Longitudinal + Reliability = Joint Modeling

Carles Serrat
Institute of Statistics and Mathematics Applied to Building

CYTED-HAROSA International Workshop
November 21-22, 2013 – Barcelona

Mainly from Rizopoulos, D (2012)
Joint Models for Longitudinal and Time-to-Event Data
with Applications in R
Chapman & Hall/CRC Biostatistics Series
Outline

1- Introduction

2- Longitudinal Data Analysis

3- Survival Analysis

4- The Standard Joint Model

5- Extensions of the Standard Joint Model
In follow-up studies, we are interested in studying the association structure between several longitudinal responses and the time until an event of interest (e.g. biomarkers with strong prognostic capabilities for even time outcomes)

- Dynamic nature (i.e. between patients and within patients across time)
Goals

- In follow-up studies, we are interested in studying the association structure between several longitudinal responses and the time until an event of interest (e.g. biomarkers with strong prognostic capabilities for even time outcomes)

- Dynamic nature (i.e. between patients and within patients across time)

- Former works by Self and Pawitan (1992) and DeGrutola and Tu (1994) in AIDS research

- Seminal papers by Faucett and Thomas (1996) and Wulfshon and Tsiatis (1997) introducing the “standard joint model”

A Motivating Dataset

- A cohort of 467 HIV-infected patients during antiretroviral treatment who had failed or were intolerant to zidovudine therapy.

- Main goal: To compare the efficacy of two alternative drugs, didanosine (ddI) and zalcitabine (ddC), in the time-to-death.

- Longitudinal information: CD4 cell counts at 0 (randomization), 2, 6, 12 and 18 months

- More details in Abrams et al. (1994)
A Motivating Dataset

- A cohort of 467 HIV-infected patients during antiretroviral treatment who had failed or were intolerant to zidovudine therapy.
- Main goal: To compare the efficacy of two alternative drugs, didanosine (ddI) and zalcitabine (ddC), in the time-to-death.
- Longitudinal information: CD4 cell counts at 0 (randomization), 2, 6, 12 and 18 months
- More details in Abrams et al. (1994)

Other Applications/Examples
- In sociology or educational testing
- In civil engineering or building construction
Inferential Objectives in Longitudinal Studies

Explicit versus implicit outcomes

- Explicit: Those variables explicitly specified in the study protocol
- Implicit: Those outcomes that are not of direct interest in the study but they condition the analysis (e.g. missing data or visit times issues)
Inferential Objectives in Longitudinal Studies

Explicit versus implicit outcomes

- Explicit: Those variables explicitly specified in the study protocol

- Implicit: Those outcomes that are not of direct interest in the study but they condition the analysis (e.g. missing data or visit times issues)

Research questions in longitudinal studies (Rizopoulos and Lesaffre, 2012)

- Effect of covariates on a single outcome

- Association between outcomes

- Complex hypothesis testing

- Prediction

- Statistical analysis with implicit outcomes
Linear Mixed-Effects Models

Let \( y_{ij} \) denote the response of subjects \( i, i = 1, \ldots, n \) at time \( t_{ij}, j = 1, \ldots, n_i \).
First linear approach:

\[ y_{ij} = \beta_{i0} + \beta_{i1} t_{ij} + \epsilon_{ij} \]

with \( \epsilon_{ij} \sim N(0, \sigma^2) \)

Second linear approach:

\[ y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{ij} + \epsilon_{ij} \]

where

- \( \beta = (\beta_0, \beta_1)' \) fixed effects
- \( b_i = (b_{i0}, b_{i1})' \) random effects with \( b_i \sim N_2(0, D) \)
- \( \epsilon_{ij} \sim N(0, \sigma^2) \)
LME formulation

\[
\begin{align*}
  y_i &= X_i \beta + Z_i b_i + \epsilon_i \\
  b_i &\sim \mathcal{N}(0, D) \\
  \epsilon_i &\sim \mathcal{N}(0, \sigma^2 I_{n_i})
\end{align*}
\]

where

- \(X_i\) and \(Z_i\) known design matrices for the fixed and random effects
- \(I_{n_i}\) denotes the \(n_i\)-dimensional identity matrix
- \(b_i\) are supposed to be independent on \(\epsilon_i\)
LME formulation

\[
\begin{cases}
y_i = X_i \beta + Z_i b_i + \epsilon_i \\
b_i \sim N(0, D) \\
\epsilon_i \sim N(0, \sigma^2 I_{n_i})
\end{cases}
\]

where

- $X_i$ and $Z_i$ known design matrices for the fixed and random effects
- $I_{n_i}$ denotes the $n_i$-dimensional identity matrix
- $b_i$ are supposed to be independent on $\epsilon_i$

