Three Study Designs

1. Cross-sectional study:
   - Information on disease status (Y) and exposure status (X) is obtained from a random sample at one time point.
   - A single observation of each variable of interest is measured from each subject: (Y_i, X_i), $(i = 1, ..., n)$.
   - Regression such as linear regression $Y_i = \beta^T X_i + \varepsilon_i$ (or logistic regression $\log\left(\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}\right) = \beta^T X_i$ if Y_i is binary) can be used to assess the association between Y and X.
   - Usually no temporal information to distinguish cause from effect.

Three Study Designs

2. Prospective cohort study:
   - A cohort with known exposure status (X) is followed over time to obtain their disease status (Y).
   - A single observation of (Y) may be observed (e.g., survival study) or multiple observations of (Y) may be observed (longitudinal study).
   - Stronger evidence for causal inference. Causal inference can be made if X is assigned randomly (if X is a treatment indicator in the case of clinical trials).

Three Study Designs

3. Retrospective (case-control) study:
   - A sample with known disease status (D) is drawn and their exposure history (E) is ascertained.
   - Assuming no bias in obtaining exposure history information on E, association between E and D can be estimated.

Longitudinal studies

- *Longitudinal study*: A longitudinal study is a prospective cohort study where repeated measures are taken over time for each individual.

Goal of longitudinal study: to characterize the change in response over time and the factors influence change. To be specific, a longitudinal study is usually designed to answer the following questions:

- How does the outcome variable of interest change over time?
- How is the (change of) variable of interest associated with treatment and other baseline covariates?
- How do the (time-varying) variables of interest relate to each other over time?

Remark 1: Longitudinal data from longitudinal studies are *clustered*. The clusters are composed of the repeated measurements obtained from a single individual at different occasions.

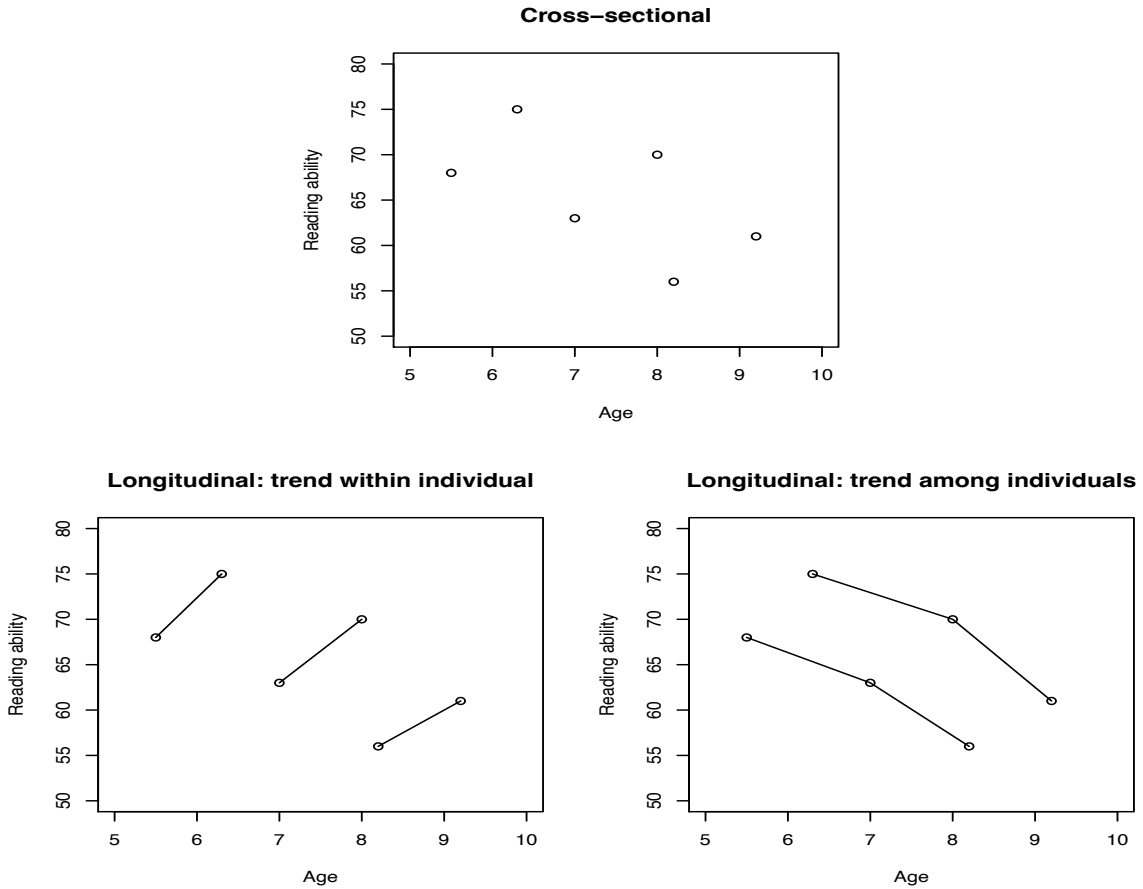
Remark 2: Longitudinal data have a temporal order and correlation within each cluster. While clustered and cluster-correlated data are not necessary longitudinal data.

Longitudinal *vs* Cross-sectional study

1. The cross-sectional study only allow comparisons among sub-populations that happen to differ at one time point. Often, all observations may be treated as independent.

2. The major advantage of longitudinal study is its capacity to capture and assess
   - within-individual changes in the outcome variable over time (longitudinal changes)
   - differences among individuals in their outcome values (cross-sectional changes, cohort effect)

Toy Example: relationship between reading ability and age



- Longitudinal studies can distinguish changes over times within individuals (aging effects) from differences among subjects in their baseline levels (cohort effects).

Remark 3: Longitudinal data analysis requires special statistical methods because the observations on any one subject tend to be positively correlated. Analyzing the repeated measurements just as though they were independent measurements falsely inflates the sample size and results in a failure to preserve type I error and confidence interval coverage.

Example of Longitudinal Study: Framingham Heart Study

- In the original Framingham study, each participant was examined every 2 years for a 10 year period for his/her cholesterol level.
- Study objectives:
  1. How does cholesterol level change over time on average as subjects get older?
  2. How is the change of cholesterol level associated with sex and baseline age?
  3. Do males have more stable baseline cholesterol level and change rate than females?
  4. Various ancillary studies augment the original study (e.g., testing genetic effect on the change of BMI)

Table: A glimpse of the data.

