Análisis de datos longitudinales

Grado en Estadística

Tema 2 – Sesión 5

Análisis de Supervivencia

Eventos recurrentes (I)

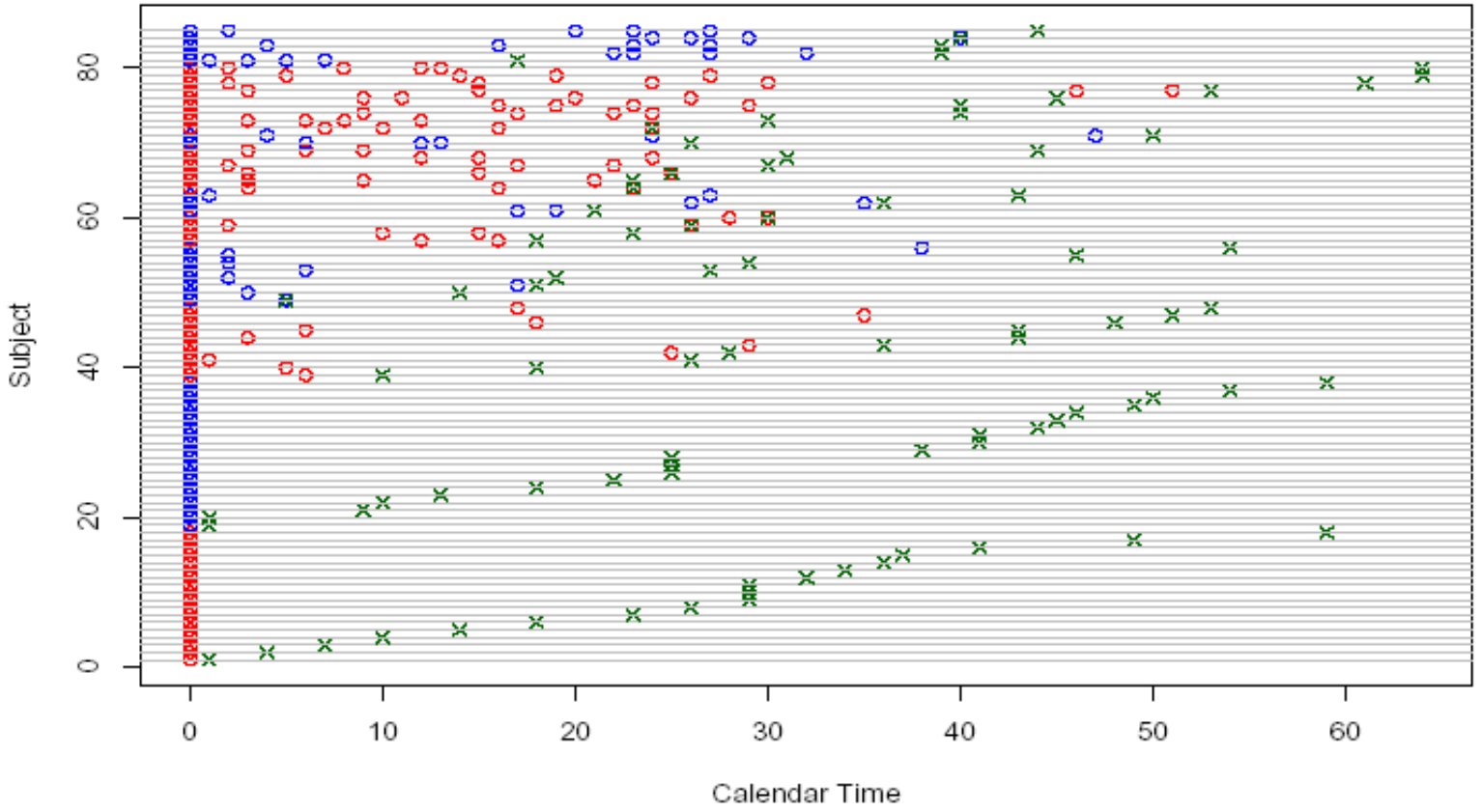
Juan R González

Departamento de Matemáticas, UAB

Instituto de Salud Global de Barcelona, ISGlobal

Data motivation (I)