Main advantages

- It allows to describe how the mean response changes in the population
- It allows to estimate individual response profiles over time
- It can accommodate any degree of imbalanced data
- The random effects part accounts for the correlation structure between the repeated measurements for each subject in a relative parsimonious way
- Errors can be modeled like $\epsilon_i \sim N(0, \Sigma_i)$, if it is necessary (Verbeke and Molenberghs, 2000; Pinheiro and Bates, 2000)
LME estimation

The conditional (hierarchical) formulation implies the marginal model for $y_i$

$$y_i = X_i \beta + \epsilon_i^* \text{ with } \epsilon_i^* \sim \mathcal{N}(0, V_i = Z_i D Z_i' + \sigma^2 I_{n_i})$$

▶ If $V_i$ is known $\beta$ can be estimated by generalized least squares.

▶ If $V_i$ is not known, $\beta$ is estimated by REML (Harville, 1974)

▶ Standard errors for the fixed-effects via robust estimation by sandwich estimator

EM algorithm (Dempster et al., 1977) and Newton-Raphson algorithms (Lange, 2004) are needed.
Implementations can be found in Laird and Ware (1982) and Lindstrom and Bates (1988).
Two main packages has been implemented

- **nlme** package (Pinheiro et al., 2012; Pinheiro and Bates, 2000) for continuous data and complex error structures.

- **lme4** package (Bates et al., 2011) for continuous and categorical responses and correlation in the repeated measurements only using random effects.

JM package by Dimitris Rizopoulos has been implemented considering the **lme** class of objects coming from the **lme()** function in the **nlme** package.
Illustration in R
Notation and definitions

- Let $T_i^*$ be a true survival time of interest with density function $f$

- Survival function: $S(t) = P(T^* > t) = \int_t^\infty f(s)ds$

- Hazard function: $h(t) = \lim_{dt \to 0} \frac{P(t \leq T^* < t + dt | T^* \geq t)}{dt}$

Consequently, $S(t) = \exp \left\{ -\int_0^t h(s)ds \right\}$.

Under the presence of right censoring....

- Let $C_i$ be the censoring time

- $\delta_i = I(T_i^* \leq C_i)$ the event indicator

- $T_i$ the observed survival time, i.e. $T_i = \min\{T_i^*, C_i\}$
Non-parametric approach: K-M estimator (Kaplan and Meier, 1958; Greenwood, 1926)

Semi-parametric approach: Proportional Hazards model (Cox, 1972), by maximizing the partial loglikelihood function

Under the Relative risk regression models

\[ h_i(t|w_i) = h_0(t) \exp(\gamma' w_i) \]

where

- \( w_i' = (w_{i1}, \ldots, w_{ip}) \) is a vector of covariates
- \( \gamma' = (\gamma_1, \ldots, \gamma_p) \) is the corresponding regression coefficients

and the ratio of hazards for two subjects \( i \) and \( k \) is

\[ \frac{h_i(t|w_i)}{h_k(t|w_k)} = \exp\{\gamma'(w_i - w_k)\} \]
Time dependent covariates

**Exogenous versus Endogenous covariates**

- Exogenous or external: when the covariate vector $y(.)$ is associated with the rate of failure over time, but its future path up to time $t > s$ is not affected by the occurrence of failure at time $s$. It is a predictable process *(Kalbfleisch and Prentice, 2002)* *(e.g. time of the day, season of the year, predetermined administrative therapy, environmental factors,...)*

- Endogenous or internal: otherwise. *(e.g. often measurements taken on the subjects under study, like biomarkers and clinical parameters)*
  - Typically measured with error
  - Their complete path up to time $t$ is not fully observed
Extended Cox Model: Implementation

The Cox model can be extended to handle exogenous time-dependent covariates (Andersen and Gill, 1982)

\[ h_i(t|\mathcal{Y}_i(t), w_i) = h_0(t)R_i(t) \exp(\gamma'w_i + \alpha y_i(t)) \]

where

- \( \mathcal{Y}_i(t) \) is the covariate history of \( y_i \) up to time \( t \)
- \( R_i(t) \) is a left continuous at risk process
  - \( (R_i(t) = 1 \) iff subject \( i \) is at risk a time \( t \)

and parameters \( \gamma \) and \( \alpha \) are again estimated by partial loglikelihood maximization.

