failed in Longitudinal data!

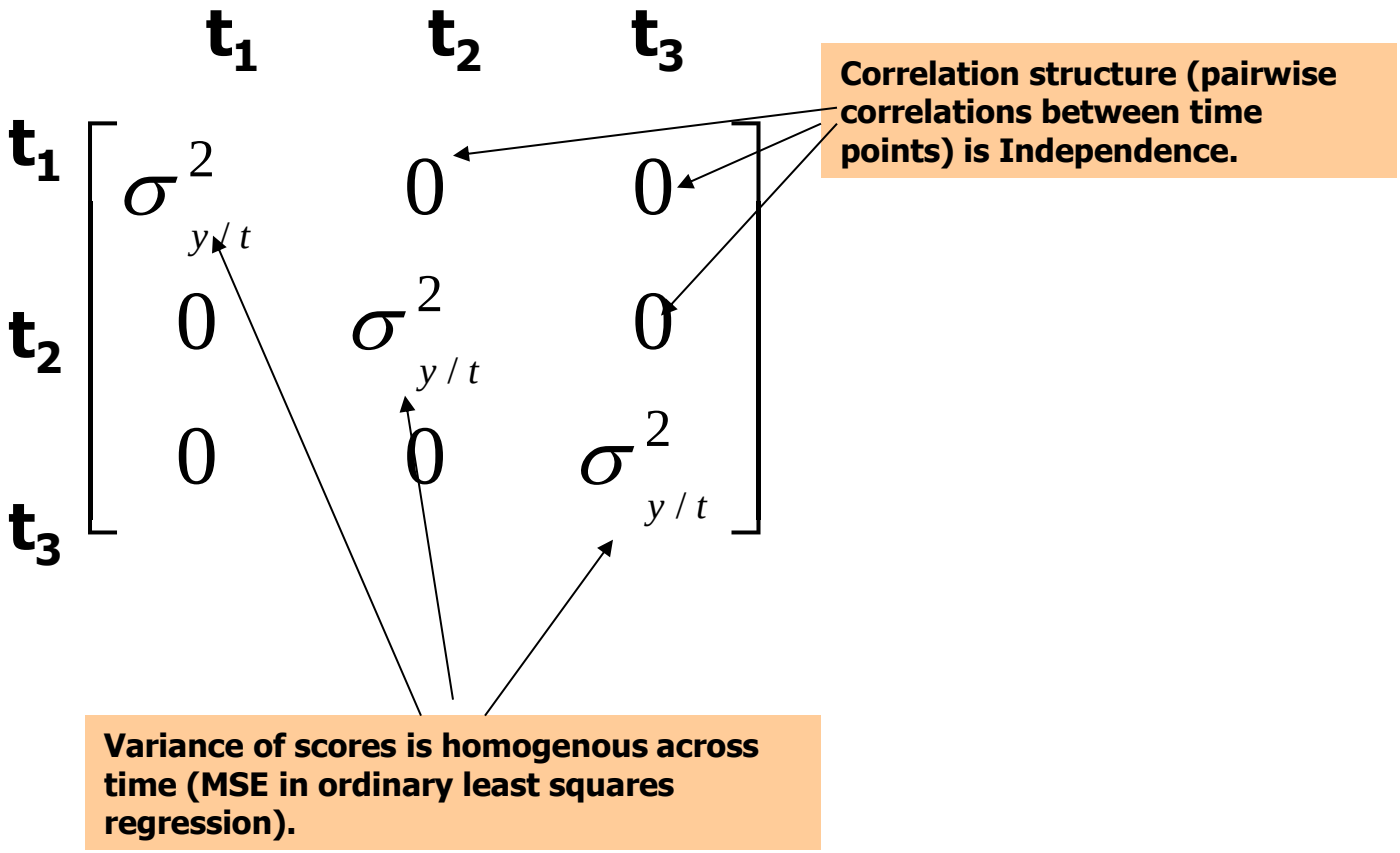
observations are correlated within subjects

This means:

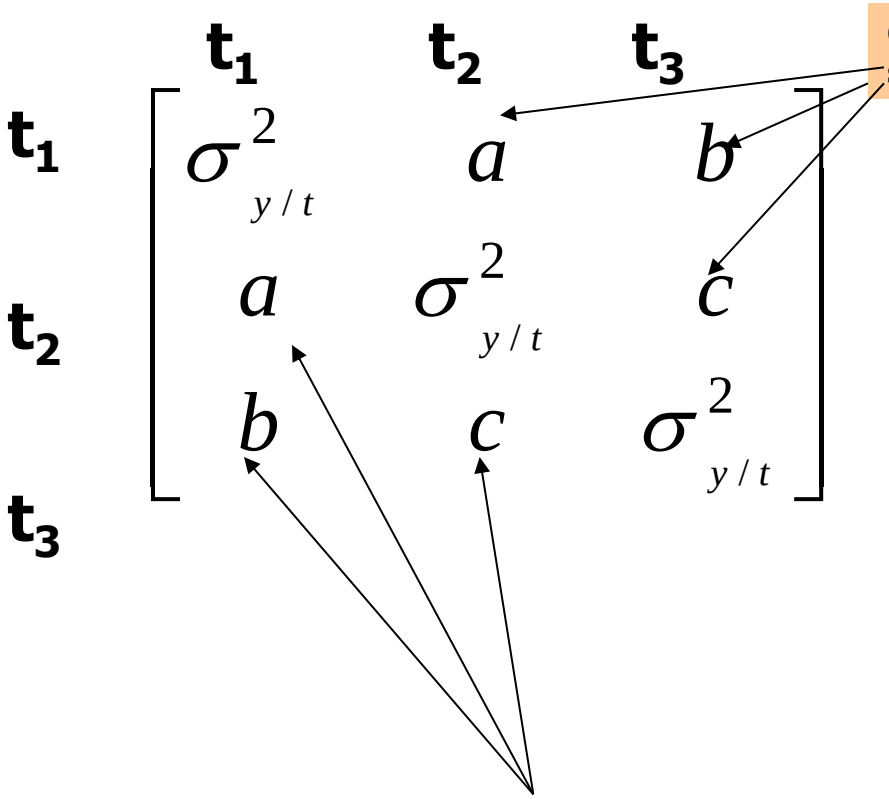
Each person measured several times and its current measurement is dependent to her/his previous measurement.

Between subject variance+
Within subject variance

OLS regression variance-covariance matrix



variance-covariance matrix



Choice of the correlation structure

- Independent (naïve analysis)
- Exchangeable
- Autoregressive
- Unstructured

Independence

	t_1	t_2	t_3
t_1	–	0	0
t_2	0	–	0
t_3	0	0	–

Exchangeable

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array}
\begin{array}{c}
\begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\
\begin{bmatrix} - & \rho & \rho \\ \rho & - & \rho \\ \rho & \rho & - \end{bmatrix}
\end{array}
$$

Autoregressive

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\
\begin{bmatrix} - & \rho & \rho^2 & \rho^3 \\ \rho & - & \rho & \rho^2 \\ \rho^2 & \rho & - & \rho \\ \rho^3 & \rho^2 & \rho & - \end{bmatrix}
\end{array}
$$

Unstructured

$$
\begin{array}{c} \\ \mathbf{t_1} \\ \mathbf{t_2} \\ \mathbf{t_3} \\ \mathbf{t_4} \end{array}
\begin{array}{c}
\begin{array}{cccc} \mathbf{t_1} & \mathbf{t_2} & \mathbf{t_3} & \mathbf{t_4} \end{array} \\
\begin{bmatrix} - & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & - & \rho_5 & \rho_4 \\ \rho_2 & \rho_5 & - & \rho_6 \\ \rho_3 & \rho_4 & \rho_6 & - \end{bmatrix}
\end{array}
$$

How do you analyze these data?

- ✓ Marginal or population averaged models
- ✓ Random-effects or subject-specific models

Marginal models

- They directly model the mean response at each occasion $E(Y_{ij}|X_{ij})$ using an appropriate link function.
- Marginal models do not necessarily require full distributional assumptions for the vector of repeated responses, only a regression model for the mean response.

- Because marginal models separately parameterize the model for the mean responses from the model for the within-subject association, Liang and Zeger (1986) recognized that it is possible to estimate the regression parameters in the former without making full distributional assumptions, therefore they proposed the Generalized Estimating Equations (GEE) approach.

Generalized Estimating Equations (GEE) approach

- ✓ GEE takes into account the dependency of observations by specifying a "working correlation structure."

- ✓ The application of the GEE approach is that it only requires specification of that part of the probability mechanism that is of scientific interest, the marginal means.