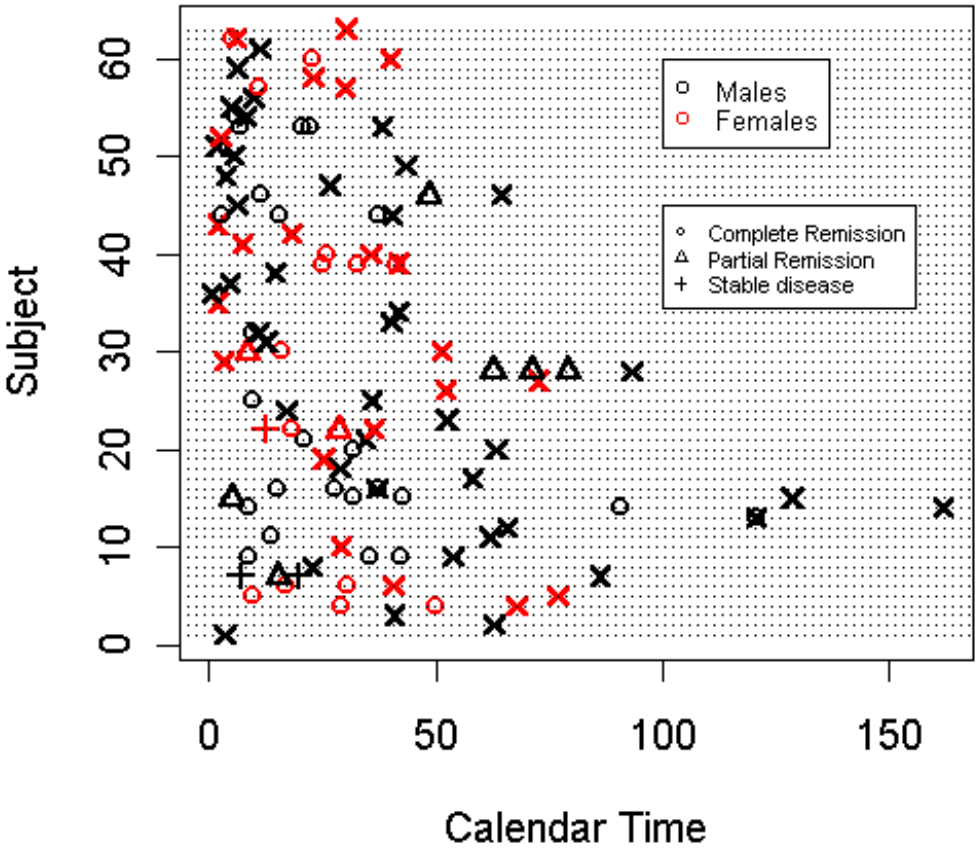
Wei, Lin, Weidsfeld'89 Bladder cancer data



Questions: Are there differences in disease-free survival between placebo and thiotepa groups? Heterogeneity? Impact of more events?

Data motivation (II)

González and Peña'04 Low grade lymphoma



Question: The sames than previous and How can affect the effect of intervention after each relapse

Background

Single event

- Applied in many areas
  - Sociology, Demography, …
  - **Biomedicine** (Survival analysis)
  - **Engineering** (Reliability analysis)
- Data characteristic
  - Time-to-event data
  - Censored data
- Statistical methods
  - Survival function (Kaplan-Meier, Nelson-Aalen)
  - Models: Parametric, Semi-parametric (Cox)

Repeated events

- Applied in many areas
  - Sociology, Demography, …
  - **Biomedicine** (Survival analysis)
  - **Engineering** (Reliability analysis)
- Data characteristic
  - Multiple time-to-events data
  - Censored data
- Statistical methods
  - Survival function (Peña-Strawderman-Hollander, Wang-Chang, Reliability models, …)
  - Models: Semi-parametric (Cox-extended models, Frailty models, Peña-Hollander Model)

Proceso contador

- Un proceso contador es un proceso estocástico $\{N(t), t \geq 0\}$ con valores positivos, enteros y crecientes tales que:
  - $N(t) \geq 0$
  - $N(t)$ es entero
  - Si $s \leq t$ entonces $N(s) \leq N(t')$
- Si $s<t$ entonces $N(t)-N(s)$ es el número de eventos que ocurren en el intervalo $(s, t]$. Ejemplos son los procesos de Poisson y los procesos de renovación
- Dada la tercera propiedad, un proceso contador es creciente, por lo tanto es una martingala. Usando el teorema de Dobb-Meyer, podemos escribir

$$N(t) = M(t) + A(t)$$

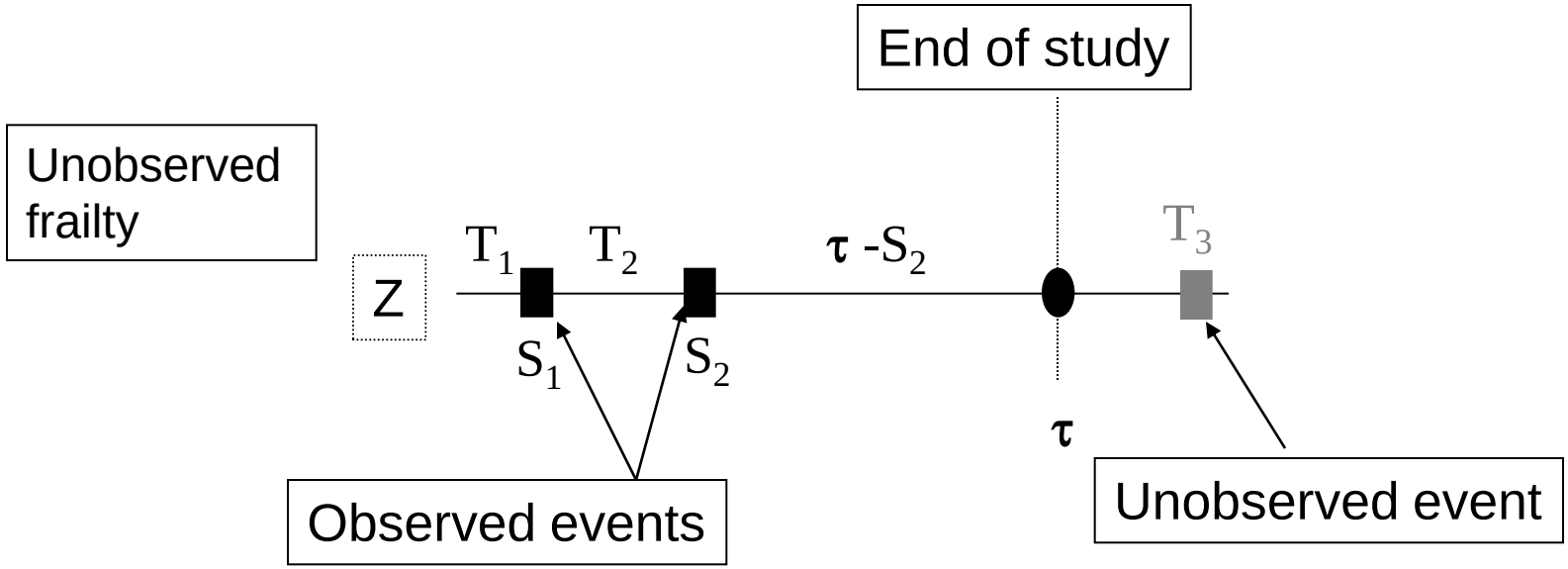
Donde $M(t)$ es una martingala y $A(t)$ es un proceso predecible

