lccsm.ado -- Summary of Statistical Methodology

Roseanne McNamee, Matthew Carr 12oct2013

Description

lccsm fits a two-stage survival model to longitudinal data with a time-to-event outcome. A stage 1 longitudinal mixed effects model is specified for a time-dependent covariate. Best linear unbiased (empirical Bayes) predictions of the underlying covariate trend are then obtained. Time-dependent (non-global) estimates of covariate variance are also calculated with the option of adjusting for the underlying covariate trend. In the second stage, a nested case-control study dataset is generated from the cohort-study dataset by sampling controls from the risk sets. For each case, the controls are chosen randomly from those members of the cohort who are at risk at the event time of the case. The resulting dataset is then analysed as a matched case-control study with odds ratios estimating the corresponding rate-ratio parameters in a proportional hazards model for the cohort study. Randomness in the matching is obtained using Stata's **runiform()** function. To ensure that the sample is random, the user should set the random-number seed; see **[R] set seed**.

Background on two-stage modelling

- Modelling a longitudinal covariate and associated risk relationship: two-stage approach (Tsiatis *et al.* (1995), Boshuizen *et al.* (2006)).
- Risk analyses can be subject to estimation attenuation or bias if original measures are used directly as covariate information due to measurement error and natural fluctuations (Prentice (1982)).

Example

To illustrate the statistical methodology, we use an example where a 3rd order polynomial function of time (with subject-specific random effects on each term) and baseline variables (basevar1 and basevar2) are used to model the longitudinal covariate. Trend adjusted time-dependent measures of covariate variance are calculated and used alongside the estimated covariate trend in a stage 2 nested case control model for a survival outcome.

Stage 1. Longitudinal model

The specified covariate is modelled using a longitudinal random effects model (with Stata's **xtmixed** command) as a polynomial function of observation/measurement time and baseline covariates. Defining t_{ij} as the observation time and y_{ij} as the measured covariate for the ith subject at the jth measurement point, we have

$$y_{ij} = (\alpha_0 + \beta_{0i}) + (\alpha_1 + \beta_{1i})t_{ij} + (\alpha_2 + \beta_{2i})t_{ij}^2 + (\alpha_3 + \beta_{3i})t_{ij}^3 + \alpha_4 \text{basevar1}_i + \alpha_5 \text{basevar2}_i + e_{ij}$$

- Fixed (population level) effects: $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_5)'$
- Random (subject level) effects: $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{3i})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi})$
- Measurement (random) error: $e_{ij} \sim \mathrm{N}(0, \theta)$

Matrix Notation

In vector form: $y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\alpha} + \mathbf{w}_{ij}'\boldsymbol{\beta}_i + e_{ij}$ with design vectors

- $\mathbf{x}_{ij} = (1, t_{ij}, t_{ij}^2, t_{ij}^3, \text{basevar1}, \text{basevar2})'$
- $\mathbf{w}_{ij} = (1, t_{ij}, t_{ij}^2, t_{ij}^3)'$

In matrix form: $\mathbf{y}_i = \mathbf{X}_i'\boldsymbol{\alpha} + \mathbf{W}_i'\boldsymbol{\beta}_i + \mathbf{e}_i$

- $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{im_i})$ and $\mathbf{W}_i = (\mathbf{w}_{i1}, \mathbf{w}_{i2}, \ldots, \mathbf{w}_{im_i})$

  where m_i is the number of observations for the ith subject.
- $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a diagonal matrix with entries θ

Obtaining Estimates of the Covariate Trend

- Fit the proposed model using maximum likelihood estimation with bootstrapping:

  $$\mathrm{L}(\boldsymbol{\alpha}, \boldsymbol{\psi}, \theta; \mathbf{Y}) = \prod_{i=1}^{N} \int_{\Omega_{\boldsymbol{\beta}}} \Pr(\mathbf{y}_i \mid \boldsymbol{\beta}_i; \boldsymbol{\alpha}, \boldsymbol{\psi}, \theta) \Pr(\boldsymbol{\beta}_i; \boldsymbol{\psi}) d\boldsymbol{\beta}_i$$

  - $\Pr(\mathbf{y}_i \mid \boldsymbol{\beta}_i; \boldsymbol{\alpha}, \boldsymbol{\psi}, \theta) \equiv \Pr(\mathbf{y}_i \mid \boldsymbol{\beta}_i; \boldsymbol{\alpha}, \theta)$

    $$= \frac{\exp\{-(\mathbf{y}_i - \mathbf{X}_i'\boldsymbol{\alpha} - \mathbf{W}_i'\boldsymbol{\beta}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{X}_i'\boldsymbol{\alpha} - \mathbf{W}_i'\boldsymbol{\beta}_i)/2\}}{(2\pi)^{m_i/2}\sqrt{|\boldsymbol{\Sigma}|}}$$

  - $\Pr(\boldsymbol{\beta}_i; \boldsymbol{\psi}) = \dfrac{\exp(-\boldsymbol{\beta}_i'\boldsymbol{\psi}^{-1}\boldsymbol{\beta}_i/2)}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\psi}|}}$