The model…

CORR

$$\begin{bmatrix} SBP1 \\ SBP2 \\ SBP3 \\ SBP4 \end{bmatrix} = \beta_0 + \begin{bmatrix} Wrist1 \\ Wrist2 \\ Wrist3 \\ Wrist4 \end{bmatrix} \beta_{12}(time) + Error$$

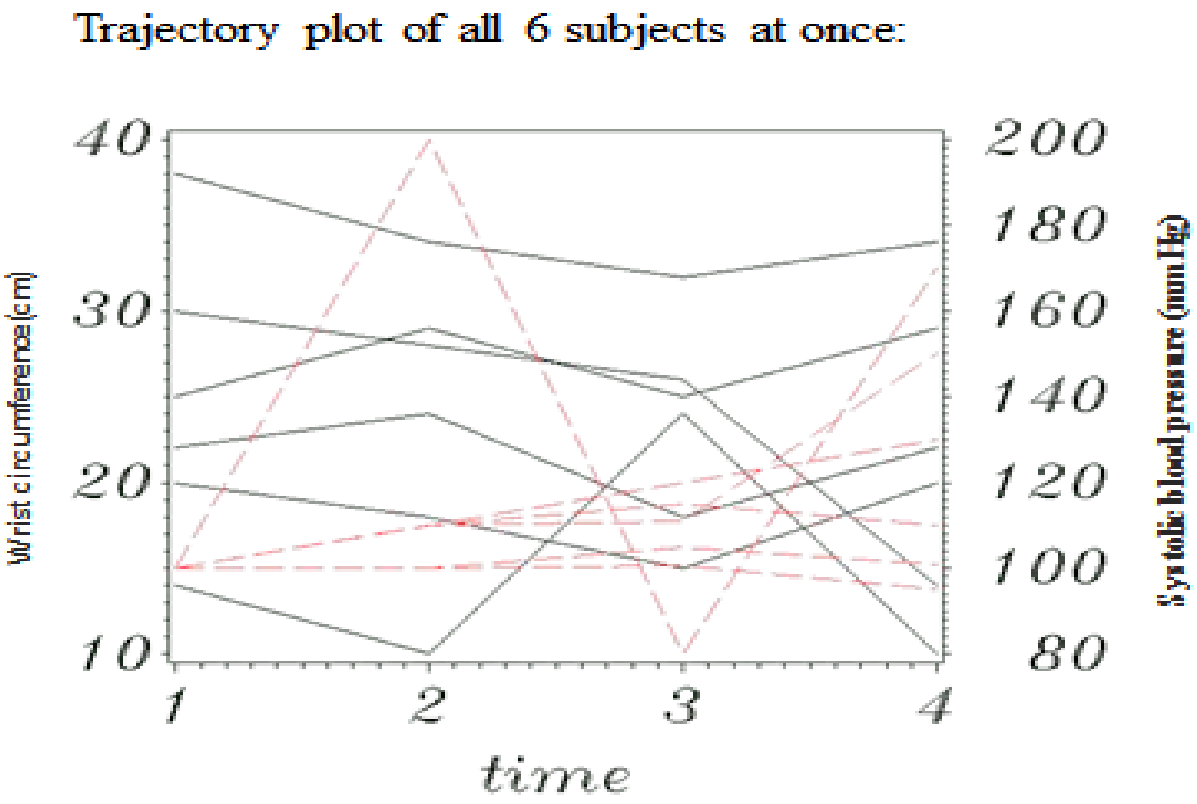
Measures linear correlation between SBP and wrist across all 4 time periods. Vectors!

Measures linear correlation between time and wrist.

CORR represents the correction for correlation between observations.

$$\hat{Y}_i = 38.24 - .0129(wrist_i) + .0775(time_i)$$

Introduction to Mixed Models



- ✓ response depends not only on covariates but also on a vector of random effects, where the mean response depends also on previous responses.

Mixed models = fixed + random effects

1- With random effect for intercept

--Rather than assuming there is a single intercept for the population, assume that there is a distribution of intercepts. Every person's intercept is a random variable from a shared normal distribution.

--A random intercept for wrist means that there is some average wrist in the population, but there is *variability between subjects*.

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

Generally, this is a "nuisance parameter"—we have to estimate it for making statistical inferences, but we don't care so much about the actual value.

Compare to OLS regression:

Compare with ordinary least squares regression (no random effects):

$$Y_{it} = \beta_{0(\,fixed\,)} + \beta_{1t(\,fixed\,)} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim N(0, \sigma^2_{y/t})$$