Procesos contadores

- El análisis de supervivencia se puede expresar como procesos contadores. Observamos $T=min(T,Z)$ $C=\{0,1\}$
  - Proceso contador: $N(t) = I(T \leq t, C=1)$
  - Proceso a riesgo: $Y(t) = I(Y \geq t)$
  - Proceso de intensidad: $\lambda(t)dt = Y(t)h(t)dt$
    con $h(t)= Pr(t \leq T < t+dt, C=1 \mid T \geq t)$
- El estimador de Kaplan-Meier sería:

$$S(t) = \prod_{s<t} 1 - \frac{dN(s)}{Y(s)}$$

Recurrent Events



- $T_1, T_2, T_3, \ldots$ = inter-event or gap times
- $S_1, S_2, S_3, \ldots$ = calendar times of event occurrences
- $X(s)$ = covariate vector, possibly time-dependent
- τ = end of observation period
- Accrued History: $F^{\dagger} = \{F^{\dagger}(s) : s \geq 0\}$
- Z = unobserved frailty variable
- $N^{\dagger}(s)$ = number of events in $[0, s]$
- $Y^{\dagger}(s)$ = at-risk indicator at time s

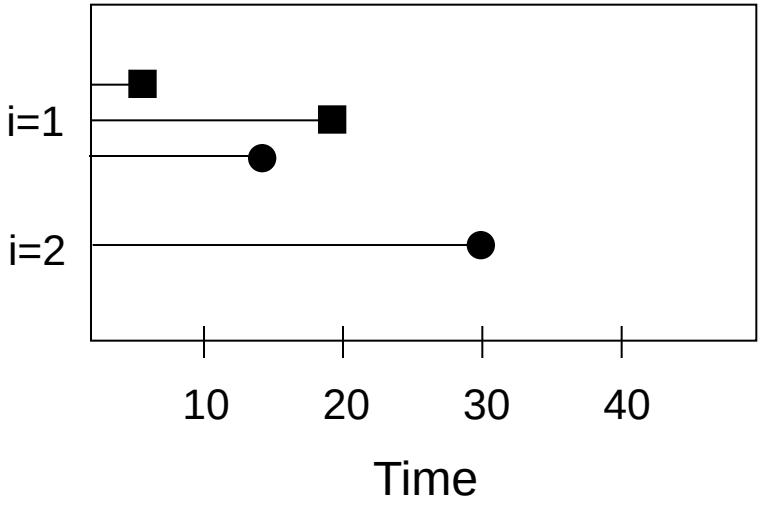
Background

Repeated nature causes:

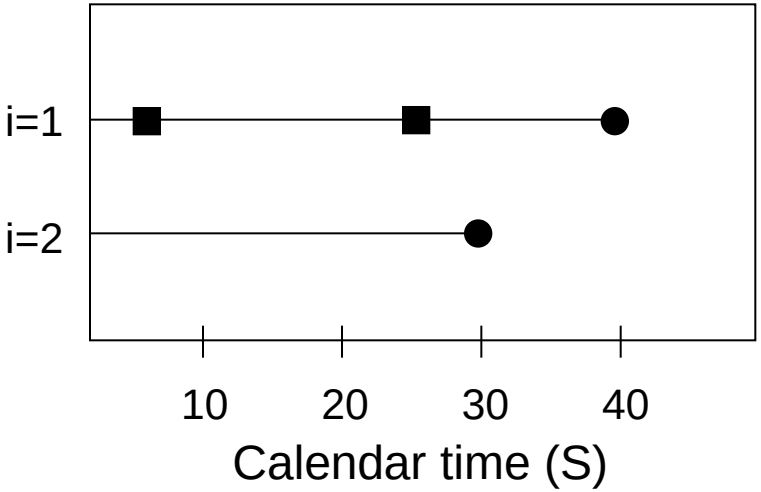
- At risk process
  - Gap time
  - Calendar time
- Within-subject correlation (no i.i.d.)
  - Heterogeneity across individuals
  - Event dependence
- Doubly-indexed processes

(Gill ‘81, Sellke 88, Peña 00)

