Introduction to mixed model and missing data issues in longitudinal studies

Hélène Jacqmin-Gadda

INSERM, U897, Bordeaux, France

Inserm workshop, St Raphael

Outline of the talk I

Longitudinal data : definition

Definition :

Variables measured at several times on the same subjects

Examples :

- repeated measures of biological markers (CD4, HIV RNA) in HIV patients
- repeated measures of neuropsychological tests to study cognitive aging
- Repeated events : dental caries, absences from school or job, ...

Longitudinal data analysis

Objective :

- Describe change of the variable with time
- Identify factors associated with change

Problem : Intra-subject correlation

Example : HIV clinical trial

X_i=1 if treatment A,

X_i=0 if treatment B

Criterion : Change over time of CD4

Repeated measures of CD4 over the follow-up period.

$t = 0$ at initiation of treatment.

Y_{ij} = CD4 measure for subject i at time t_{ij}, $i = 1, ..., N$, $j = 1, ..., n_i$.

Analysis assuming independence

$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij} + \epsilon_{ij}$
with $\epsilon_{ij} \sim \mathcal{N}(O, \sigma^2)$ and $\epsilon_{ij} \perp \epsilon_{ij'}$

Intra-subject correlation

$\rightarrow$ $\hat{Var}(\hat{\beta})$ biased

$\rightarrow$ Tests for β biased

For time-independent covariate :

- $var(\hat{\beta}_2)$ under-estimated
- Tests for $H_0 : \beta_2 = 0$ anti-conservative (p value too small)

Linear mixed model with random intercept

$$Y_{ij} = (\beta_0 + \gamma_{0i}) + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij} + \epsilon_{ij}$$

with $\gamma_{0i} \sim \mathcal{N}(O, \sigma_0^2)$, and $\epsilon_{ij} \sim \mathcal{N}(O, \sigma^2)$ and $\epsilon_{ij} \perp \epsilon_{ij'}$

- γ_{0i} are random variables
- Only one additional parameter : σ_0^2

Linear mixed model with random intercept (2)

- Population (marginal) mean :

$$E(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij}$$

- Subject-specific (conditional) mean :

$$E(Y_{ij}|\gamma_{0i}) = (\beta_0 + \gamma_{0i}) + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij}$$

- Assume common correlation between all the repeated measures

Linear mixed model with random intercept and slope

$$Y_{ij} = (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i})t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij} + \epsilon_{ij},$$

$$\gamma_{0i} \sim \mathcal{N}(O, \sigma_0^2),\ \gamma_{1i} \sim \mathcal{N}(O, \sigma_1^2),\ \epsilon_{ij} \sim \mathcal{N}(O, \sigma^2),\ \epsilon_{ij} \perp \epsilon_{ij'}$$

- Population (marginal) mean :

$$E(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij}$$

- Subject-specific (conditional) mean :

$$E(Y_{ij}|\gamma_i) = (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i})t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij}$$

- The correlation between repeated measures depend on measurement times

Linear mixed model : general formulation

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T \gamma_i + \epsilon_{ij}$$

$$\gamma_i \sim \mathcal{N}(0, B) \text{ and } \epsilon_i \sim \mathcal{N}(0, R_i).$$

X_{ij} : vector of explanatory variables
β : vector of fixed effects
Z_{ij} : sub-vector of X_{ij} (including functions of time)
γ_i : vector of random effects.

Population (marginal) mean : $E(Y_{ij}) = X_{ij}^T \beta$
Subject-specific (conditional) mean : $E(Y_{ij}|\gamma_i) = X_{ij}^T \beta + Z_{ij}^T \gamma_i$

Linear mixed model : example

Linear mixed model with AR Gaussian error

$$Y_{ij} = (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i})t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij} + w_{ij} + e_{ij}$$

with $\gamma_i^t = (\gamma_{0i}, \gamma_{1i}) \sim \mathcal{N}(\mathbf{0}, B)$,

$e_{ij} \sim \mathcal{N}(O, \sigma^2)$, $e_{ij} \perp e_{ij'}$,

$w_{ij} \sim \mathcal{N}(O, \sigma_w^2)$ and $Corr(w_{ij}, w_{ij'}) = exp(-\delta|t_{ij} - t_{ij'}|)$

Linear mixed model : Estimation

- Maximum likelihood estimator
- $Y_i = (Y_{i1}, ..., Y_{ij}, ..., Y_{in_i})^T$ multivariate Gaussian with
  - mean $X_i\beta$
  - and covariance matrix $V_i = Z_i B Z_i^T + R_i$
- Softwares : SAS Proc mixed, R lme, stata