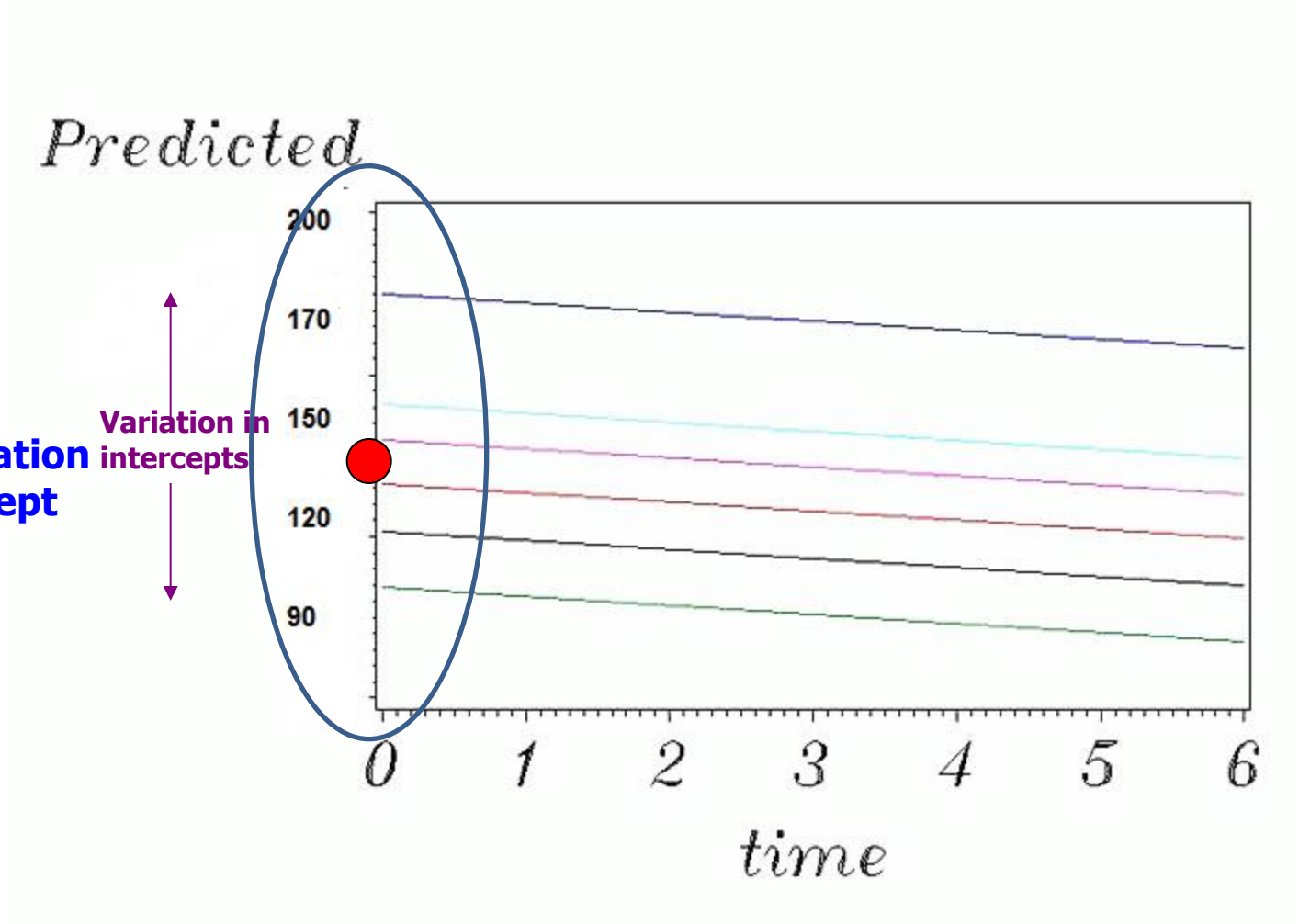
Unexplained variability in Y.

LEAST SQUARES ESTIMATION FINDS THE BETAS THAT MINIMIZE THIS VARIANCE (ERROR)