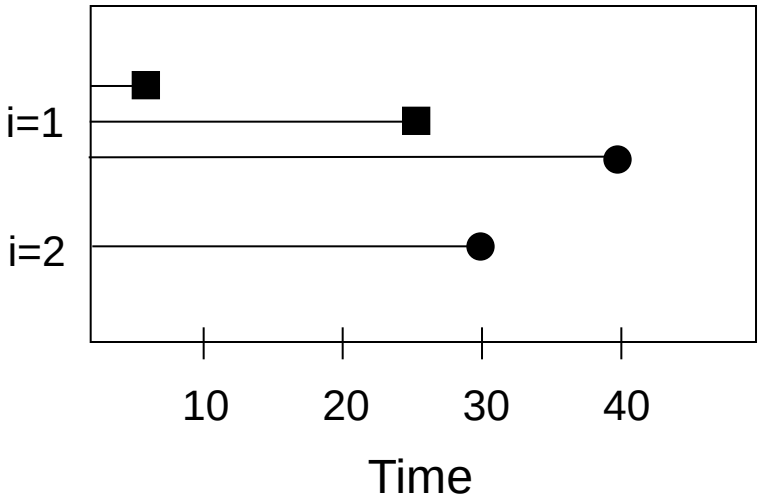
Gap time formulation (PWP-GT)



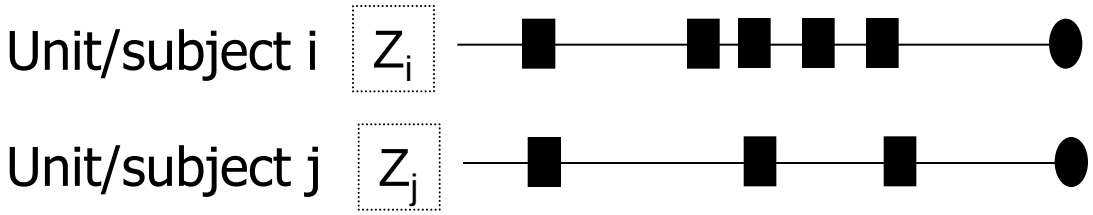
Observed recurrent events



Total time formulation (WLW, PWP-CP)



Within-subject correlation



- Biomedical data (uncontrolled variables, non-measurable variables: genetic susceptibility, …)
- F non i.i.d.
- There exists a random variable Z with known distribution. If we condition to Z=z the interocurrence times are i.i.d.

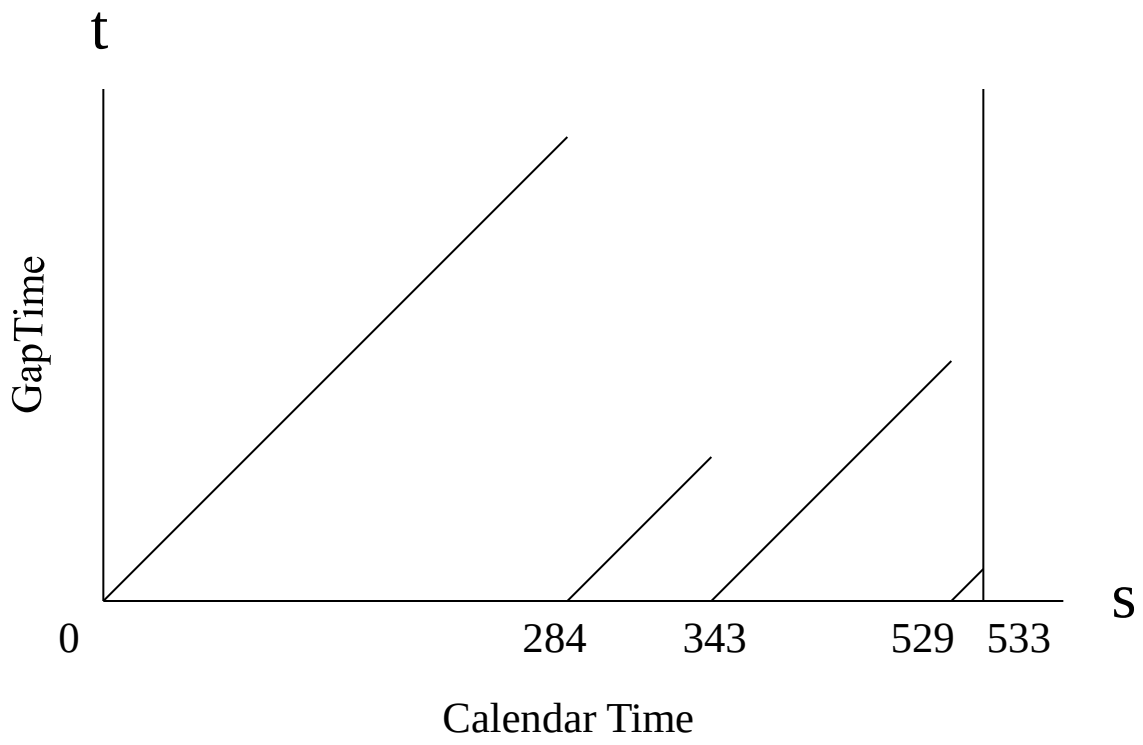
- Approaches:
  - Variance-corrected models (Cox-based models)
  - Frailty models

Survival function estimators

- Consider only the first, possibly right-censored, observation per unit and use the product-limit estimator (PLE).
  - Loss of information
  - Inefficient
- Ignore the right-censored last observation, and use empirical distribution function (EDF).
  - Leads to bias (“biased sampling”)
  - Estimator actually inconsistent

Survival function estimators

Needed: Calendar-Gaptime Space (double indexed Processes)



