EXERCISES with R: Advanced survival models and prediction for correlated data

Virginie Rondeau, Agnieszka Krol
Short Course - Karolinska Institutet
Stockholm, October 24-25, 2016

The package `frailtypack` provides various functions for models for correlated outcomes and survival data. Illustrations on the models included in the package, ie. Cox proportional hazard models, shared frailty models for recurrent events (clustered data), nested frailty models, additive frailty models, (multivariate) joint frailty models for recurrent events and a terminal event, joint models for longitudinal data and a terminal model, trivariate joint models for longitudinal data, recurrent events and a terminal event, were presented in papers dedicated to the package.[5,7,8]

In this section we focus on extended models for correlated data using datasets: `cdg`, `kidney` (from the package `survival`), `readmission`, `bcos`, `colorectal` and `colorectalLongi` (from the package `frailtypack`).

1 Shared frailty models for recurrent events

1.1 Dataset `readmission`

For this illustration we will use the dataset `readmission`. It contains rehospitalization data of patients diagnosed with colorectal cancer.[4] The data describe the calendar time (in days) of the successive hospitalizations after the date of surgery. The first readmission time was considered as the time between the date of the surgical procedure and the first rehospitalization after discharge related to colorectal cancer. Each subsequent readmission time was defined as the difference between the current hospitalization date and the previous discharge date. It contains information on patients characteristics: type of treatment, sex, Dukes' tumoral stage, comorbidity Charlson's index and survival status. A total of 861 rehospitalization events were recorded for the 403 patients included in the analysis. Several readmissions can occur for the same patient, and an individual frailty may influence the occurrence of subsequent rehospitalizations. Among the patients 112 (28%) died during the study.

Questions:

1.1.1 Make a description of the datafile, especially for the recurrent events: number of events, time between events, number of censored patients.

1.1.2 Fit a simple Cox model (using `frailtypack`) and interpret the effects of the chosen prognostic factors.

1.1.3 Fit a suitable shared frailty model with the timescale of your choice and interpret the effects of the chosen prognostic factors.

1.1.4 Do you observe an intra-subject correlation?

1.1.5 Evaluate the fitted model using the martingale residuals.

1.1.6 Present the frailties predicted from the model in function of individual number of events to detect outlying individuals.

1.1.7 Which distributions for the frailties do you prefer in terms of goodness-of-fit?

1.1.8 Is a stratification on gender justified here?

1.2 Dataset `kidney`

The dataset `kidney` includes information on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored. The data contains information on 38 patients, each individual has exactly 2 observations. Following patients' characteristics were included in the data: age, sex, disease type, and frailty estimates from the original paper.[6]

Questions:

1.2.1 Make a description of the datafile.

1.2.2 Fit a shared frailty model and interpret the effects of the chosen prognostic factors.

1.2.3 Do you observe an intra-subject correlation?

1.2.4 Plot the estimated baseline survival function.

1.2.5 Compare the fit of models with different estimation of baseline hazards: semiparametric models using splines (with equidistant or percentile intervals) and parametric models using Weibull distribution or piecewise-constant functions (with equidistant or percentile intervals).

2 Shared frailty models for clustered data

2.1 Dataset `bcos`

The dataset `bcos` includes interval-censored data for 94 breast cancer patients who were randomized to either radiation therapy with chemotherapy or radiation therapy alone.[2] The outcome is time until the onset of breast cosmetic deterioration which is interval-censored between the last clinic visit before the event was observed and the first visit when the event was observed. Patients without breast deterioration were right-censored. No covariates are included in the data.

Questions:

2.1.1 Make a description of the datafile.

2.1.2 Fit a simple Cox model with interval-censoring data and interpret the effect of the treatment.

2.1.3 Create 20 artificial groups (considered as hospitals or cancer centers) of individuals and fit a shared frailty model with interval censoring.

2.1.4 Compare the approach with a naive model that considers the interval midpoints as event times.

2.1.5 Do you observe a correlation in these artificial groups?

3 Nested frailty models for recurrent events in clustered data

3.1 Dataset `cgd`

In the dataset `cgd` recurrent infection times on 128 patients from 13 hospitals were observed.[3] These hierarchical data have two levels of clustering: the patient level (`id`) and hospital level (`center`). Data are from a placebo controlled trial of gamma interferon in chronic granulotomous disease (CGD). The event at the individual level correspond to a serious infection. The dataset includes patients' characteristics measured at baseline: treatment (placebo or gamma interferon), sex, age, height, weight, pattern of inheritance, use of steroids and use of prophylactic antibiotics.