Generalized linear mixed model

$Y_{ij} \sim$ exponential family of distribution and

$$g(E(Y_{ij}|\gamma_i)) = X_{ij}^T\beta + Z_{ij}^T\gamma_i \text{ with } \gamma_i \sim \mathcal{N}(O, B).$$

- Example : Logistic mixed model

$$logit(Pr(Y_{ij} = 1|\gamma_i)) = X_{ij}^T\beta + Z_{ij}^T\gamma_i \text{ with } \gamma_i \sim \mathcal{N}(0, B).$$

- Maximum likelihood estimation : Numerical integration
- Softwares : SAS Proc nlmixed, R nlme, stata

Typology of missing data in longitudinal studies

Notation :

$Y_i = (Y_{obs,i}, Y_{mis,i})$
with $Y_{obs,i}$ the observed part of Y_i and $Y_{mis,i}$ the missing part,

$R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing
$R_i = (R_{i1}, ..., R_{ij}, ..., R_{in_i})'$

X_i explanatory variables completely observed

Typology of missing data (2)

Monotone missing data = dropout : $P(R_{ij} = 0 | R_{ij-1} = 0) = 1$
R_i may be summarized by the time to dropout T_i
and an indicator for dropout δ_i

Intermittent missing data : $P(R_{ij} = 0 | R_{ij-1} = 0) < 1$

Typology of missing data (3)

Missing Completely at random (MCAR) :

$P(R_{ij} = 1)$ is constant

The observed sample is representative of the whole sample.

$\rightarrow$ Loss of precision, no bias

Covariate-dependent missingness process :

$P(R_{ij} = 1) = f(X_i)$

$\rightarrow$ Loss of precision, no bias if analyses are adjusted on X_i

Typology of missing data (4)

Missing at random (MAR) : $P(R_{ij} = 1) = f(Y_{obs,i}, X_i)$

Example : Probability of dropout depends on past observed values

$\rightarrow$ Loss of precision, no bias with appropriate statistical methods

Informatives or MNAR : $P(R_{ij} = 1) = f(Y_{mis,i}, Y_{obs,i}, X_i)$

Example : Probability that Y be observed depends on current Y value

$\rightarrow$ Loss of precision, biases

$\rightarrow$ Sensitivity analyses

Exploring incomplete data

- Describe missing data frequency
- Cross classify missing data patterns with covariates
- Compare mean evolution for available data and complete cases
- Compare mean evolution until time t given observation status at time $t+1$
- Logistic regression for $P(R_{ij}=1)$ given covariates and $Y_{ik}, k<j$
- Cox regression for time to dropout given covariates

$\rightarrow$ Impossible to distinguish MAR from MNAR

An example : Paquid data set

The Paquid Cohort in Gironde

- 2792 subjects of 65 years and older at baseline
- Living at home at the beginning of the study (1988) in Gironde (France)
- Seen at home at 1, 3, 5, 8, and 10 years after the baseline visit
- Cognitive measure : Digit Symbol Substitution Test of Wechsler (attention, limited time to 90s)