Survival function estimators

$$Z_i(s,t) = I\{s - S_{iN_i^\dagger(s-)} \le t\} \qquad N_i(s,t) = \int_0^s Z_i(v,t) N_i^\dagger(\mathrm{d}v)$$

$$A_i(s,t) = \int_0^s Z_i(v,t) A_i^\dagger(\mathrm{d}v)$$

$$M_i(s,t) = \int_0^s Z_i(v,t) M_i^\dagger(\mathrm{d}v) = N_i(s,t) - A_i(s,t)$$

$$Y_i(s,t) = \sum_{j=1}^{N_i^\dagger(s-)} I\{T_{ij} \ge t\} + I\{(s \wedge \tau_i) - S_{iN_i^\dagger(s-)} \ge t\}$$

- $N_i(s,t)$ = # of events in calendar time [0,s] for the ith unit with gaptimes at most t

- $Y_i(s,t)$ = number of events in [0,s] for the ith unit with gaptimes at least t.

Survival function estimators

Ver artículo Gonzalez and Peña, 2004

- Peña-Strawderman-Hollander'01 (GPLE) (i.i.d. model)

$$\hat{\bar{F}}(s,t) = \prod_{w \le t} \left[1 - \hat{\Lambda}(s, \mathrm{d}w)\right] = \prod_{w \le t} \left[1 - \frac{N(s, \Delta w)}{Y(s,w)}\right]$$

- Peña-Strawderman-Hollander'01 (FRMLE) (frailty model)
  - Frailties $Z_1, Z_2, \ldots, Z_n$ i.i.d. H_z. Si Z_i son i.i.d Gamma(α,α)

$$\bar{F}(t) = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^{\alpha}$$

Survival function estimators

- Wang-Chang'99 (WC): Includes both i.i.d and gamma frailty models

$$\hat{S}(t) = \prod_{i=1}^{n} \prod_{\{j:\ T_{ij} \le t\}} \left[1 - \frac{d^*(T_{ij})}{R^*(T_{ij})}\right]$$

$$K_i^* = \begin{cases} 1 & \text{if} \quad K_i = 0 \\ K_i & \text{if} \quad K_i > 0 \end{cases} \qquad d^*(t) = \sum_{i=1}^{n} \left\{ \frac{I\{K_i > 0\}}{K_i^*} \sum_{j=1}^{K_i} I\{T_{ij} = t\} \right\}$$

$$R^*(t) = \sum_{i=1}^{n} \frac{1}{K_i^*} \left[\sum_{j=1}^{K_i} I\{T_{ij} \ge t\} + I\{\tau_i - S_{iK_i} \ge t\} I\{K_i = 0\} \right]$$

Comparing survival curves

- Comparing survival curves
  - There exist asymptotic forms for GPLE and WC variances and NOT for FRMLE
  - Variability of median survival may be computed using resampling techniques (Efron'82, Bickel'81, Beran'82, Singh'81)

Comparing survival curves

Ver artículo Gonzalez, Delicado, Peña 2010

- Study several bootstrapping schemes for estimating the sampling distribution of estimators of the median survival with recurrent events.
- Construct bootstrap confidence intervals
- Mechanism for comparing median survival for different groups

Bootstrapping Schemes

Bootstrapping the observed data

Obtain B i.i.d samples from

$$\{(K_i^*, \tau_i^*, T_{i1}^*, T_{i2}^*, \ldots, T_{iK_i}^*, \tau_i^* - S_{iK_i}^*), i = 1, 2, \ldots, n\}$$

with replacement, from the observed sample

$$\{(K_i, \tau_i, T_{i1}, T_{i2}, \ldots, T_{iK_i}, \tau_i - S_{iK_i}), i = 1, \ldots, n\}$$

Bootstrapping Schemes

Bootstrapping T_{ij}^*'s from F (or S)

Step 1. Take $\tau_i^* = \tau_i$

Step 2. From the distribution (or), continue generating an i.i.d sequence of T_{ij}^*'s until K_i^* where

Step 3. The bootstrap sa $\sum_{j=1}^{K_i^*} T_{ij}^* \le \tau_i^* < \sum_{j=1}^{K_i^*+1} T_{ij}^*.$

Step 4. For this b ated median estimate $\left(K_i^*, \tau_i^*, T_{i1}^*, T_{i2}^*, \ldots, T_{iK_i^*}^*, \tau_i^* - S_{iK_i^*}^*\right)$

Bootstrapping Schemes

Plan VI. Semiparametric bootstrap

Step 1. Given the data, estimate $\hat{\alpha}$ and $\hat{\Lambda}_0$. Then estimate the distribution using

$$\hat{\bar{F}}_0(t) = \prod_{\{j:\ t_j \le t\}} \left[1 - \Delta\hat{\Lambda}_0(t_j)\right]$$

Step 2. Generate $Z_1^*, Z_2^*, \ldots, Z_n^*$ according to a $\text{Gamma}(\hat{\alpha}, \hat{\alpha})$

Step 3. Take $\tau_i^* = \tau_i$

Step 4. From $\hat{\bar{F}}_0(t)$ continue generating an i.i.d sequence of T_{ij}^*'s until K_i^* where

$$\sum_{j=1}^{K_i^*} T_{ij}^* \le \tau_i^* < \sum_{j=1}^{K_i^*+1} T_{ij}^*.$$

Bootstrapping Schemes

Plan VI. Semiparametric bootstrap (cont.)