Newid	ID	Cholst	Sex	Baseline Age	Year
1	1244	175	1	32	0
1	1244	198	1	32	2
1	1244	205	1	32	4
1	1244	228	1	32	6
1	1244	214	1	32	8
1	1244	214	1	32	10
2	835	299	0	34	0
2	835	328	0	34	4
2	835	374	0	34	6
2	835	362	0	34	8
2	835	370	0	34	10
3	176	250	0	41	0
3	176	277	0	41	2
3	176	265	0	41	4
3	176	254	0	41	6
3	176	263	0	41	8
3	176	268	0	41	10

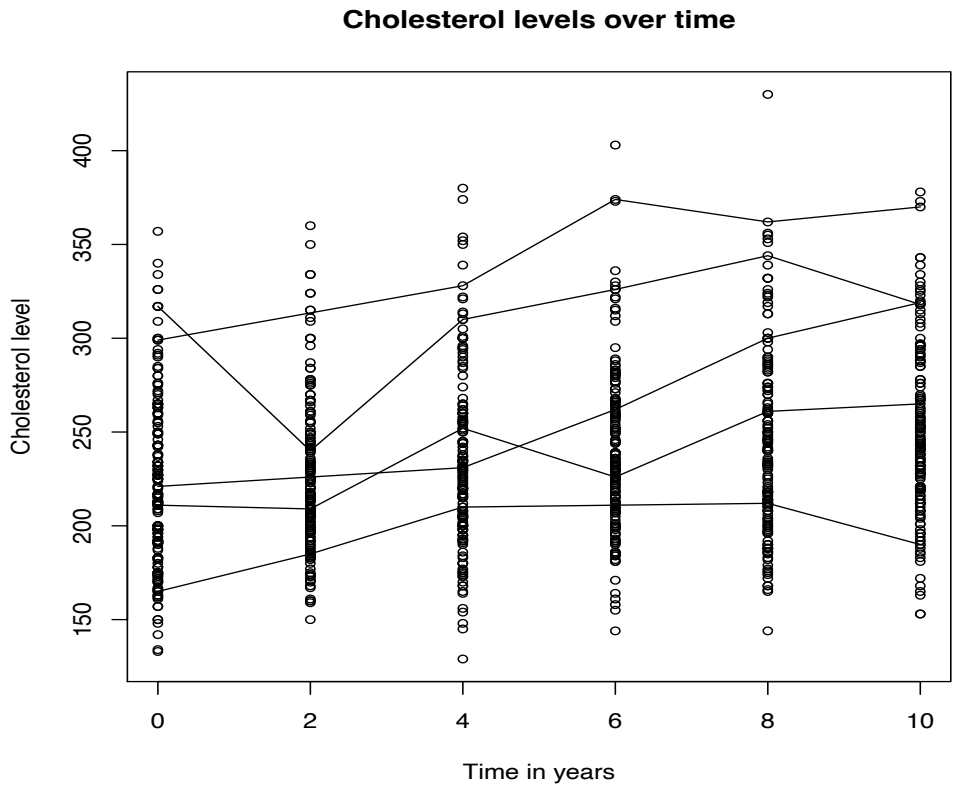


Figure: Cholesterol level over time for a subset of 200 subjects from Framingham study.

Framingham Example

- Cholesterol levels increase (linearly) over time for most individuals.
- Each subject has his/her own trajectory line with a possibly different intercept and slope, implying two sources of variations: within and between subject variations.
- Each subject has on average 5 observations (as opposed to one observation per subject for a cross-sectional study)
- The data is not balanced. Some individuals have missing observations.
- The inference is NOT limited to these 200 individuals. Instead, the inference is for the target population and each subject is viewed as a random person drawn from the target population.

Sources of Correlation in Longitudinal Data

- Between-individual heterogeneity
- Within-individual biological variation (e.g. blood pressure, self-reported pain)
- Random measurement error

Sources of variation and correlation in longitudinal data:

1. Between-subject variation: For the blood pressure example, if each subject's blood pressures were measured within a relatively short time, then the following model may be a reasonable one:

$$y_{ij} = b_i + e_{ij},$$

where b_i is the true blood pressure of subject i, e_{ij} is the independent (random) measurement error, independent of b_i.

For $j \neq k$,

$$\begin{aligned}\text{corr}(y_{ij}, y_{ik}) =& \frac{\text{cov}(y_{ij}, y_{ik})}{\sqrt{\text{var}(y_{ij})\text{var}(y_{ik})}} \\ =& \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}\end{aligned}$$

Therefore, if the between-subject variation $\sigma_b^2 \neq 0$, then data from the same subjects are correlated.

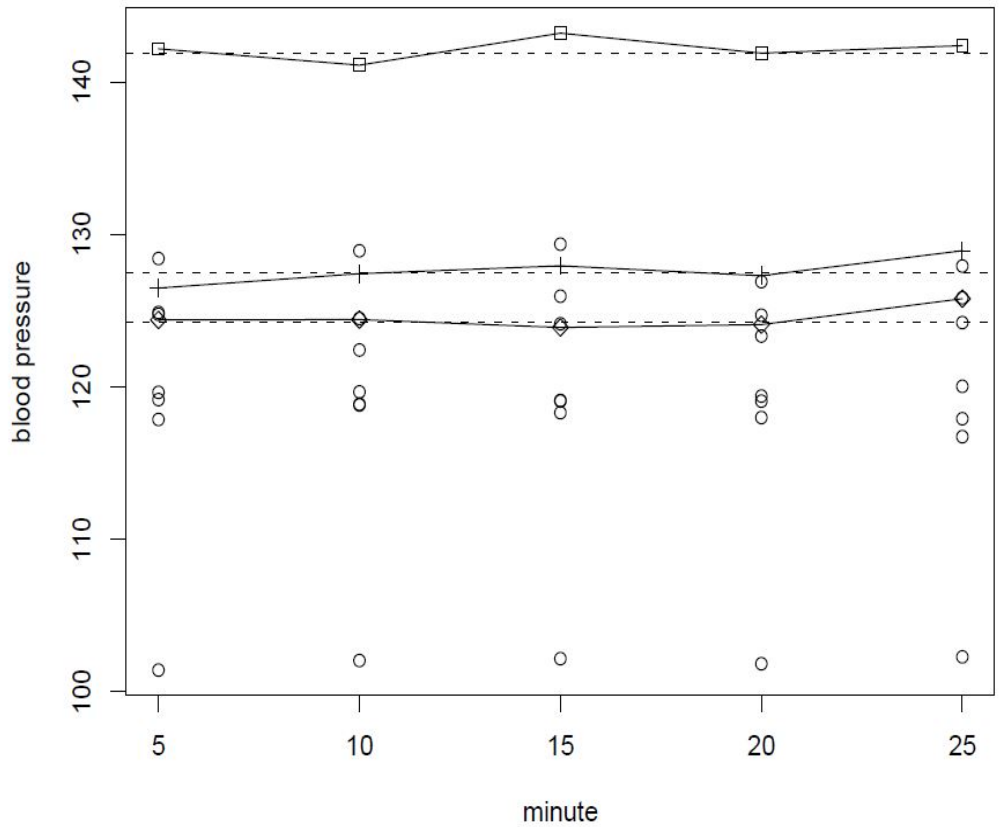


Figure: The blood pressure example.