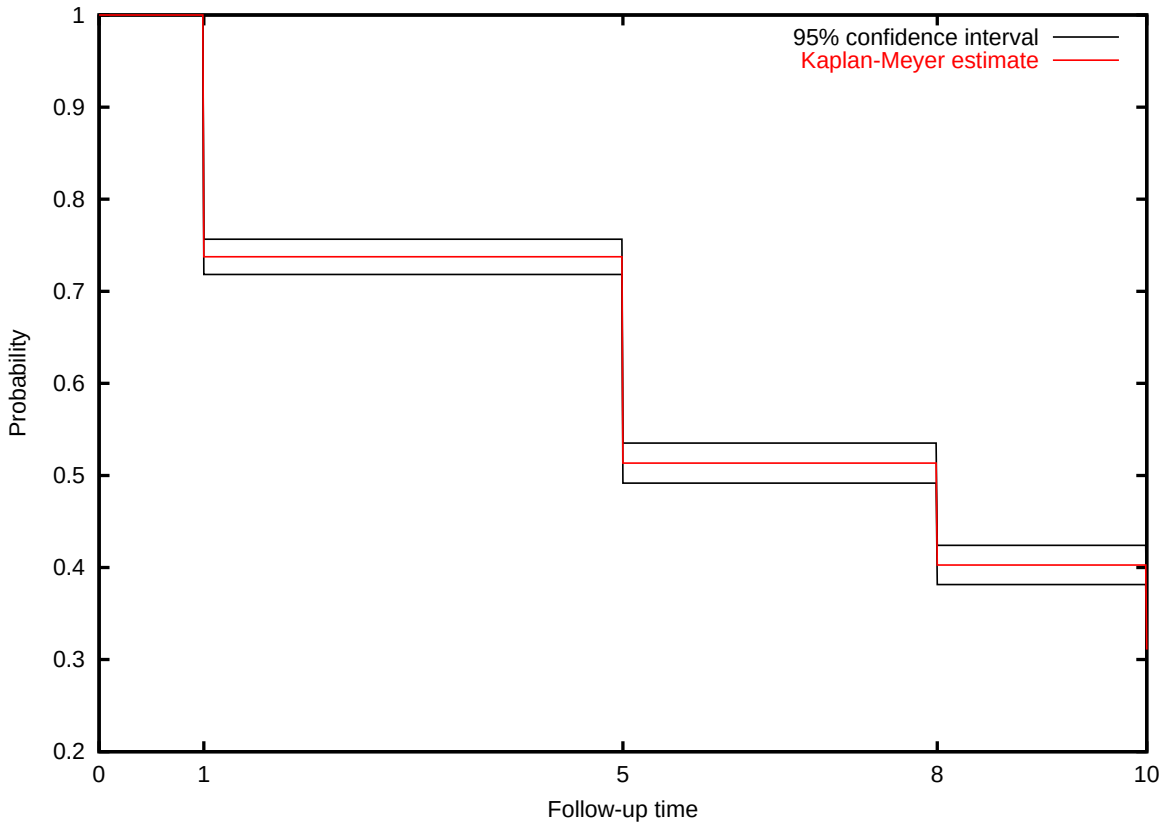
Sample :

- 2026 subjects
- without diagnosis of dementia between T0 and T10
- with the test completed at least once (at T0)