$$\beta_0 = \text{constant}$$

$$\beta_{time} = \text{constant}$$

Meaning of random intercept



2- With random effect for time, but fixed intercept…

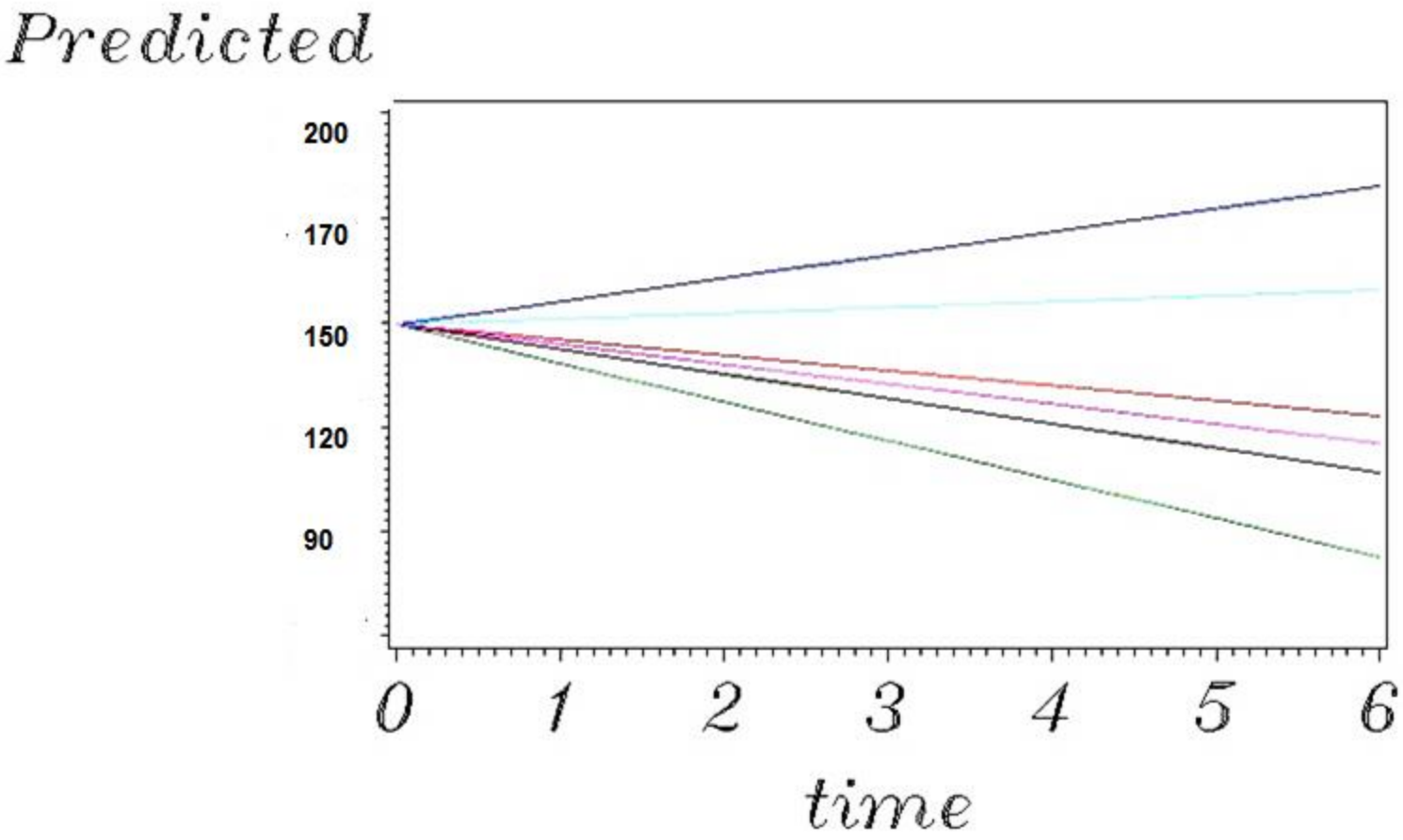
$$Y_{it} = \beta_{0(\,fixed\,)} + \beta_{i,time(random)} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim N(0, \sigma^2_{y/t})$$