2. Serial correlation: If the time intervals between blood pressure measurements are relatively large so it may not be reasonable to assume a constant blood pressure for each subject:

$$y_{ij} = b_i + U_i(t_{ij}) + \epsilon_{ij},$$

where b_i = true long-term blood pressure, $U_i(t_{ij})$ = a stochastic process (e.g., a time series) due to biological fluctuation of blood pressure, ϵ_{ij} is the independent (random) measurement error. Here the correlation is caused by both b_i and $U_i(t_{ij})$.

3. In a typical longitudinal study for human subjects where the number of observations per subject is small $\sim$ moderate, there may not be enough information for the serial correlation and most correlation can be accounted for by (possibly complicated) between-subject variation.

Methods for analyzing longitudinal data

1. Two-stage: summarize each subject's outcome and regress the summary statistics on one-time covariates. Useful as an exploratory tool for continuous longitudinal data. This method is now replaced by the mixed effects model approach.
2. Mixed effects model approach: model fixed effects and random effects; use random effects to model correlation.
3. Generalized estimating equation (GEE) approach: model the dependence of marginal mean on covariates. Correlation is not a main interest. Sometimes considered for discrete data.
4. Transition models: use history as covariates. Good for prediction of future response using history.

Two-stage method for analyzing longitudinal data

- Outcome (continuous): $y_{i1}, ..., y_{in_i}$ measured at $t_{i1}, ..., t_{in_i}$, one-time covariates: $x_{i1}, ..., x_{ip}$.
- Two-stage analysis:
  1. Stage 1: Get summary statistics from subject i's data: $y_{i1}, ..., y_{in_i}$. For example, use mean $\bar{y}_i = (y_{i1} + \cdots + y_{in_i})/n_i$ or fit a linear regression for each subject:

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + \epsilon_{ij},$$

and get estimates $\hat{b}_{i0}$, $\hat{b}_{i1}$. Here we assume that subject i's true response at time t_{ij} is given by $b_{i0} + b_{i1} t_{ij}$, a straight line. Suppose $t = 0$ is the baseline, then b_{i0} is subject i's true response at baseline and b_{i1} is subject i's change rate of the true response (not y). The error term ϵ_{ij} can be regarded as measurement error.

2. Stage 2: Treat the summary statistics as new responses and regress the summary statistics on one-time covariates. For example, after we got $\hat{b}_{i0}$ and $\hat{b}_{i1}$, we can calculate the means of $\hat{b}_{i0}$ and $\hat{b}_{i1}$ and the standard errors of those means, or do the following regressions

$$\begin{aligned}\hat{b}_{i0} =& \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_p x_{ip} + e_{i0},\\ \hat{b}_{i1} =& \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_{i1}.\end{aligned}$$

Here, α_k is the effect of x_k on the true baseline response (not y), β_k is the effect of x_k on the change rate of of the true response.

Some remarks on two-stage analysis:

1. The first stage model should be reasonably good for the second stage analysis to be valid and make sense.
2. Two-stage analysis can only be used when the covariates considered are measured once at the baseline (constant over time).
3. When the covariates considered are time-varying covariates, two-stage analysis is not appropriate. Mixed effects models can be used.
4. Although two-stage approach can be used to make inference on the quantities of interest, it is less efficient compared to the mixed model approach. Therefore, mixed effects model approach should be used whenever possible.

Linear mixed effects model

A linear mixed model is an extension of a linear regression model to model longitudinal (correlated) data. It contains fixed effects and random effects where random effects are subject-specific and used to model between-subject variation and the correlation induced by this variation.

What are fixed effects? Fixed effects are the covariate effects that are fixed across subjects in the study sample. These effects are the ones of our particular interest. E.g., the regression coefficients in usual regression models are fixed effects:

$$y = \alpha + x\beta + \epsilon.$$

What are random effects? Random effects are the covariate effects that vary among subjects. So these effects are subject-specific and hence are random (unobservable) since each subject is considered as randomly drawn from a population.

Random intercept and random slope

Table: Data from m subjects

Subject	Outcome	Time	Random intercept	Random slope
1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$t_{11}, t_{12}, \ldots, t_{1n_1}$	b_{10}	b_{11}
2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$t_{21}, t_{22}, \ldots, t_{2n_2}$	b_{20}	b_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$y_{i1}, y_{i2}, \ldots, y_{in_i}$	$t_{i1}, t_{i2}, \ldots, t_{in_i}$	b_{i0}	b_{i1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	$y_{m1}, y_{m2}, \ldots, y_{mn_m}$	$t_{m_1}, t_{m_2}, \ldots, t_{mn_m}$	b_{m0}	b_{m1}

Other covariates: $x_{ij2}, \ldots, x_{ijp}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$.

A random intercept and slope model assumes:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij}.$$

Random intercept and random slope

Random intercept and slope model:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij},$$

β_k is the same as before, random effects b_{i0}, b_{i1} are assumed to have a bivariate normal distribution

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix} \right)$$

Here, $\sigma_{01} = \sigma_{10}$. A general model does not impose constraint on σ_{ij} except covariance matrix being positive definite.

Random intercept and random slope

Interpretation of the model components:

1. Mean structure is the same as before:

$$E[y_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp}.$$

2. β_k: Average increase in y associated with one unit increase in x_k, the kth covariate.
3. $\beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij}$ = true response for subject i at t_{ij}.
4. $\beta_0 + b_{i0}$ = the intercept for subject $i \Longrightarrow b_{i0}$ = deviation of intercept of subject i from population intercept β_0.

