12. Selected topics

Sonja Greven

Summer Term 2017

Overview Chapter 12 - Selected topics

12.1 Joint models for longitudinal and event time data

12.2 Stochastic time-varying covariates

12.3 Sample size in longitudinal studies

Longitudinal and time-to-event data

Often, longitudinal and time-to-event data are collected together and the longitudinal data is only available until the event occurs. Examples:

- Longitudinal auto-antibodies after seroconversion and time to onset of type I diabetes
- CD4 cell counts after seroconversion and onset of HIV
- Longitudinal measurements of the prothrobin marker and time to death in liver cirrhosis patients

We then observe longitudinal measurements $y_{i1}, \ldots, y_{in_i}$ and the event time or censoring time T_i with $t_{in_i} \leq T_i$ and event indicator δ_i (1 if subject i experiences the event, 0 if it is censored).

Challenges in modeling this type of data

The longitudinal marker $y_i(t_{ij})$ for subject $i = 1, \cdots, n$ is

- measured at varying time points t_{ij}
- measured with error
- subject to informative dropout (no measurements after event onset)

Aim: Estimating the relationship between marker and time to event T_i

Joint models

1. submodel for the true trajectories, e.g. a mixed model

$$
\begin{aligned}
y_{ij} = y_i(t_{ij}) &= m_i(t_{ij}) + \epsilon_i(t_{ij}) \\
&= \mathbf{x}_i(t_{ij})^\top \boldsymbol{\beta} + \mathbf{z}_i(t_{ij})^\top \mathbf{b}_i + \epsilon_i(t_{ij})
\end{aligned}
$$

2. submodel for time-to-event, e.g. proportional hazards model

$$\lambda_i(u) = \lambda_0\left(u\right) \exp\left\{\alpha \cdot m_i\left(u\right)\right\}$$

3. combined in a *joint* likelihood to avoid biases in two-step estimation approach (first estimating 1., then plugging results into 2.)

$$f(T_i, y_i(t_{ij})) = \int f(T_i | b_i) f(y_i(t_{ij}) | b_i) f(b_i) db_i$$

Estimation

- Inference is based on the EM-algorithm or on Bayesian approaches.
- Joint models are a broad class of different models, e.g. different specifications of the link between longitudinal and survival.
- Joint models are implemented in different R-packages, e.g. `JM`, `JMbayes`, `bamlss`, and `lcmm` (latent class model).

Further readings

- Article on the `JM`-package (Section 1 and 2 give a clear and short overview)
  Rizopoulos, D.(2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9): 1-33.

- Standard review paper on the class of joint models
  Tsiatis, A.A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 14: 809-834.

- Overview of latent class approaches
  Proust-Lima, C., Sene, M., Taylor, J.M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods of Medical Research*, 23: 74-90.

Overview Chapter 12 - Selected topics

12.1 Joint models for longitudinal and event time data

12.2 Stochastic time-varying covariates

12.3 Sample size in longitudinal studies

Stochastic time-varying covariates

Different types of covariates:

- Time-invariant covariates for each subject, e.g. gender, race, treatment group
- Time-varying covariates:
  - design-related, e.g.
    - time since baseline and its transformations such as t^2
    - treatment in a "crossover" study
  - stochastic time-varying covariates, e.g.
    - dietary intake
    - bloodmarker
    - air pollution
    - physical activity

Stochastic time-varying covariates

- In our models, we assumed a relationship for the mean

$$g(\mathsf{E}(Y_{ij}|\mathbf{X}_i)) = \mathbf{x}_{ij}^T\boldsymbol{\beta}.$$

- This implicitly assumes that $\mathsf{E}(Y_{ij}|\mathbf{X}_i)$ depends only on $\mathbf{x}_{ij}$:

$$\mathsf{E}(Y_{ij}|\mathbf{X}_i) = \mathsf{E}(Y_{ij}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}) = \mathsf{E}(Y_{ij}|\mathbf{x}_{ij}). \tag{12.1}$$

  This is true for time-invariant variables. For time-varying stochastic covariates, however, preceding or subsequent values of $\mathbf{x}_{ij}$ can 'confound' the relationship between Y_{ij} and $\mathbf{x}_{ij}$ and $\widehat{\boldsymbol{\beta}}$ can then be biased.

External and Internal Covariates

A covariate is called exogenous or external when

$$f(\mathbf{x}_{i,j+1}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{ij},Y_{i1},\ldots,Y_{ij}) = f(\mathbf{x}_{i,j+1}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{ij}).$$

Otherwise, the covariate is called internal or endogenous.

Examples:

- air pollution measured at a central monitor is external, as it does not depend on health outcomes
- personal air pollution exposure is internal if subjects with poor health outcomes change their behavior to avoid high air pollution exposures.

For an external covariate (and automatically for design-related covariates),

$$\mathsf{E}(Y_{ij}|\boldsymbol{X}_i) = \mathsf{E}(Y_{ij}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}) = \mathsf{E}(Y_{ij}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{ij}).$$

External Covariates

For external covariates, we can focus on specifying a model for $f(Y_{ij}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{ij})$. Possible models include

- concurrent, model $\mathsf{E}(Y_{ij}|\mathbf{x}_{ij})$
- lagged, model $\mathsf{E}(Y_{ij}|\mathbf{x}_{i,j-k})$ for some k
- cumulative, model $\mathsf{E}(Y_{ij}|\sum_{k=1}^{j}\mathbf{x}_{ik})$
- distributed lags, regression coefficients for $\mathbf{x}_{ij},\ldots,\mathbf{x}_{i,j-k}$ follow some pre-specified structure (e.g. polynomial).

Note that e.g. modeling $\mathsf{E}(Y_{ij}|\mathbf{x}_{ij})$ while Y_{ij} depends on both $\mathbf{x}_{ij}$ **and** $\mathbf{x}_{i,j-1}$ can give misleading results.

Internal Covariates

When variables are internal, we have to think both about meaningful targets of inference and valid methods of inference. Methods include causal inference, and modeling of the joint process $\{Y_{ij}, \mathbf{x}_{ij}\}$.

Overview Chapter 12 - Selected topics

12.1 Joint models for longitudinal and event time data

12.2 Stochastic time-varying covariates

12.3 Sample size in longitudinal studies

Sample size in longitudinal studies

As an example, assume that we have

- $N/2$ subjects **per group**
- $n_i = n$ measurements per subject (with equal time points t_j, but not necessarily equidistant)
- Two groups: placebo and therapy
- Model: LMM with linear group-specific trend, random intercept and slope per subject
- Null hypothesis: $\delta = 0$, where δ stands for the difference between the linear trends in groups A and B.

Sample size formula

For given type 1 error α and type 2 error β, the necessary sample size N to detect a difference δ then is obtained using

$$N/2 = \frac{(Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2 2\tilde{\sigma}^2}{\delta^2},$$

where

$$\tilde{\sigma}^2 = \sigma^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + d_{22}$$

with $\bar{t} = \sum_{j=1}^{n} t_j/n$, error variance σ^2 and random slope variance d_{22}. Thus, one needs to make assumptions about δ, σ^2 and d_{22} to calculate N.

The t_j are often chosen equidistantly with the study duration limited by organizational reasons. Then, N needs to be greater the smaller n is.

Comments and extensions

- The formula can be "reversed" to e.g. derive the power as a function of N.
- The formula can easily be adapted for groups of different sizes.
- The formula can easily be adapted for comparing other coefficients.
- The extension to non-normal responses is also possible.

For further discussion, see e.g. Diggle et al (2002), Fitzmaurice et al (2004).