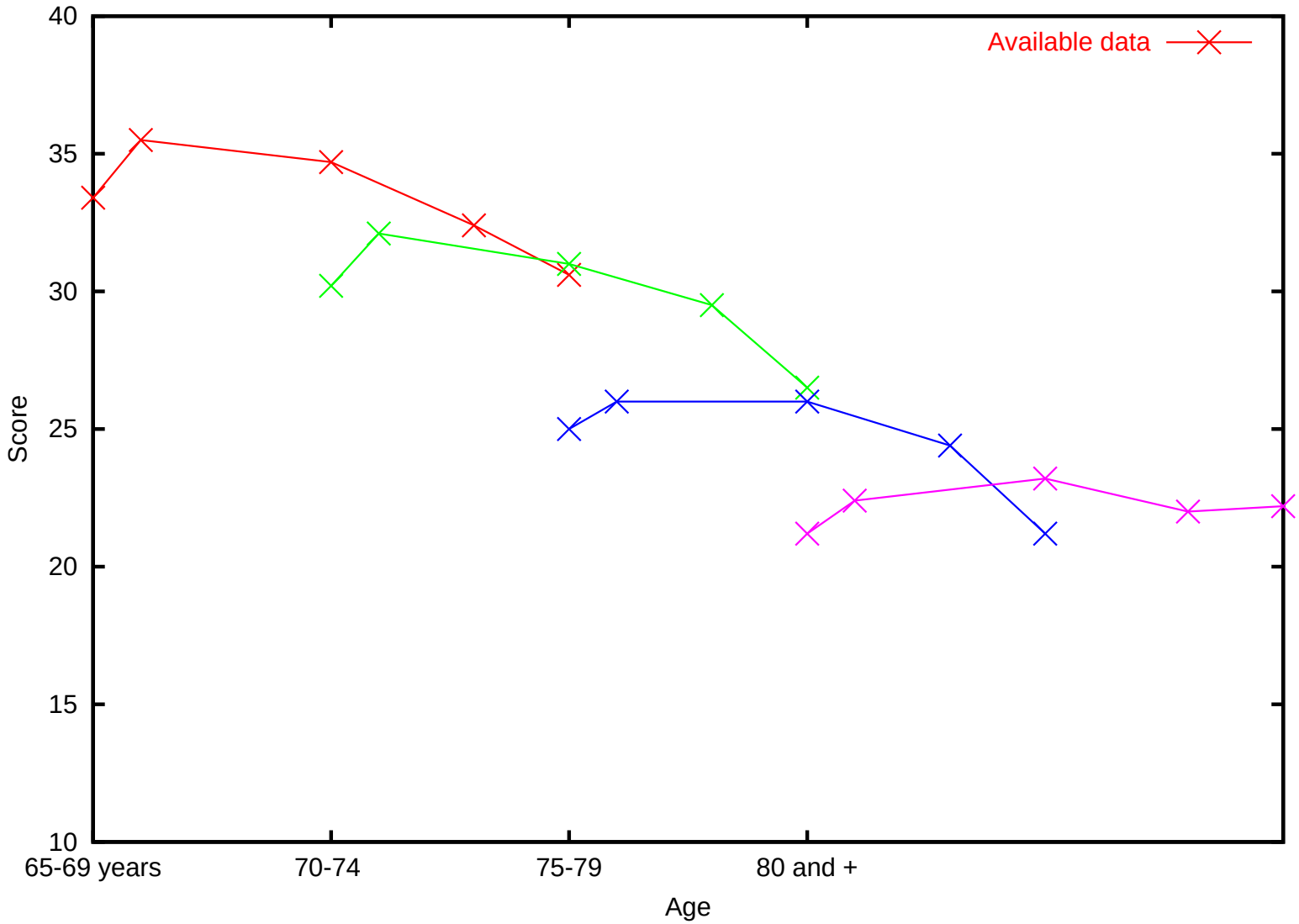
Description of dropout : Kaplan-Meyer

Dropout time (=event) : first visit with missing score

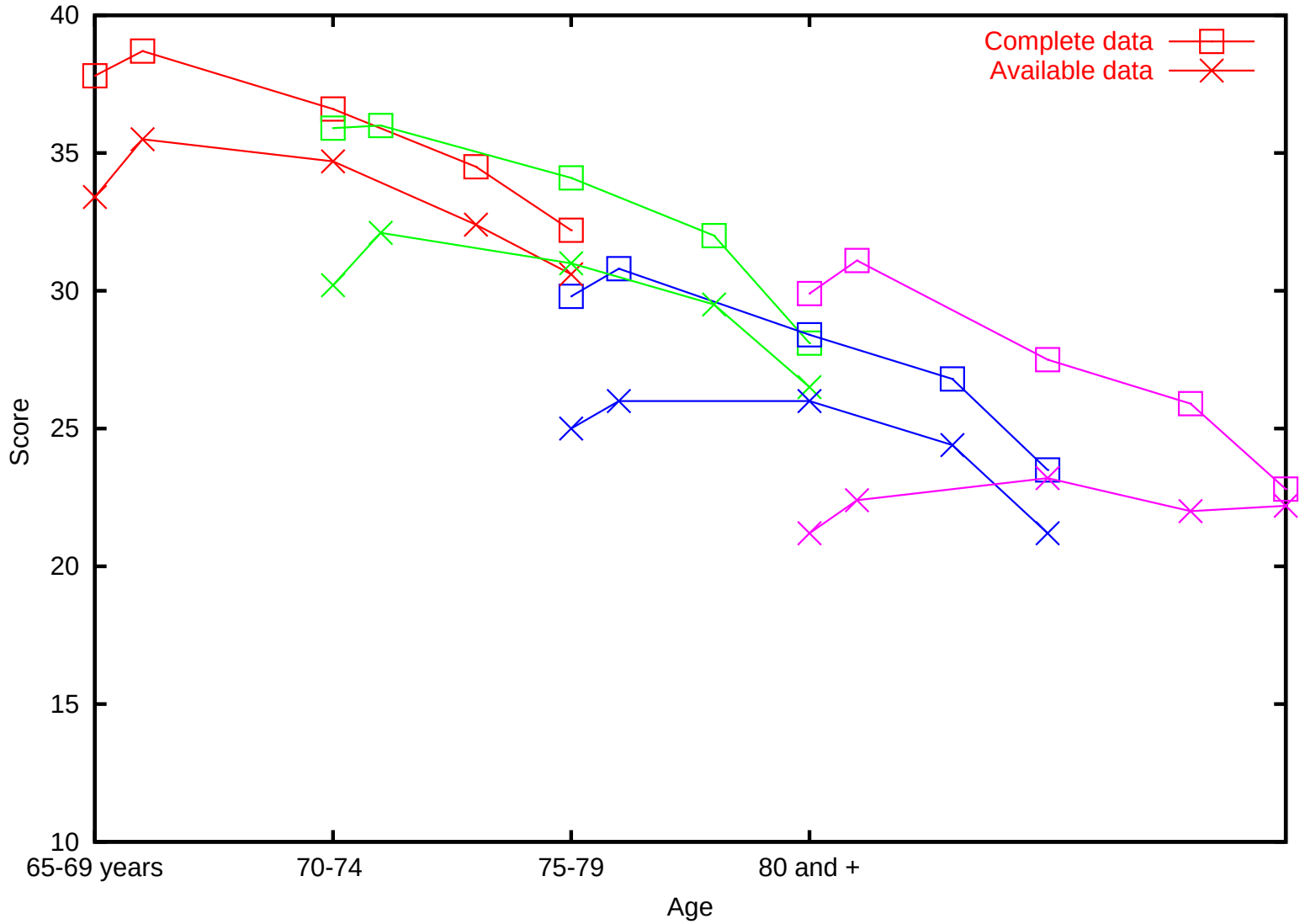
Probability to be in the cohort



Observed means of the DSST score given time



Observed means of the DSST score given time



Logistic regression model for dropout in the first 5 years

Covariates	OR	95% CI of the OR
T3	0.02	0.003 - 0.10
T5	0.01	0.001 - 0.09
age	1.01	0.99 - 1.02
age $\times$ T3	1.05	1.03 - 1.08
age $\times$ T5	1.06	1.03 - 1.09
previous MMSE score	0.91	0.88 - 0.93
men	0.86	0.75 - 0.99
Education (vs university level)		
No education	1.88	1.15 - 3.07
no diploma	2.02	1.39 - 2.93
CEP	1.67	1.17 - 2.40
high school level	1.39	0.96 - 2.00

Methods for MCAR or MAR data

- Complete case analysis (loss of precision, require MCAR)
- Imputation (require MCAR or MAR)
- Maximum likelihood using available data (require MAR)

Maximum likelihood for MAR data (1)

Objective : Estimate θ from the distribution $f(Y|\theta)$
Likelihood of the observed data : Y_{obs}, R

$$f(Y_{obs}, R|\theta, \psi) = \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis}, \psi) dY_{mis}$$

Maximum likelihood for MAR data (2)

If the data are MAR :

$$\begin{aligned} f(Y_{obs}, R|\theta, \psi) &= \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, \psi) dY_{mis} \\ &= f(R|Y_{obs}, \psi) \int f(Y_{obs}, Y_{mis}|\theta) dY_{mis} \\ &= f(R|Y_{obs}, \psi) f(Y_{obs}|\theta) \end{aligned}$$

Log-likelihood :

$$l(\theta, \psi|Y_{obs}, R) = l(\theta|Y_{obs}) + l(\psi|R, Y_{obs})$$

If ψ and θ are distinct :
$\rightarrow$ the missing data are ignorable
$\rightarrow$ θ is estimated by maximisation of $l(\theta|Y_{obs})$ using only available reponses.