5. $\beta_1 + b_{i1}$ = the slope for subject $i \Longrightarrow b_{i1}$ = deviation of slope of subject i from population slope β_1.
6. $\text{Var}(b_{i0} + b_{i1}t_{ij}) = \sigma_{00} + 2t_{ij}\sigma_{01} + t_{ij}^2\sigma_{11}$ = between-subject variance (varying over time).
7. σ_ϵ^2 = within-subject variance.
8. Total variance of y: $\text{Var}(y_{ij}) = \sigma_{00} + 2t_{ij}\sigma_{01} + t_{ij}^2\sigma_{11} + \sigma_\epsilon^2$, not a constant over time.
9. Correlation between y_{ij} and $y_{ij'}$: not a constant over time.

When no fixed effects x, model reduces to random effects model

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1}t_{ij} + \epsilon_{ij}.$$

Other longitudinal data models

- A correlated error model

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + \epsilon_{ij},$$

where ϵ_{ij} are correlated normal errors. For example,

1. Compound symmetric (exchangeable) variance matrix

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \right).$$

Here, $-1 < \rho < 1$. A random intercept model is equivalent to this model.

2. AR(1) variance matrix

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \right).$$

Here, $-1 < \rho < 1$. It assumes that the error $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})$ is an autoregressive process with order 1.

Remark: This structure is more appropriate if y is measured at equally spaced time points.

3. Spatial power variance matrix

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho^{|t_2-t_1|} & \rho^{|t_3-t_1|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_3-t_2|} \\ \rho^{|t_3-t_1|} & \rho^{|t_3-t_2|} & 1 \end{bmatrix} \right).$$

Here, $0 < \rho < 1$. This error structure reduces to AR(1) when y is measured at equally spaced time points.

Remark: This structure is appropriate if y is measured at unequally spaced time points.

4. Unstructured variance matrix

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right).$$

where $\sigma_{ij} = \sigma_{ji}$, $i, j = 1, ...3$.

Remark: This structure may be used only if (potential) time points are the same for all subjects and the number is relatively small.

General linear mixed models

General model 1:

$Y =$ fixed effects + random effects + pure measurement error.

For example,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij},$$

where ϵ_{ij} is the pure measurement error (has an independent variance structure).

Software to implement the above model:

Proc Mixed in SAS:

```
Proc Mixed data= method=;
  class id;
  model y = t x / s; /* specify t x for fixed effects */
  random intercept t / subject=id type=un; /* specify the covariance */
                                           /* for random effects */
  repeated / subject=id type=vc; /* specify the variance structure for error */
run;
```

General model 2:

$$Y = \text{ fixed effects + random effects + stochastic process,}$$

For example,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + b_{i0} + b_{i1} t_{ij} + U_i(t_{ij}),$$

where $U_i(t)$ is a stochastic process with AR(1), a spatial power variance structure or other variance structure.

Software to implement the above model: Proc Mixed in SAS:

```
Proc Mixed data= method=;
  class id;
  model y = t x / s; /* specify t x for fixed effects */
  random intercept t / subject=id type=un; /* specify the covariance */
                                           /* for random effects */
  repeated / subject=id type=sp(pow)(t); /* specify the variance structure for error */
run;
```

If the time points are equally spaced, we can use type=ar(1) in the repeated statement for AR(1) variance structure for $U_i(t)$.

General model 3:

Y = fixed effects + random effects + stochastic process + pure measurement error,

For example,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + b_{i0} + b_{i1} t_{ij} + U_i(t_{ij}) + \epsilon_{ij},$$

where $U_i(t)$ is a stochastic process with some variance structure (e.g., a spatial power variance structure), ϵ_{ij} is the pure measurement error.

Software to implement the above model: Proc Mixed in SAS:

```
Proc Mixed data= method=;
  class id;
  model y = t x / s; /* specify t x for fixed effects */
  random intercept t / subject=id type=un; /* specify the covariance */
                                           /* for random effects */
  repeated / subject=id type=sp(pow)(t) local; /* specify error variance structure */
run;
```

If the time points are equally spaced, we can use type=ar(1) in the repeated statement if assuming AR(1) for $U_i(t)$.

Estimation and Inference

Estimation and inference for linear mixed models

Let θ consist of all parameters in random effects and errors (ϵ_{ij}). We want to make inference on β and θ. There are two approaches:

1. Maximum likelihood:

$$l(\beta, \theta; y) = \log L(\beta, \theta; y).$$

   Maximize $l(\beta, \theta; y)$ jointly w.r.t. β and θ to get their MLEs.

2. Restricted maximum likelihood (REML):
   (a) Get REML of θ from a REML likelihood $l_{REML}(\theta; y)$ (take into account estimation of β). Leads to less biased $\hat{\beta}$. For example, in a linear regression model

$$\hat{\sigma}^2_{REML} = \frac{\text{Residual Sum of Squares}}{n - p - 1}.$$

   (b) Estimate β by maximizing $l(\beta, \widehat{\theta}_{REML}; y)$.

Hypothesis testing

- After we fit a linear mixed model such as

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij},$$

  SAS will output a test for each β_k, including the estimate, SE, p-value (for testing $H_0 : \beta_k = 0$), etc.
- If we want to test a contrast between β_k, we can use 'Estimate' statement in Proc Mixed. Then SAS will output the estimate, SE for the contrast and the p-value for testing the contrast is zero.

How to choose random effects and the error structure?

1. Use graphical representation to identify possible random effects.
2. Use biological knowledge to identify possible error structure.
3. Use information criteria to choose a final model:
   (a) Akaike's Information Criterion (AIC):

$$AIC = -2\{l(\hat{\beta}, \hat{\theta}; y) - q\}$$

   where $q = \#$ of elements in θ. Smaller AIC is preferred.
   (b) Bayesian Information Criterion (BIC):

$$BIC = -2\left\{l(\hat{\beta}, \hat{\theta}; y) - \frac{q \log(n)}{2}\right\},$$

   $n = \#$ of subjects. Again, smaller BIC is preferred.

Analyze Framingham Heart Study data using linear mixed models

Model to address objective 1:

How does cholesterol level change over time on average as subjects get older?