**Implementation:** `survival` package (Therneau and Lumley, 2012)
- \( \text{Surv()} \) and \( \text{coxph()} \) functions.
Extended Cox Model: Illustration in R
The survival submodel: Notation and definitions

- **Aim:** To measure the association between the longitudinal marker level and the risk for an event

- Let $m_i(t)$ be the true and unobserved value of the longitudinal outcome at time $t$ (Remark: $m_i(t) \neq y_i(t)$)

- Let $M_i(t) = \{m_i(s), 0 \leq s < t\}$ be the longitudinal process up to time $t$

- The relative risk model is formulated in the form

$$h_i(t|M_i(t), w_i) = h_0(t) \exp(\gamma'w_i + \alpha m_i(t))$$

**Remark:** To let $h_0(t)$ without specifying may lead to an underestimation of the standard errors of the parameters (Hsieh *et al.*, 2006)

**Solution:** Explicitly define $h_0(t)$. 
The survival submodel (cont’)

Options for specifying the baseline risk

- To use known parametric distributions
- To use parametric but flexible specifications of baseline hazard
  - Step functions and linear splines (Whittemore and Killer, 1986)
  - B-splines (Rosenberg, 1995)
  - Restricted cubic splines (Herndon and Harrell, 1996)

Under the piecewise-constant model we formulate

\[ h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q) \]

where

- \( 0 = v_0 < v_1 < \ldots < v_Q \) denotes a partition of the time scale, with \( v_Q \) larger than the larger observed time
- \( \xi_q \) constant hazard in the interval \((v_{q-1}, v_q]\)
The longitudinal submodel

By using the linear mixed effects paradigm $y_i(t)$ is modeled like

$$
\begin{align*}
  y_i(t) & = m_i(t) + \epsilon_i(t) \\
  m_i(t) & = X_i'(t)\beta + Z_i' b_i \\
  b_i & \sim \mathcal{N}(0, D) \\
  \epsilon_i(t) & \sim \mathcal{N}(0, \sigma^2)
\end{align*}
$$

where

- $x_i(t)$ and $z_i(t)$ are time-dependent design vectors and $\epsilon_i(t)$ is also time-dependent
- errors terms are mutually independent and independent of the random effects.
The longitudinal submodel

Intuitive representation of joint models

![Graph showing the longitudinal submodel](image-url)
Implementation of Joint Models in R

- JM package by Dimitris Rizopoulos (2010, 2012) follows the random effects strategy. Currently only works with linear mixed-effects submodels with iid error terms and no serial correlation structure.

- The main function is jointModel() that needs and lme class of mixed-effects model under an unstructured variance-covariance matrix for the random effects and a coxph model for the survival submodel. method argument in jointModel() allows piecewise-PH-GH, spline-PH-GH, Cox-PH-GH, weibull-PH-GH and weibull-AFT-GH specifications for the baseline hazard function.
Further reading


- Asymptotic properties under unspecified baseline hazard (Zeng and Cai, 2005)

- Bayesian estimation of joint models using MCMC (Hanson \textit{et al.}, 2011; Chi and Ibrahim, 2006, Xu and Zeger, 2001)

- Conditional score approach for the random effects as a nuisance parameter (Tsiatis and Davidian, 2001)
Parameterizations (1/3)

- Interaction effects

\[ h_i(t) = h_0(t) \exp(\gamma'w_{i1} + \alpha'\{w_{i2} \times m_i(t)\}) \]

- Lagged effects

\[ h_i(t) = h_0(t) \exp(\gamma'w_i + \alpha m_i\{\max(t - c, 0)\}) \]
Parameterizations (2/3)

- Time-Dependent slopes parameterization

\[ h_i(t) = h_0(t) \exp(\gamma'w_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)) \]
Parameterizations (3/3)

- Cumulative effects parameterization

\[ h_i(t) = h_0(t) \exp\{\gamma w_i + \alpha \int_0^t m_i(s) \, ds\} \]

- Random effects parameterization

\[ h_i(t) = h_0(t) \exp(\gamma' w_i + \alpha' b_i) \]
More on the Standard Joint Model

- To handle Exogenous time-dependent covariates
- To fit stratified relative risk models
- Allows for Multiple failure times (e.g. competing risks or recurrents events)
- To fit accelerated failure time models
- Diagnostics and Prediction
Prediction examples

Subject 2

Subject 25

log (serum Bilirubin) vs Time

Survival Probability vs Time
Prediction examples
Prediction examples

Subject 2

Subject 25

\[ \log(\text{serum Bilirubin}) \]

Time

Survival Probability

\[ 0 \quad 5 \quad 8 \]

\[ 0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1.0 \]