Questions:

3.1.1 Make a description of the datafile.

3.1.2 Fit a standard frailty model for the recurrent events using a chosen timescale. Do you observe an intra-subject correlation? Which factors are prognostic for recurrent events?

3.1.3 Fit a standard frailty model for the clustered data using a chosen timescale. Do you observe an intra-group correlation? Which factors are prognostic for the grouped events?

3.1.4 Fit a nested frailty model for hierarchical data using a chosen timescale and estimation method for the baseline hazard function and compare the model to the standard frailty models from questions 3.1.2 and 3.1.3.

3.1.5 Interpret the model by considering the prognostic factors, variance of the frailties and the estimation of the baseline hazard function.

3.1.6 Plot the baseline hazard functions obtained from models with the calendar and gap time scales on one figure.

4 Joint frailty models for recurrent events and a terminal event

4.1 Dataset `readmission`

We use the dataset readmission from Section 1.1.

Questions:

4.1.1 Could you justify why do we need to consider a joint frailty model here?

4.1.2 Fit a joint frailty model for rehospitalizations and death using time scale of your choice. If you apply splines for the baseline hazard functions, you can use the values of kappa from the reduced models from Section 1.1.

4.1.3 Which factors are associated with the recurrent events and the terminal events?

4.1.4 Are the recurrent events and the terminal event processes associated?

4.1.5 Could you evaluate the goodness-of-fit of the model?

4.1.6 Could you check if the variable `sex` has a time-constant effect (check the proportional hazards assumption for this variable)?

4.1.7 Compare predicted risks of death for two specific subjects with histories of rehospitalizations. Consider two cases for predictions for intervals $[t, t + w]$: predictions with fixed prediction time t and moving window w, and predictions with moving prediction time t and fixed window w.

4.1.8 Compare predictive accuracy (using EPOCE) of two joint models with different baseline hazard functions estimation (eg. splines vs. Weibull).

4.1.9 Compare predicted risks of a new recurrent event for two subjects with different histories of rehospitalizations. Again, consider different settings for predictions horizons (as in question 4.1.7).

4.1.10 Fit a model that assumes the same effect of the frailty on both processes.

4.1.11 Fit a model that assumes two independent frailties (a general joint frailty model). Give an interpretation of the dependencies present in the model.

5 Joint frailty models for survival processes of clustered data

5.1 Dataset `readmission`

We use the dataset readmission from Section 1.1. We artificially create clusters on individuals. The first survival event will be the first observed rehospitalization and the second event, death. The framework of semi-competing risks is used here, thus individuals' follow-up stops at time of the rehospitalization, death or in case when none of these events are observed, the censoring time. We consider 6 clusters defined by a new variable group:

```
readmission <- transform(readmission, group = id %% 6 + 1 )
readmcluster <- subset(readmission, (t.start == 0 & event == 1)
 | event == 0)
```

Questions:

5.1.1 Fit a joint frailty model accounting for the clustering of the data. Assume the same effect of the frailty on both processes.

5.1.2 Interpret this model, especially the variance of the random effects.

5.1.3 Fit a shared frailty model for clustered data for the event of death. Compare two individuals in terms of predicted risk of death using marginal and conditional predictions.

6 Joint models for longitudinal data and survival processes

6.1 Datasets `colorectal` and `colorectalLongi`

For this illustration we will use the dataset `colorectal` and `colorectal.Longi`. The datasets include a random selection of 150 patients from a multi-center randomized phase III clinical trial FFCD 2000-05 of patients diagnosed with metastatic colorectal cancer.[1] The data contains a follow-up of tumor size measure (sum of the longest diameters of target lesions) and times of apparition of new lesions as recurrent events. Moreover, some baseline characteristics (age, WHO performance status and previous resection), treatment arm

(combination vs. sequential) and time of death (or last observed time for a right-censored individual) are included in the data. Dataset `colorectal` provides information on recurrent event and death and dataset `colorectalLongi` on the measurements of tumor size.