Step 5. The bootstrap sample for the *i* th unit is

$$(K_i^*, \tau_i^*, T_{i1}^*, T_{i2}^*, \ldots, T_{iK_i^*}^*, \tau_i^* - S_{iK_i^*}^*)$$

Step 6. For this bootstrap sample, compute FRMLE, and the associated median estimate

Bootstrapping Schemes

To take into account the length of period (censored time) we bootstrapping τ_i^*'s from G_n instead of take the same τ_i's

Bootstrapping Schemes

Plan I. Bootstrapping the observed data

Plan II. Bootstrapping T_{ij}^{*}'s from PSH estimator

Plan III. Bootstrapping T_{ij}^{*}'s from PSH estimator and τ_{ij}^{*}'s from G_n

Plan IV. Bootstrapping T_{ij}^{*}'s from WC estimator

Plan V. Bootstrapping T_{ij}^{*}'s from WC estimator and τ_{ij}^{*}'s from G_n

Plan VI. Semiparametric bootstrap

Plan VII. Semiparametric bootstrap and bootstrapping τ_{ij}^{*}'s from G_n

Simulation study

Simulation data:

- i.i.d model $\tau_i \sim Exp(\upsilon)$ $T_{ij} \sim Exp(\theta)$
- correlated model $T_{ij} \sim F_0(t \mid \theta) \sim Exp(\theta)$
- 2,000 samples and 500 bootstrap replicates (B=500)

For each sample:

- MSE (mean square error)
- 95% bootstrap percentile confidence interval (BPCI)
  - Empirical coverage
  - Mean, median and variance length of BPCI
- Generate using:

  $n \in \{15,50,80\}$, $\theta \in \{1/3,1/6\}$, $\nu=1$, $\alpha \in \{\infty,6,2\}$

Results. Case i.i.d, θ=1/3 and τ=1

	MSE ($\times 10^6$)	% due to Bias	95% bootstrap percentile confidence interval			
			% Emp. Cov.	Length		
				Mean	Median	Var. ($\times 10^6$)
n=15						
Plan I	2,830	8.1	88.4	0.19	0.17	10,625
Plan II	2,914	11.3	94.3	0.21	0.19	10,760
Plan III	2,916	11.9	95.2	0.22	0.20	11,748
Plan IV	7,037	10.3	93.6	0.32	0.28	33,831
Plan V	6,879	10.4	94.2	0.32	0.28	35,398
n=50						
Plan I	667	3.1	93.3	0.10	0.10	772
Plan II	662	3.7	94.6	0.10	0.10	653
Plan III	662	3.7	94.8	0.10	0.10	668
Plan IV	1418	3.2	94.9	0.15	0.15	2050
Plan V	1425	3.2	94.5	0.15	0.15	2026
n=80						
Plan I	391	2.1	94.1	0.08	0.08	357
Plan II	385	2.5	95.4	0.08	0.08	293
Plan III	387	2.6	95.2	0.08	0.08	290
Plan IV	847	1.9	95.3	0.12	0.12	903
Plan V	847	1.9	95.4	0.12	0.12	941

Results. Correlated case, θ=1/3 and τ=1

		MSE ($\times 10^6$)	% due to Bias	95% bootstrap confidence interval: % Emp. Cov.	Length: Mean	Length: Median	Length: Var. ($\times 10^6$)
α=2							
n=15	Plan IV	25,760	15.2	93.4	0.57	0.45	169,858
	Plan V	23,622	15.6	93.8	0.58	0.46	169,358
	Plan VI	18,764	7.2	92.1	0.45	0.34	131,154
	Plan VII	18,858	7.3	92.3	0.45	0.34	135,060
n=50	Plan IV	4,569	5.4	94.2	0.26	0.24	15,422
	Plan V	4,569	5.5	94.2	0.26	0.23	16,193
	Plan VI	2,582	0.1	93.8	0.20	0.19	5,187
	Plan VII	2,563	0.1	94.1	0.20	0.19	5,262
n=80	Plan IV	2,653	5.6	94.8	0.20	0.19	4,262
	Plan V	2,679	5.8	95.1	0.20	0.19	4,346
	Plan VI	1,676	0.2	94.4	0.16	0.15	2,069
	Plan VII	1,684	0.2	94.8	0.16	0.15	2,009
α=6							
n=15	Plan IV	12,034	12.7	92.9	0.40	0.33	71,607
	Plan V	11,549	12.9	93.1	0.41	0.33	70,850
	Plan VI	9,276	16.7	92.2	0.31	0.26	38,316
	Plan VII	9,217	16.9	92.7	0.31	0.26	46,671
n=50	Plan IV	2,093	4,0	93.7	0.18	0.17	3,192
	Plan V	2,085	4.1	93.9	0.18	0.17	3,523
	Plan VI	1,274	2.0	93.7	0.14	0.13	1,537
	Plan VII	1,271	2.0	93.5	0.14	0.13	1,514
n=80	Plan IV	1,208	3.7	95.3	0.14	0.14	1,579
	Plan V	1,217	3.9	95.4	0.14	0.14	1,584
	Plan VI	727	1.0	95.3	0.11	0.10	788
	Plan VII	725	1.0	95.2	0.11	0.10	777