- Consider the following basic model suggested by the data:

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + \epsilon_{ij}, \tag{1}$$

where y_{ij} is the jth cholesterol level measurement from subject i, t_{ij} is year from the beginning of the study (or baseline) and b_{i0}, b_{i1} are random variables distributed as

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix} \right)$$

and ϵ_{ij} are independent errors distributed as $N(0, \sigma_\epsilon^2)$.

- Model (1) assumes that
  1. The true cholesterol level for each individual changes linearly over time with different intercept and slope, which are both random (since the individual is a random subject drawn from the population).
  2. Since $t = 0$ is the baseline, so b_{i0} can be viewed as the true but unobserved cholesterol level for subject i at the baseline, and b_{i1} can be viewed as the change rate of the true cholesterol level for subject i.

3. β_0 is the population average of the true baseline cholesterol level of all individuals in the population, β_1 is the population average change rate of true cholesterol level and it tells us how cholesterol level changes on average as people get older. So β_1 is the longitudinal effect or aging effect on cholesterol level.
4. σ_{00} is the variance of the true baseline cholesterol level b_{i0}; σ_{11} is the variance of the change rate b_{i1} of the true cholesterol level; and σ_{01} is the covariance between true baseline cholesterol level b_{i0} and the change rate b_{i1} of true cholesterol level.

- The random variables b_{i0} and b_{i1} can be re-written as

$$b_{i0} = \beta_0 + a_{i0}, \;\; b_{i1} = \beta_1 + a_{i1},$$

where a_{i0}, a_{i1} have the following distribution:

$$\begin{pmatrix} a_{i0} \\ a_{i1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix} \right)$$

- Model (2.1) then can be re-expressed as

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + a_{i0} + a_{i1} t_{ij} + \epsilon_{ij}. \tag{2}$$

Therefore, β_0, β_1 are fixed effects and a_{i0}, a_{i1} are random effects.

Matrix Form of Linear Mixed Effects Model

A general linear mixed model is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix}\right)$$

G matrix in SAS specifies structure for random effects $\boldsymbol{\gamma}$ (between subject effects), R matrix in SAS specifies covariance matrix for residuals $\boldsymbol{\varepsilon}$ (within subject effects).

The following is the SAS program for fitting model (1):

```
title "Framingham data: mixed model without covariates";
proc mixed data=cholst;
  class newid;
  model cholst = year / s;
  random intercept year / type=un subject=newid g;
  repeated / type=vc subject=newid;
run;
```

Part of outputs for fixed effects

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	220.57	2.9305	199	75.26	<.0001
year	2.8170	0.2408	191	11.70	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
year	1	191	136.83	<.0001

Part of outputs for random effects

The Mixed Procedure

Estimated G Matrix

Row	Effect	newid	Col1	Col2
1	Intercept	1	1467.30	-2.2259
2	year	1	-2.2259	3.8409

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	newid	1467.30
UN(2,1)	newid	-2.2259
UN(2,2)	newid	3.8409
Residual	newid	434.11

Fit Statistics

-2 Res Log Likelihood	9960.1
AIC (smaller is better)	9968.1
AICC (smaller is better)	9968.2
BIC (smaller is better)	9981.3

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	939.63	<.0001

From output, we see that:

1. $\widehat{\sigma}_{00} = 1467$, as compared to $\widehat{\text{var}}(\widehat{b}_0) = 1738$ from the two-stage approach.
2. $\widehat{\sigma}_{11} = 3.84$, as compared to $\widehat{\text{var}}(\widehat{b}_1) = 13.2$ from the two-stage approach.
3. $\widehat{\text{corr}}(b_0, b_1) = -2.2259/\sqrt{1467 \times 3.84} = -0.03$, as compared to $\widehat{\text{corr}}(\widehat{b}_0, \widehat{b}_1) = -0.27$.
4. The estimated mean of true baseline cholesterol level is $\widehat{\beta}_0 = 220.57$ with SE = 2.93, as compared to the sample mean 220.69 of $\widehat{b}_0$ with SE = 2.94 from the two-stage approach.
5. The estimated change rate (longitudinal effect) $\widehat{\beta}_1 = 2.82$ with SE = 0.24, as compared to the sample mean 2.55 of $\widehat{b}_1$ with SE = 0.26 from the two-stage approach.
6. $\widehat{\sigma}^2_\epsilon = 434.11$.

- Model to investigate the cross-sectional age effect and longitudinal age effect on cholesterol level:
  - Re-write the true baseline cholesterol level b_{i0} and the change rate b_{i1} in terms of conditional distributions given age:

$$b_{i0} = \beta_0 + \beta_C \text{age}_i + a_{i0} \tag{3}$$

$$b_{i1} = \beta_1 + \beta_A \text{age}_i + a_{i1} \tag{4}$$

    Where age_i is individual i's baseline age. Then β_C is the cross-sectional age effect and $\beta_1 + \beta_A \text{age}_i$ is the longitudinal effect for the population with baseline age equal to age_i.
  - The average longitudinal effect is

$$\beta_1 + \beta_A E(\text{age}),$$

    which can be estimated by

$$\widehat{\beta}_1 + \widehat{\beta}_A \overline{\text{age}},$$

    where $\overline{\text{age}}$ is the sample average age.

- We can center age and use the centered age (denoted by $\text{c\_age}_i = \text{age}_i - \overline{\text{age}}$) in (4), so

$$b_{i1} = \beta_1 + \beta_A \text{c\_age}_i + a_{i1}.$$

  Then β_1 is the average longitudinal effect
- We are interested in testing whether the cross-sectional age effect and longitudinal age effect are the same:

$$H_0 : \beta_C = \beta_1.$$

- Assume the usual distribution for (a_{i0}, a_{i1}):

$$\begin{pmatrix} a_{i0} \\ a_{i1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix} \right).$$

  Here both σ_{00} and σ_{11} are the remaining variances in b_{i0} and b_{i1} after the baseline age effect has been taken into account. So they should be smaller than those corresponding values in model (1).

- Basic model (1) becomes

$$y_{ij} = \beta_0 + \beta_C \text{c\_age}_i + \beta_1 t_{ij} + \beta_A \text{c\_age}_i * t_{ij} + a_{i0} + a_{i1} t_{ij} + \epsilon_{ij}, \quad (5)$$

  where $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent errors.

- The following is the SAS program for fitting model (5):

```
title "Framingham data: longitudinal effect vs. cohort effect";
proc mixed data=cholst;
  class newid;
  model cholst = year cage cage*year / s;
  random intercept year / type=un subject=newid g;
  repeated / type=vc subject=newid;
  estimate "long-cross" year 1 cage -1;
run;
```

- The relevant output of the above SAS program is

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	10826.01576300	
1	2	9929.74817925	0.00000516
2	1	9929.72729664	0.00000000

Convergence criteria met.