  An expectation maximisation (EM) algorithm is utilised to optimise the likelihood function with respect to the model parameters and non-parametric bootstrapping is used to estimate the standard errors.
- Test the fit of the model.
- Refit the model using REML to obtain estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\psi}$ and θ and unbiased estimates of the variance parameters.
- EB/BLUP estimation of subject-specific $\boldsymbol{\beta}_i$:

  $$\Pr(\boldsymbol{\beta}_i \mid \mathbf{y}_i; \hat{\boldsymbol{\alpha}}, \hat{\theta}, \hat{\boldsymbol{\psi}}) \propto \mathrm{L}(\boldsymbol{\beta}_i \mid \mathbf{y}_i; \hat{\boldsymbol{\alpha}}, \hat{\theta}, \hat{\boldsymbol{\psi}}) \Pr(\boldsymbol{\beta}_i; \hat{\boldsymbol{\psi}})$$

  where

  $$\mathrm{L}(\boldsymbol{\beta}_i \mid \mathbf{y}_i; \hat{\boldsymbol{\alpha}}, \hat{\theta}, \hat{\boldsymbol{\psi}}) = \prod_j \Pr(y_{ij} \mid \boldsymbol{\beta}_i; \hat{\boldsymbol{\alpha}}, \hat{\theta}, \hat{\boldsymbol{\psi}})$$
- EB/BLUP predictions of covariate trend:

  $$\hat{y}_{ij}^* = \mathbf{x}_{ij}'\hat{\boldsymbol{\alpha}} + \mathbf{w}_{ij}'\hat{\boldsymbol{\beta}}_i$$

Estimation in Continuous Time

- Predicting the covariate trend on $0 < t \le T_i$ where T_i is the event/censoring time:

  $$\hat{y}_i^*(t) = \mathbf{x}_i(t)'\hat{\boldsymbol{\alpha}} + \mathbf{w}_i(t)'\hat{\boldsymbol{\beta}}_i$$

  - $\mathbf{x}_i(t) = (1, t, t^2, t^3, \text{basevar1}_i, \text{basevar2}_i)'$

Estimating Covariate Variance

Many studies use global measures of variability or standard deviation as covariates in survival models. We argue that current measures of variability should only be based on current and historical observations on the covariate of interest. As such, we employ the 'Standard Residual' (adjusted for the modelled trend): $SR_i(t) = SR_{ij}$ for $t_{ij} \le t < t_{i,j+1}$ and $j \ge 3$ where

- $$SR_{ij} = \frac{\sum_{k=1}^{j} (y_{ik} - y_i^*(t_{ik}))^2}{j-1}$$

Alternative measures include:

(i) the 'Coefficient of Variation' (adjusted for the modelled trend):
$CV_i(t) = CV_{ij}$ for $t_{ij} \le t < t_{i,j+1}$ and $j \ge 3$ where

- $CV_{ij} = SR_{ij} / y_i^*(t_{ij})$

(ii) the 'Root Successive Variance' (adjusted for the modelled trend):
$RSV_i(t) = RSV_{ij}$ for $t_{ij} \le t < t_{i,j+1}$ and $j \ge 3$ where

- $$RSV_{ij} = \frac{\sum_{k=2}^{j} \{(y_{ik} - y_i^*(t_{ik})) - (y_{i,k-1} - y_i^*(t_{i,k-1}))\}^2}{j-2}$$

Stage 2: Survival model

- Using predicted values of the underlying covariate trend from the stage 1 longitudinal model.
- Nested case control method: cases are matched with *q* controls on time, basevar1 and the number of measurement points (numvisit) using Stata's **stset** and **sttocc** commands.
- Inclusion-exclusion criteria used in order to to avoid confounding; see Lubin & Gail (1984) and Greenland & Thomas (1982).
- The controls are selected so that they are identical to a case in terms of the potentially confounding factors.

Conditional logistic regression and 1:*q* case-control matching are used where set k has case event time A_k. We define Z_{kl} as an event indicator for individual l in set k (where $Z_{k0} = 1$ and $Z_{kl} = 0$ for $l = 1, 2, \ldots, q$) and specify

$$\Pr(Z_{kl} = 1) = \frac{\exp(c_k + \gamma y_{kl}^*(A_k))}{1 + \exp(c_k + \gamma y_{kl}^*(A_k))}$$

The c_k are conditioned out as

$$\Pr(Z_{k0} = 1 \mid V_k = 1) = \frac{\exp(\gamma y_{k0}^*(A_k))}{\sum_{l=1}^{q} \exp(\gamma y_{kl}^*(A_k))}$$

where $V_k = \sum_{l=0}^{q} Z_{kl}$. The coefficient γ is estimated using MLE and the conditional likelihood;

$$L(\gamma) = \prod_k \frac{\exp(\gamma y_{k0}^*(A_k))}{\sum_{l=1}^{q} \exp(\gamma y_{kl}^*(A_k))}$$

The likelihood function is of the same form as a stratified Cox partial likelihood, and is maximised using Cox model estimation with a common arbitrary time value used for all individuals and case-control set stratification.