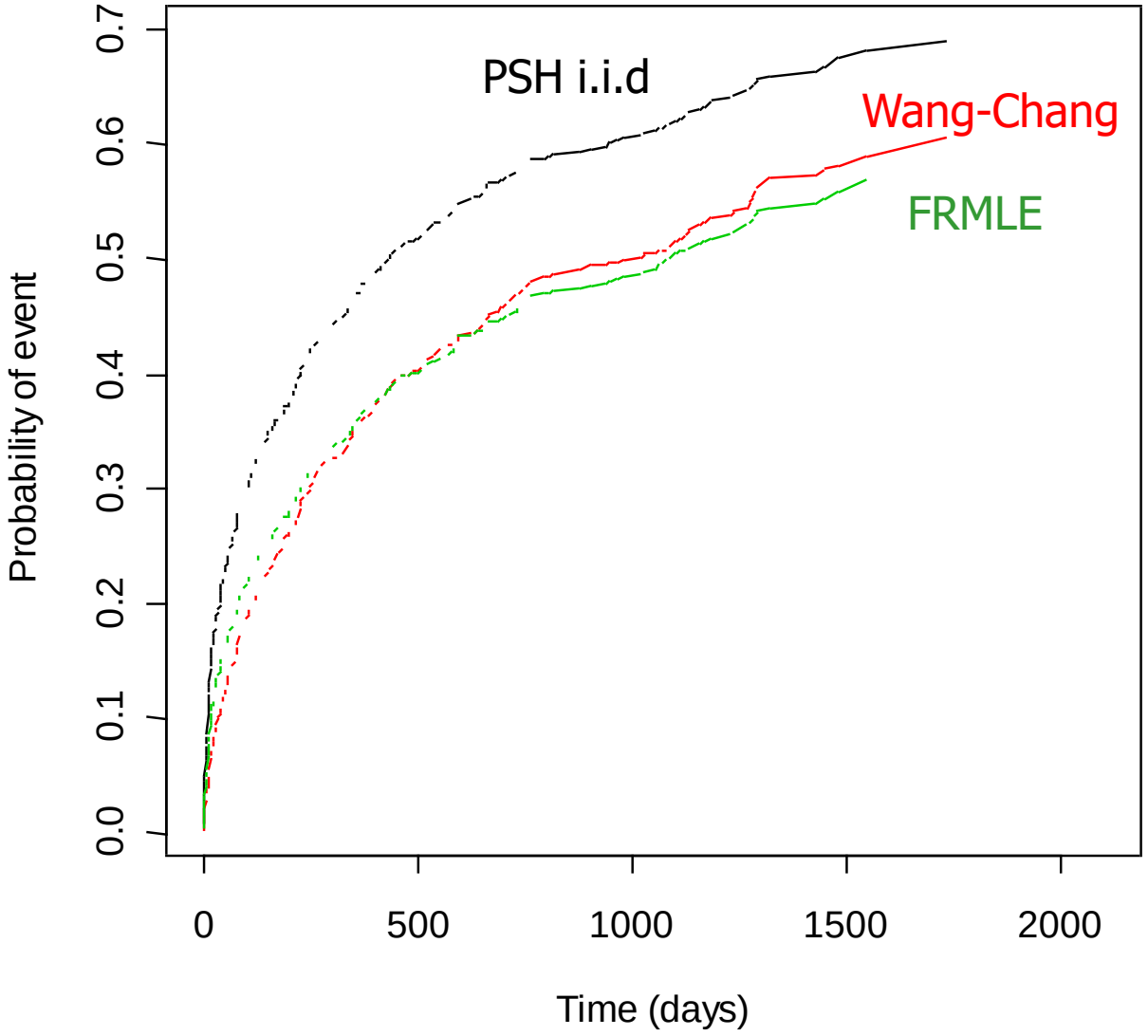
Application to a Hospital readmission data

- 403 cases with colorectal cancer (de *nuovo*) that have been operated
- Main variable: Time until the readmission related with colorectal cancer.
- Covariants: Tumoral stage (Dukes), age, sex.