6.1.1 Define the three outcomes of interest.

6.1.2 Make a description of the datafile, especially of the outcomes of interest.

6.1.3 What kind of transformation could be used for the biomarker?

6.1.4 Create a dataset for the process of terminal event.

6.1.5 Fit an adapted joint model for the left-censored biomarker and death and evaluate the goodness-of-fit.

6.1.6 Which are prognostic factors for the risk of death and for the evolution of the tumor size?

6.1.7 How the longitudinal biomarker and the survival process are associated?

6.1.8 Could you predict the risk of death for a specific subject given the history of the biomarker?

6.1.9 Fit the trivariate joint model using initial values for covariates regression coefficients of the appropriate reduced models.

6.1.10 What are prognostic factors for the risk of appearance of new lesions and death, and for the evolution of the tumor size?

6.1.11 How the longitudinal marker and the survival processes are associated?

6.1.12 Evaluate the goodness-of-fit of the trivariate model.

6.1.13 Could you predict the risk of death for a specific subject given the history of the biomarker and recurrent event process?

6.1.14 Compare the predictive accuracy of the bivariate and trviariate joint models using the Brier Score (with modified functions ipcw and pecMethods) and the EPOCE.

References

[1] M. Ducreux, D. Malka, J. Mendiboure, P. L. Etienne, P. Texereau, D. Auby, P. Rougier, M. Gasmi, M. Castaing, M. Abbas, P. Michel, D. Gargot, A. Azzedine, C. Lombard-Bohas, P. Geoffroy, B. Denis, J. P. Pignon, L. Bedenne, and O. Bouché. Sequential versus combination chemotherapy for the treatment of advanced colorectal cancer (FFCD 2000-05): an open-label, randomised, phase 3 trial. *The Lancet Oncology*, 12(11):1032–44, 2011.

[2] Dianne M Finkelstein and Robert A Wolfe. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, pages 933–945, 1985.

[3] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

[4] J. R. Gonzalez, E. Fernandez, V. Moreno, J. Ribes, P. Merce, M. Navarro, M. Cambray, and J. M. Borras. Sex differences in hospital readmission among colorectal cancer patients. *Journal of Epidemiology and Community Health*, 59(6):506–511, 2005.

[5] Agnieszka Król, Audrey Mauguen, Yassin Mazroui, Alexandre Laurent, Stefan Michiels, and Virginie Rondeau. Tutorial in joint modeling and prediction: a statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software (In press)*, 2016.

[6] CA McGilchrist and CW Aisbett. Regression with frailty in survival analysis. *Biometrics*, pages 461–466, 1991.

[7] V. Rondeau and J. R. Gonzalez. frailtypack : A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine*, 80:154–164, 2005.

[8] V. Rondeau, Y. Mazroui, and J. R. Gonzalez. frailtypack : An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4), 2012.

	Cox	Shared	Nested	Additive	Joint standard *(Bivariate: 1 RE + 1 TE)*	Joint cluster *(Bivariate: 1 RE + 1 TE)*	Joint general *(Bivariate: 1 RE + 1 TE)*	Joint nested *(Bivariate: 1 RE + 1 TE)*	Joint longitudinal *(Bivariate: 1 LO + 1 TE)*	Joint trivariate *(Trivariate: 1 LO + 1 RE + 1 TE)*	Joint Multivariate *(Trivariate: 2 RE + 1 TE)*
`Available options`											
Gamma distribution		✕	✕		✕	✕	✕	✕			
Log-Normal distribution		✕		✕	✕				✕	✕	✕
Left-truncation	✕	✕	✕								
Interval Censoring	✕	✕			✕	✕					
Two strata	✕	✕	✕	✕							
More strata (max=6)	✕	✕			✕						
Time-dependant covariates	✕	✕			✕	✕					
Calendar timescale	✕	✕	✕		✕			✕		✕	✕
Weibull	✕	✕	✕	✕	✕	✕		✕	✕	✕	✕
Piecewise	✕	✕	✕	✕	✕	✕					✕
`Available output`											
Predicted frailties		✕	✕	✕	✕				✕	✕	
Variances of the frailties		✕									
Martingale residuals	✕	✕	✕	✕	✕				✕	✕	
`Model evaluation`											
Prediction of a terminal event	✕	✕			✕				✕	✕	
Prediction of a new recurrent event					✕						
Cmeasures	✕	✕									
Epoce					✕				✕	✕	

Table 1: Package characteristics. Blue cross is for option available for a given type of model in the package on CRAN, orange cross is for option included in the package but not on CRAN yet. Empty cells mean that an option is not available for a given type of model (either not coded yet or simply not applicable). RE = Recurrent Event, TE = Terminal Event, LO = Longitudinal Outcome