$$\beta_{i,time} \sim N(\beta_{time,population}, \sigma^2_{\beta_t})$$

$$\beta_0 = \text{constant}$$

Meaning of random beta for time



3- With both random…

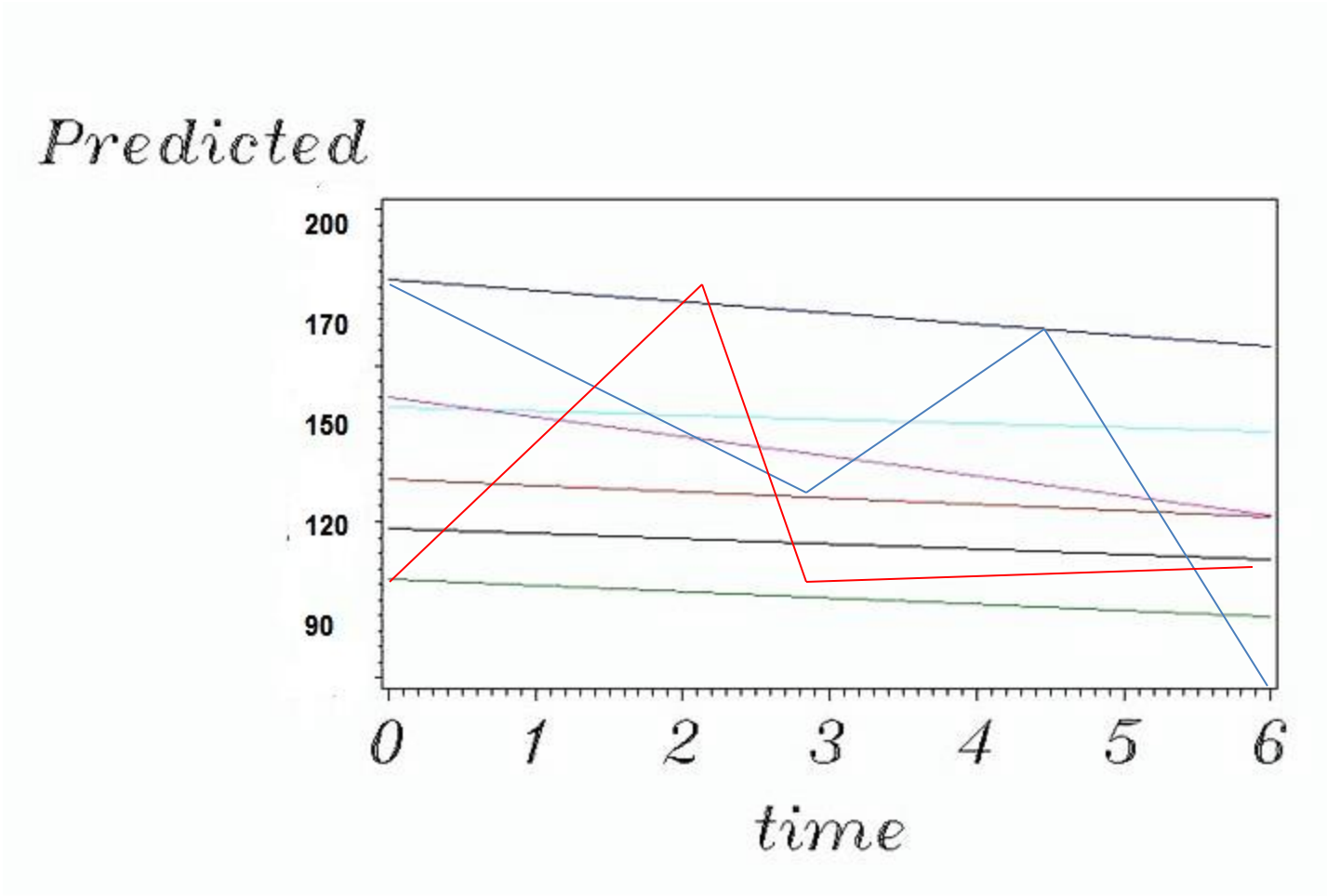
With a random intercept and random time-slope:

$$Y_{it} = \beta_{0i(random)} + \beta_{i,time(random)} + \varepsilon_{it}$$

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

$$\beta_{i,time} \sim N(\beta_{time,population}, \sigma^2_{\beta_t})$$

Meaning of random beta for time and random intercept



Choosing the best model

Aikake Information Criterion (AIC) : a fit statistic penalized by the number of parameters