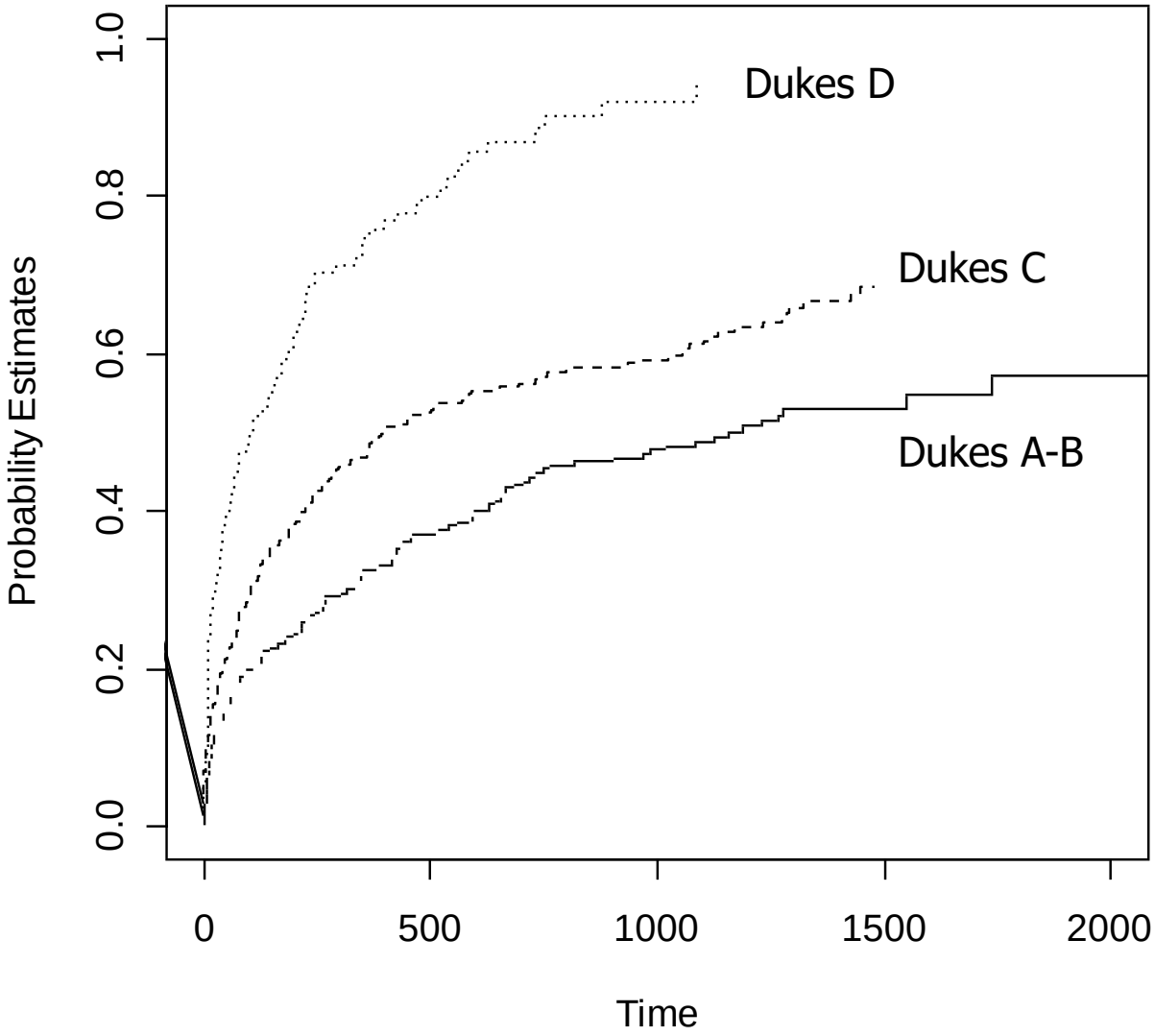
Patients' characteristics

	Number of readmissions (%)						
	0	**1**	**2**	**3**	**4**	**≥5**	**mean**
Sex							
Males	112 (46.9)	57 (23.8)	34 (14.2)	13 (5.4)	10 (4.2)	13 (5.4)	2.3
Females	87 (53.0)	48 (29.3)	11 (6.7)	8 (4.9)	5 (3.0)	5 (3.0)	1.9
Age							
<60	47 (42.3)	32 (28.8)	11 (9.9)	7 (6.3)	8 (7.2)	6 (5.4)	2.4
60-74	98 (50.5)	44 (22.7)	27 (13.9)	12 (6.2)	7 (3.6)	6 (3.1)	2.1
≥75	54 (55.1)	29 (29.6)	7 (7.1)	2 (2.0)	0 (0.0)	6 (6.1)	1.8
Dukes							
A-B	103 (57.2)	43 (23.9)	16 (8.9)	8 (4.4)	7 (3.9)	3 (1.7)	1.8
C	67 (45.3)	40 (27.0)	20 (13.5)	7 (4.7)	6 (4.1)	8 (5.4)	2.2
D	29 (38.7)	22 (29.3)	9 (12.0)	6 (8.0)	2 (2.7)	7 (9.3)	2.7

Validation correlated model



Results


Probability Estimates
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Time
Dukes D
Dukes C
Dukes A-B

Results

	Semiparametric (plan VII)			T^*_{ij} from WC (plan V)	
	α	Median (days)	CI95%	Median (days)	CI95%
Sex					
Male	0.99	799	(539,1171)	909	(524,1230)
Female	1.50	1427	(755,2175)	1222	(721,2175)
Age					
<60	1.22	799	(415,983)	718	(474,1134)
60-74	1.05	1230	(597,1427)	1104	(646,1547)
≥75	0.94	1188	(551,2175)	1188	(510,2175)
Dukes					
A-B	1.11	2175	(1188,∞)	1736	(1188,2175)
C	1.45	1073	(450,1288)	1028	(489,1325)
D	2.19	199	(109,297)	199	(161,350)

Concluding Remarks

- Plans anchored in PSH estimator are the best plans under i.i.d. model (plans I,II, and III)
- Semiparametric plans are the best plans under corelated model (plans VI and VII)
- Plans anchored in WC estimator (IV and V) offer a robust procedure when model that generated the data is not known.
- Bootstrapping from empirical distribution of the monitoring times do not provide improvements

The survrec Package

October 24, 2002

Date 2002-October-24

Title Survival analysis for recurrent event data

Author F77 original by Edsel A Peña <pena@stat.sc.edu> and Robert L Strawderman <rls@cornell.edu>. Added Fortran routines, R code and packaged by Juan R González <jrgonzalez@ico.scs.es>.

Maintainer Juan R González <jrgonzalez@ico.scs.es>

Depends none

Description Estimation of survival function for recurrent event data using Peña-Strawderman-Hollander, Whang-Chang estimators and MLE estimation under a Gamma Frailty model.