Estimated G Matrix

Row	Effect	newid	Col1	Col2
1	Intercept	1	1226.69	9.7829
2	year	1	9.7829	3.2598

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	newid	1226.69
UN(2,1)	newid	9.7829
UN(2,2)	newid	3.2598
Residual	newid	434.15

Fit Statistics

-2 Res Log Likelihood	9929.7
AIC (smaller is better)	9937.7

AICC (smaller is better)	9937.8
BIC (smaller is better)	9950.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	896.29	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	220.57	2.7172	198	81.18	<.0001
year	2.8157	0.2343	190	12.02	<.0001
cage	1.9861	0.3455	652	5.75	<.0001
year*cage	-0.1024	0.02930	652	-3.50	0.0005

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
year	1	190	144.42	<.0001
cage	1	652	33.05	<.0001
year*cage	1	652	12.22	0.0005

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
long-cross	0.8296	0.4174	652	1.99	0.0473

- What we learn from this output:
  1. $\widehat{\sigma}_{00} = 1226.7$, much smaller than the corresponding estimate 1467 from model (1) when baseline age was not used to explain the variability in the true baseline cholesterol level.
  2. $\widehat{\sigma}_{11} = 3.26$, much smaller than the corresponding estimate 3.84 from model (1) when baseline age was not used to explain the variability in the true baseline cholesterol change rate.
  3. $\widehat{\beta}_0 = 220.57$ is the estimate of mean true baseline cholesterol level for the individuals whose baseline age = 42.56 (the average age), which is the same as the one from model (1) but with a smaller SE (2.71 vs. 2.93).
  4. The estimate of the longitudinal age effect is $\widehat{\beta}_1 = 2.8157$ with SE = 0.2343, which is similar to $\widehat{\beta}_1 = 2.8170$ with SE = 0.24 from model (1).

5. The estimate of the cross-sectional age effect is $\widehat{\beta}_C = 1.99$ with SE = 0.3455, which is very different from the estimate of the longitudinal age effect $\widehat{\beta}_1 = 2.82$.
6. The p-value for testing $H_0 : \beta_C = \beta_1$ is 0.0473, significant at level 0.05.
7. $\sigma^2_\epsilon = 434.15$ is basically the same as the corresponding estimate from model (1), which is 434.11.
8. Similarly, we can test $i.i.d$ ϵ_{ij} by considering correlated errors such as AR(1).

- Model to address objective 2:

  How is the change of cholesterol level associated with sex and baseline age?

  - Re-write the true baseline cholesterol level b_{i0} and the change rate b_{i1} in model (1) in terms of conditional distribution given gender and baseline age:

$$b_{i0} = \beta_0 + \beta_{0,\text{sex}}\text{sex}_i + \beta_{0,\text{age}}\text{age}_i + a_{i0} \tag{6}$$

$$b_{i1} = \beta_1 + \beta_{1,\text{sex}}\text{sex}_i + \beta_{1,\text{age}}\text{age}_i + a_{i1} \tag{7}$$

    where we assume that a_{i0}, a_{i1} have the following

$$\begin{pmatrix} a_{i0} \\ a_{i1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix}\right).$$

  - Then $\beta_{0,\text{sex}}$, $\beta_{0,\text{age}}$ are the sex effect and baseline age effect on the baseline cholesterol level.

- Similarly, $\beta_{1,\text{sex}}$, $\beta_{1,\text{age}}$ are the sex effect and baseline age effect on the change rate of the true cholesterol level.
- Substituting the above expressions into model (1) to obtain

$$y_{ij} = \beta_0 + \beta_{0,\text{sex}}\text{sex}_i + \beta_{0,\text{age}}\text{age}_i + \beta_1 t_{ij}$$

$$+\beta_{1,\text{sex}}\text{sex}_i * t_{ij} + \beta_{1,\text{age}}\text{age}_i * t_{ij} + a_{i0} + a_{i1}t_{ij} + \epsilon_{ij}. \qquad (8)$$

- Suppose we also want to test whether or not the change rates between 30 years old males and 40 years old females are the same using above model.
- From (6) and (7), the (average) change rate of 30 years old males is

$$\beta_1 + 1 \times \beta_{1,\text{sex}} + 30 \times \beta_{1,\text{age}} = \beta_1 + \beta_{1,\text{sex}} + 30\beta_{1,\text{age}}.$$

The (average) change rate of 40 years old females is

$$\beta_1 + 0 \times \beta_{1,\text{sex}} + 40 \times \beta_{1,\text{age}} = \beta_1 + 40\beta_{1,\text{age}}.$$

The difference between these two rates is

$$\beta_1 + \beta_{1,\text{sex}} + 30\beta_{1,\text{age}} - (\beta_1 + 40\beta_{1,\text{age}}) = \beta_{1,\text{sex}} - 10\beta_{1,\text{age}}.$$

Therefore, we need only to test H_0: $\beta_{1,\text{sex}} - 10\beta_{1,\text{age}} = 0$.

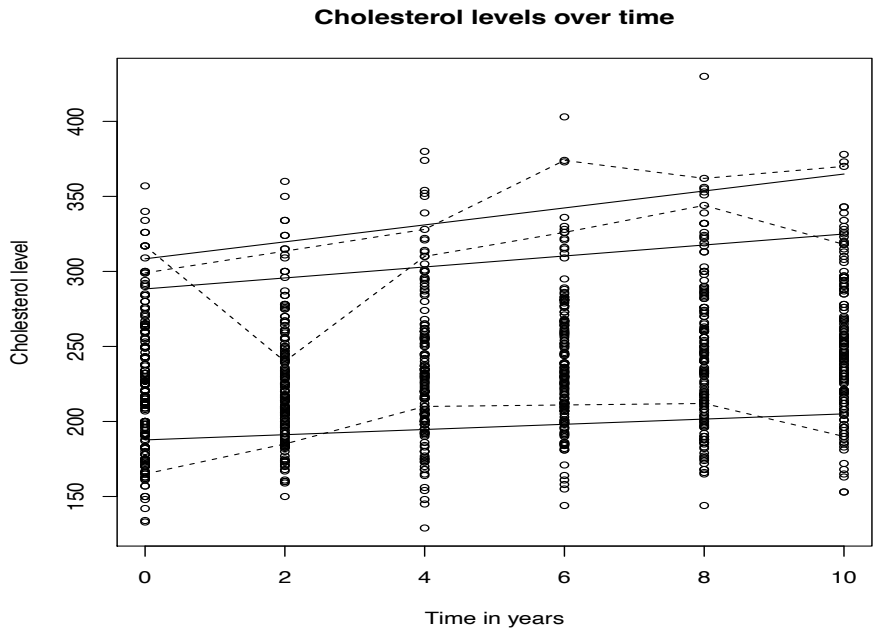
We can use the following SAS program to answer our questions.

```
title "Framingham data: how baseline cholesterol level and";
title2" change rate depend on sex and baseline age";
proc mixed data=cholst;
  class newid;
  model cholst = sex age year sex*year age*year / s;
  random intercept year / type=un subject=newid g s;
  repeated / type=vc subject=newid;
  estimate "rate-diff" sex*year 1 age*year -10;
run;
```