$$AIC = -2*\log \text{likelihood} + 2*(\#\text{parameters})$$

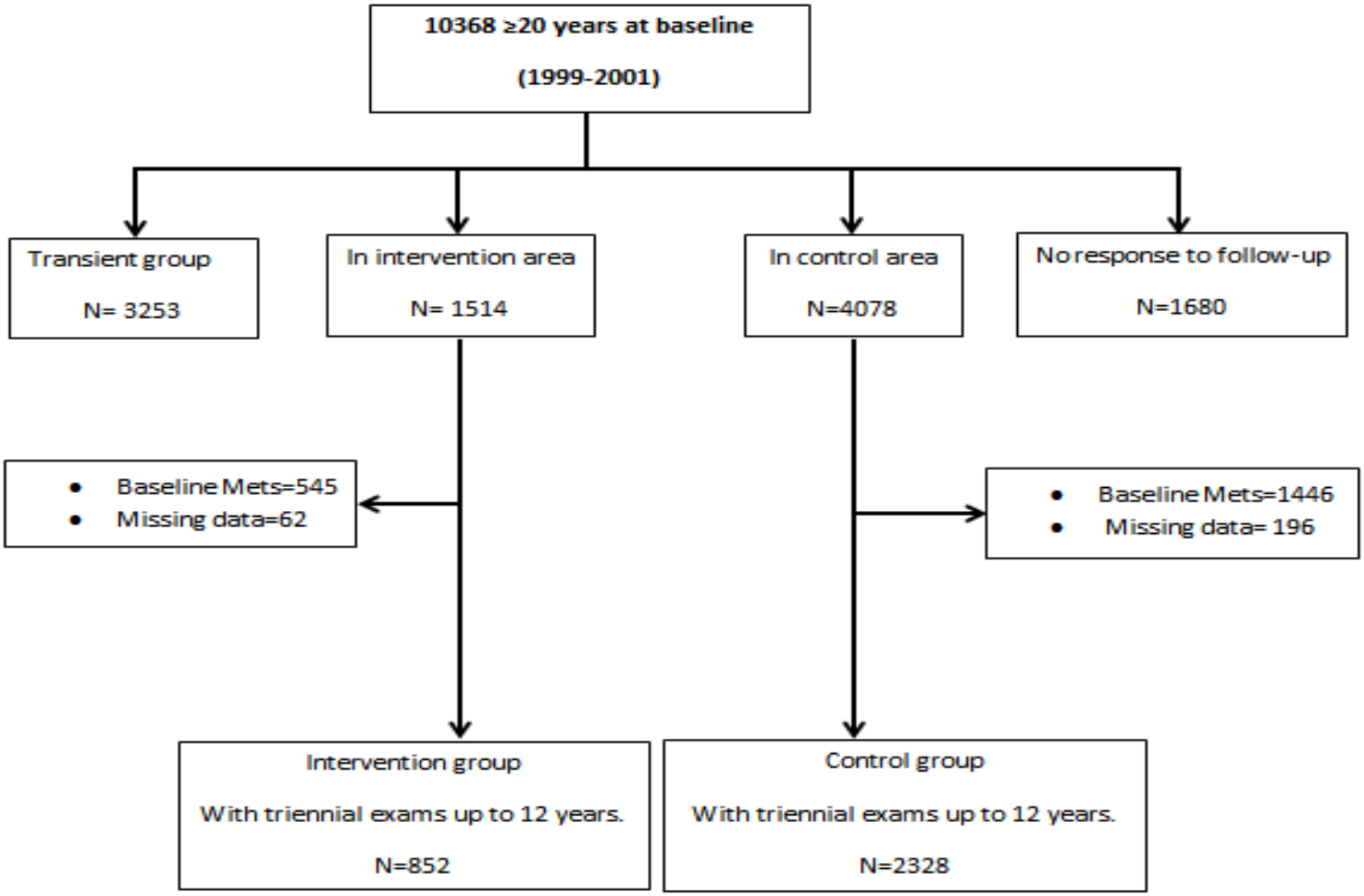
→ Values closer to zero indicate better fit and greater parsimony.

→Choose the model with the smallest AIC.

AICs for the four models

MODEL	AIC
All fixed	162.2
Intercept random Time slope fixed	150.7
Intercept fixed Time effect random	161.4
All random	152.7

Example: The effect of Intervention on METS over 5 phases of TLGS



Total study population: 3180

Table S1 : Effect of intervention on metabolic syndrome at each examination cycle in participants without metabolic syndrome at baseline Based on GEE method

	Exam 2	Exam 3	Exam 4	Exam 5	p-trend*
Model 1					
Intervention effect	0.76(0.61-0.96)	0.76(0.61-0.95)	0.99(0.82-1.21)	1.04(0.85-1.27)	0.02
p-value	0.019	0.015	0.96	0.67	
Model 2					
Intervention effect	0.76(0.61-0.95)	0.76(0.6-0.94)	0.99(0.81-1.21)	1.038(0.85-1.26)	0.02
p-value	0.017	0.013	0.92	0.7	
Model 3					
Intervention effect	0.78(0.62-0.99)	0.74(0.59-0.64)	1.01(0.82-1.24)	1.06(0.86-1.30)	0.01
p-value	0.039	0.012	0.93	0.56	

Model 1: intervention + time (exam 2 to exam 5) + intervention*time + gender + age at each phase