Example : MAR analysis of Paquid data

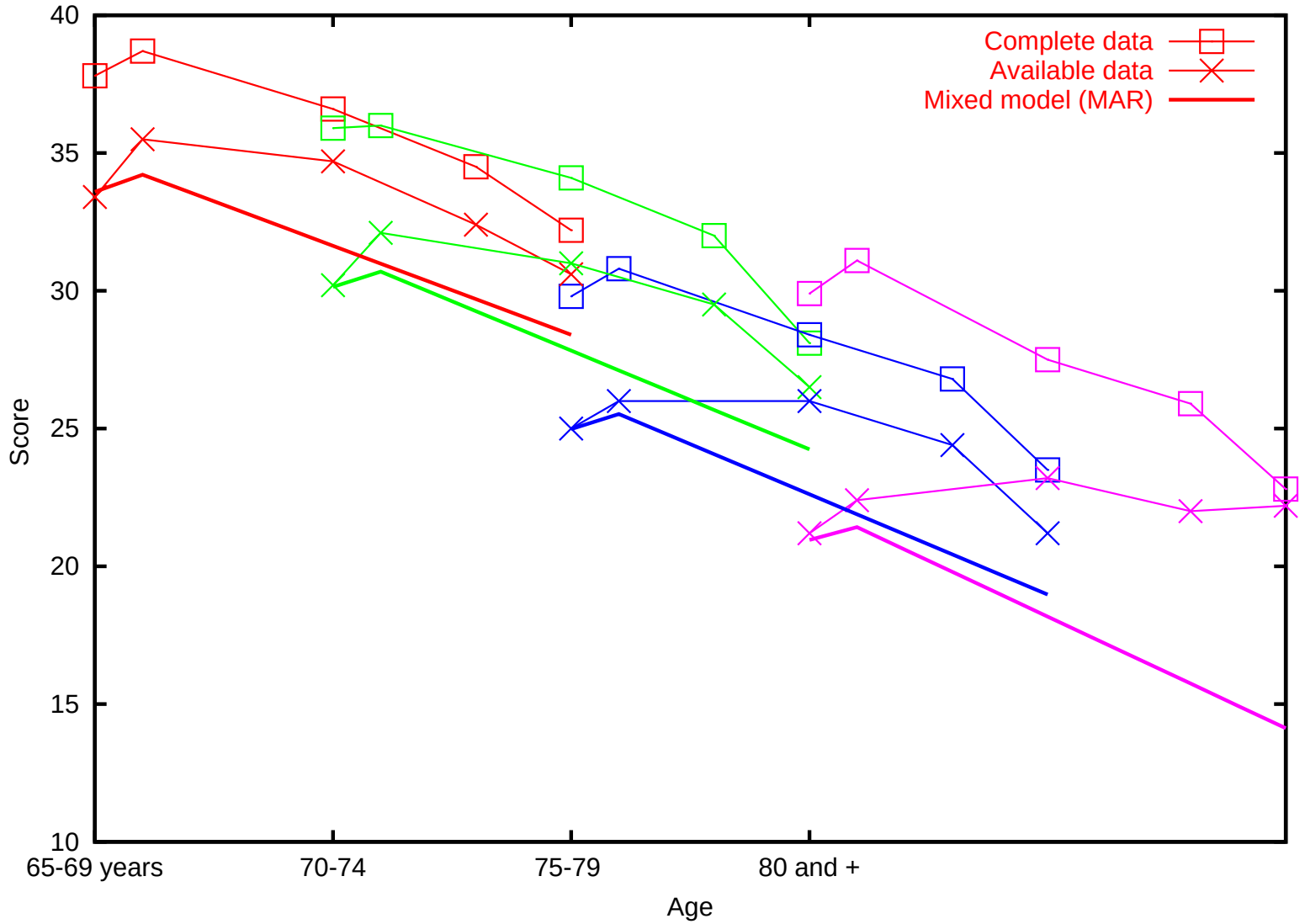
Mixed effect model

Y_{ij} test score for subject i at time t_{ij}

$$Y_{ij} = (\beta_0 + age_i'\gamma_0 + \alpha_{0i}) + (\beta_1 + age_i'\gamma_1 + \alpha_{1i}) \times t_{ij} + \beta_3 I_{\{t_{ij}=0\}} + e_{ij}$$

with
$\alpha_i = (\alpha_{0i}\alpha_{1i})^T \sim N(0, G)$, $e_{ij} \sim N\left(0, \sigma_e^2\right)$
age_i vector of indicators for baseline age classes (70-74, 75-79, 80 years and older , ref= 65-69)
$I_{\{t_{ij}=0\}}$ indicator of the baseline visit

Observed and predicted means of the score given time



Conclusion

Advantages of mixed models

- use all the available information (repeated measures)
- Flexibly handle intra-subject correlation (unbiased inference)
- Any number and times of measurements
- Robust to missing at random data
- Available in most softwares

Limits of mixed models

- Assume homogeneous population
  $\longrightarrow$ extended models included latent classes(mixture)
- As the MAR assumption is uncheckable, complete the study by a sensitivity analysis
  $\longrightarrow$ extended models for MNAR data

References

Chavance, M. et Manfredi R. Modélisation d'observation incomplètes . Revue d'Epidémiologie et Santé Publique 2000,48,389-400.

Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data .2nd Edition. Oxford Statistical Science series 2002, Oxford University Press.

Jacqmin-Gadda H, Commenges D, Dartigues JF. Analyse de données longitudinales gaussiennes comportant des données manquantes sur la variable à expliquer. Revue d'Epidémiologie et Santé Publique 1999, 47,525-534.

Little R.J.A. et Rubin D.B. Statistical Analysis with Missing Data , New York : John Wiley & Sons, 1987.

Verbeke G and Molenberghs G Linear mixed models for longitudinal data . Springer Series in Statistics, Springer-Verlag,2000, New-York.