(a) Effect on baseline cholesterol level:
Model (8): $\hat{\beta}_0 = 138.18$ (SE = 14.9),
$\hat{\beta}_{0,sex} = -9.64$ (SE = 5.43), $\hat{\beta}_{0,age} = 2.05$ (SE = 0.35)

(b) Effect on change rate of cholesterol level:
Model (8): $\hat{\beta}_1 = 6.80$ (SE = 1.22),
$\hat{\beta}_{1,sex} = 1.80$ (SE = 0.45), $\hat{\beta}_{1,age} = -0.11$ (SE = 0.03).

We can also estimate the individual random effects and estimate their trajectory lines.



Model to address Objective 3:

Do males have more stable (true) baseline cholesterol level and change rate than females? Hypothesis on variance components.

- From model (1), assume b_{i0}^*, b_{i1}^* have different distributions for males and females:

$$Males: \begin{pmatrix} b_{i0}^* \\ b_{i1}^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{m0} \\ \mu_{m1} \end{pmatrix}, \begin{pmatrix} \sigma_{m00} & \sigma_{m01} \\ \sigma_{m01} & \sigma_{m11} \end{pmatrix}\right)$$

$$Females: \begin{pmatrix} b_{i0}^* \\ b_{i1}^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{f0} \\ \mu_{f1} \end{pmatrix}, \begin{pmatrix} \sigma_{f00} & \sigma_{f01} \\ \sigma_{f01} & \sigma_{f11} \end{pmatrix}\right) \quad (9)$$

- We would like to test

$$H_0 : \sigma_{m00} = \sigma_{f00}, \sigma_{m01} = \sigma_{f01}, \sigma_{m11} = \sigma_{f11}$$

(i.e., the above two variance-covariance matrices are the same).

The SAS program and its output for fitting above model are as follows:

```
data cholst; set cholst;
  gender=sex;
run;

title "Framingham data: do males have more stable (true) baseline";
title2 "cholesterol level and change rate than females?";
proc mixed data=cholst;
   class newid gender;
   model cholst = sex time sex*time / s;
   random intercept time / type=un subject=newid group=gender g;
   repeated / type=vc subject=newid;
run;
```

Framingham data: do males have more stable (true) baseline cholesterol level and change rate than females?

The Mixed Procedure

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	10889.09479529	
1	3	9939.57691271	0.00000317
2	1	9939.56399905	0.00000000

The Mixed Procedure
Convergence criteria met.

Estimated G Matrix

Row	Effect	newid	gender	Col1	Col2	Col3	Col4
1	Intercept	1	0	1402.47	-4.7015		
2	time	1	0	-4.7015	1.8279		
3	Intercept	1	1			1532.81	3.6119
4	time	1	1			3.6119	4.7970

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate
UN(1,1)	newid	gender 0	1402.47
UN(2,1)	newid	gender 0	-4.7015
UN(2,2)	newid	gender 0	1.8279
UN(1,1)	newid	gender 1	1532.81
UN(2,1)	newid	gender 1	3.6119
UN(2,2)	newid	gender 1	4.7970
Residual	newid		433.71

Fit Statistics

-2 Res Log Likelihood	9939.6
AIC (smaller is better)	9953.6
AICC (smaller is better)	9953.7
BIC (smaller is better)	9976.7

In order to test H_0: the two variance matrices are the same using the likelihood ratio test (LRT), we need to fit a model with the same fixed and random effects but under H_0. The following is the SAS program and its output under H_0. Refer the null model as (9').

```
title "Framingham data under H0: males and females have the same variance";
title2 "matrices of baseline cholesterol level and change rate";
proc mixed data=cholst;
  class newid gender;
  model cholst = sex time sex*time / s;
  random intercept time / type=un subject=newid g;
  repeated / type=vc subject=newid;
run;
```

Framingham data under H0: males and females have the same variance matrices of baseline cholesterol level and change rate

The Mixed Procedure

Model Information

Data Set	WORK.CHOLST
Dependent Variable	cholst
Covariance Structures	Unstructured, Variance

The Mixed Procedure
Convergence criteria met.

Estimated G Matrix

Row	Effect	newid	Col1	Col2
1	Intercept	1	1465.85	-0.2516
2	time	1	-0.2516	3.2618

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	newid	1465.85
UN(2,1)	newid	-0.2516
UN(2,2)	newid	3.2618
Residual	newid	434.17

Fit Statistics

-2 Res Log Likelihood	9943.0
AIC (smaller is better)	9951.0
AICC (smaller is better)	9951.1
BIC (smaller is better)	9964.2

The difference of -2*restricted log likelihood is $9943 - 9939.6 = 3.4$ (between models (9) and (9′)) and the p-value = $P[\chi^2_3 \geq 3.4] = 0.33$. There is no sufficient statistical evidence to reject the null hypothesis.

Robust Standard Error Estimates

When the covariance structure is not correctly specified, the inference may be incorrect.

Use the model you posed to estimate the fixed effects (β's) and calculate the robust sandwich variance estimate for the fixed effect estimates. These SE estimates will be valid regardless of the validity of the random effects structure.