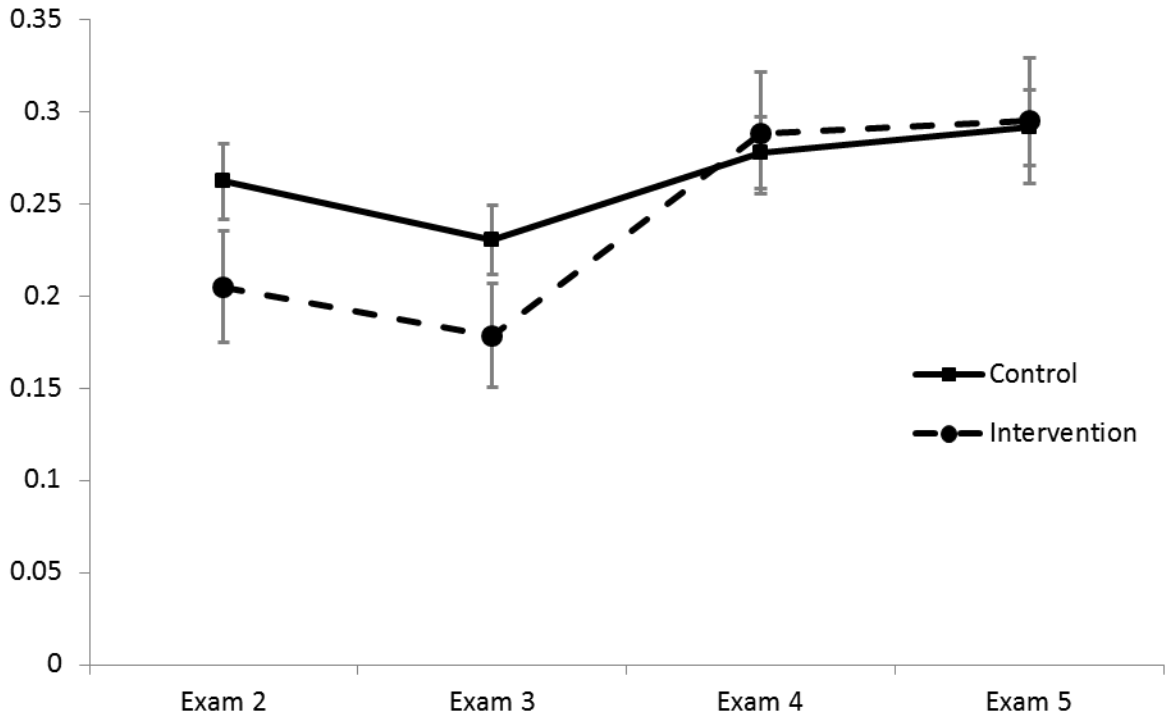
Model 2: model 1+ education, medication, smoking and physical activity at baseline

Model 3: model 2 with sampling weighting for propensity score for response to follow-up

ł Propensity score was calculated as the probability of being followed-up based on baseline covariates including BMI, WC, SBP,DBP, FPG, 2h-PCG,TC,TG,HDL-C, current smoking, low physical activity, hypertension drugs, lipid lowering drugs and diabetes drugs in intervention and control group.

* p for trend was estimated time as continues variable

Body mass index, BMI; Waist circumference, WC; Fasting plasma glucose, FPG; 2-h post challenge plasma glucose, 2h-PCG, Total Cholesterol, TC; Triglycerides, TG; High density lipoprotein, HDL-C.



Age and gender adjusted prevalence of metabolic syndrome at each follow-up examination (every 3 years), among participants without metabolic syndrome at exam 1.

Table 3 : Effect of intervention on metabolic syndrome at each examination cycle in participants without metabolic syndrome at baseline Based on random effect method

	Exam 2	Exam 3	Exam 4	Exam 5	p-trend*
Model 1					
Intervention effect	0.61(0.43-0.87)	0.63(0.44-0.89)	1.08(0.78-1.48)	1.02(0.74-1.41)	0.001
p-value	0.006	0.01	0.65	0.9	
Model 2					
Intervention effect	0.61(0.43-0.86)	0.62(0.44-0.88)	1.06(0.77-1.46)	1.0 (0.73-1.39)	0.001
p-value	0.005	0.008	0.71	0.97	
Model 3					
Intervention effect	0.63(0.48-0.83)	0.6(0.46-0.79)	1.09(0.85-1.4)	1.03(0.79-1.33)	<0.001
p-value	0.001	<0.001	0.5	0.82	

Model 1: intervention + time (exam 2 to exam 5) + intervention*time + gender + age at each phase

Model 2: model 1+ education, medication, smoking and physical activity at baseline

Model 3: model 2 with sampling weighting for propensity score for response to follow-up

ł Propensity score was calculated as the probability of being followed-up based on baseline covariates including BMI, WC, SBP,DBP, FPG, 2h-PCG,TC,TG,HDL-C, current smoking, low physical activity, hypertension drugs, lipid lowering drugs and diabetes drugs in intervention and control group.

* p for trend was estimated time as continues variable

Body mass index, BMI; Waist circumference, WC; Fasting plasma glucose, FPG; 2-h post challenge plasma glucose, 2h-PCG, Total Cholesterol, TC; Triglycerides, TG; High density lipoprotein, HDL-C.

Marginal models or Mixed model?

The choice of model for a particular application would depend on the relevant questions being addressed, which in turn informs the type of design and data collection that would be relevant.

Thanks for your attention