- For example, you can use the following model to estimate β's:

$$\begin{aligned} y_{ij} =& \beta_0 + \beta_1 t_{ij} + \beta_2 \text{sex}_i + \beta_3 \text{age}_i + \beta_4 \text{sex}_i t_{ij} + \beta_5 \text{age}_i t_{ij} \\ &+ b_{i0} + b_{i1} t_{ij} + \epsilon_{ij}. \end{aligned}$$

  Specify empirical in Proc mixed, you will get robust SE estimates.

```
proc mixed data=cholst empirical;
  class newid;
  model cholst = time sex age sex*time age*time / s;
  random intercept time / type=un subject=newid;
  repeated / type=vc subject=newid;
run;
```

Design Issues

Design a longitudinal study: Sample size estimation

In the classical setting, sample size estimation is posed as a hypothesis testing problem

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 \neq \mu_2.$$

Assume $y_{1k}, ..., y_{mk} \sim N(\mu_k, \sigma^2)$, $k = 1, 2$. Given significance level α, power γ, and the difference $\Delta = (\mu_1 - \mu_2)/\sigma$ we wish to detect, the required total sample size (number of subjects) in each group should be

$$m = 2\left[\frac{z_{\alpha/2} + z_{1-\gamma}}{\Delta}\right]^2.$$

Design a longitudinal study

I. Compare time-averaged means between two groups.

Assume model for the data to be collected:

$$\text{Group A} : y_{ij} = \mu_A + \epsilon_{ij},\ i = 1, ..., m,\ j = 1, ..., n$$

$$\text{Group B} : y_{ij} = \mu_B + \epsilon_{ij},\ i = 1, ..., m,\ j = 1, ..., n$$

$m =$ number of subjects, $n =$ number of observations per subject, ϵ_{ij} normally distributed errors with mean zero, variance σ^2 and correlation ρ.

We want to test

$$H_0 : \mu_A = \mu_B \quad vs \quad H_A : \mu_A \neq \mu_B$$

at level α with power γ to detect difference $\Delta = (\mu_A - \mu_B)/\sigma$. The quantities m and n have to satisfy

$$m = 2(1 + (n-1)\rho)\frac{(z_{\alpha/2} + z_{1-\gamma})^2}{n\Delta^2}.$$

The sample size formula depends on σ^2 and ρ.

$$m = 2(1 + (n-1)\rho)\frac{(z_{\alpha/2} + z_{1-\gamma})^2}{n\Delta^2}$$

Remarks:

1. When $n = 1$, the study reduces to a cross-sectional study and the sample size formula reduces to the classical one.
2. When $\rho = 0$ (responses are independent), the required sample size is $1/n$ of that for classical study.
3. When $\rho = 1$, required sample size is the same as that of the classical study.
4. For fixed n, smaller ρ gives smaller sample size.
5. If correlation is high, use more subjects and less obs/subject; if correlation is low, use less subjects and more obs/subject.

An example:

If $n = 3$, $\alpha = 0.05$, $\gamma = 0.8$, then the number of subjects (m) per group is

$$m = 2(1 + 2\rho)(1.96 + 0.84)^2/3\Delta^2$$

	Δ							
ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
0.2	733	184	82	46	30	21	15	12
0.3	838	210	94	53	34	24	18	14
0.4	942	236	105	59	38	27	20	15
0.5	1047	262	117	66	42	30	22	17
0.6	1152	288	128	72	47	32	24	18
0.7	1256	314	140	79	51	35	26	20
0.8	1361	341	152	86	55	38	28	22

Design a longitudinal study (cont'd)

II. Compare slopes between two groups.

Model for the data to be collected:

$$\text{Group A}: y_{ij} = \beta_{0A} + \beta_{1A} t_j + \epsilon_{ij},\ i = 1, ..., m,\ j = 1, ..., n$$

$$\text{Group B}: y_{ij} = \beta_{0B} + \beta_{1B} t_j + \epsilon_{ij},\ i = 1, ..., m,\ j = 1, ..., n$$

m = number of subjects, n = number of observations/subject, ϵ_{ij} normally distributed errors with mean zero, variance σ^2 and correlation ρ.

We are interested in testing

$$H_0 : \beta_{1A} = \beta_{1B} \quad vs \quad H_A : \beta_{1A} \neq \beta_{1B}$$

at level α with power γ to detect difference $\Delta = (\beta_{1A} - \beta_{1B})/\sigma$. The quantities m and n have to satisfy

$$m = 2(1-\rho)\frac{(z_{\alpha/2} + z_{1-\gamma})^2}{n\Delta^2 s_t^2},\ s_t^2 = \frac{1}{n}\sum_{j=1}^{n}(t_j - \bar{t})^2.$$

The sample size formula depends on information on σ^2 and ρ and the placement of time points t_j's.

$$m = 2(1-\rho)\frac{(z_{\alpha/2} + z_{1-\gamma})^2}{n\Delta^2 s_t^2},\ s_t^2 = \frac{1}{n}\sum_{j=1}^{n}(t_j - \bar{t})^2$$

Remarks:

1. For fixed time points t_j, larger ρ gives smaller sample size m.
2. If correlation is low, use more subjects and less obs/subject; if correlation is high, use less subjects and more obs/subject.
3. The sample size formula depends on information on σ^2 and ρ and the placement of time points t_j's.
4. For more general covariance structure, replace $\sigma^2(1-\rho)/ns_t^2$ by $X^T R^{-1} X$, where $X = (\mathbf{1}, \mathbf{t})$, $\mathbf{t} = (1, t_1, \cdots, t_n)^T$.

An example:
If $n = 3$, $\alpha = 0.05$, $\gamma = 0.8$, $t = (0, 2, 5)$ so $s_t^2 = 4.222$, then the number of subjects (m) per group is

$$m = 2(1 - \rho)(1.96 + 0.84)^2/(3 \times 4.222\Delta^2)$$

	Δ								
ρ	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.2	2479	1102	620	397	276	203	155	123	100
0.3	2169	964	543	348	241	178	136	108	87
0.4	1859	827	465	298	207	152	117	92	75
0.5	1550	689	388	248	173	127	97	77	62
0.6	1240	551	310	199	138	102	78	62	50
0.7	930	414	233	149	104	76	59	46	38
0.8	620	276	155	100	69	51	39	31	25

Summary

- Longitudinal studies have advantages over cross-sectional studies.
- Challenges in analyzing data from longitudinal studies: correlation, within-subject and between-subject variation.
- Linear mixed models for analyzing continuous longitudinal data: random effects are explicitly used to model the between-subject variation.
- Generalized linear mixed model can be used to analyze discrete longitudinal data where random effects are used to model the correlation. Subject-specific interpretation.