3-Lecture Minicourse on Statistics of Survival Data

Eric Slud

I. (11/6) **Death Hazards & Competing Risks**

Concepts:

(i) Statistical Estimation as mathematical problem,

(ii) Identifiability, nonparametric vs. nonparametric.

II. (11/13) **Population Cohorts & Martingales**

Concepts:

(iii) Counting process models,

(iv) "Innovations" and Statistics

III. (11/20) **Models and Likelihoods with ∞-Dimensional Parameters**

Concepts:

(v) Nuisance parameters,

(vi) Statistical Efficiency.

Lecture Slides (incl. annotated references) at :

www.math.umd.edu/∼evs/SurvSlid.pdf

Statistical Models for Causative Factors

Imagine a population cohort of individuals observed through a common window of time until either a discrete event ('endpoint') of interest occurs **or** the study ends.

Window may be defined by:

- chronological time-origin,
- individual time-origin e.g. entry into study, surgery,
- another time-scale, such as 'operational' time (reliability) or 'exposure' (epidemiology).

DATA:

(1) **explanatory** part: initial or *baseline* variables Z_i, including group-membership labels, together with time-varying measurements $V_i(t)$ (e.g., cumulative indicators of EKG anomalies or family disease history or blood pressure etc., and maybe age), plus

(2) **at-risk process** $Y_i(t)$ indicator $\{$ alive and under observation at time $t\}$; and

(3) **response** $N_i(t)$ cumulative count of observed events, such as 'death', 'recurrence of tumor', cumulative count of polyps, etc.

Response process initially 0, jumps only at times when $Y_i(t) = 1$.

PROBLEM: formulate, fit, test with data, a model for *prognosis* (probability or rate of event occurrence) as function of explanatory covariates.

Terminologies: prognostic-index, failure intensity, risk factor, survival regression model.

In setting of previous lecture:

$$N_i(t) = \Delta_i \, I_{[T_i \le t]} \quad , Y_i(t) \; = \; I_{[T_i \ge t]}$$

Now explicitly consider covariates in conditional prob's

$$P(T < t + \delta \,, \; \Delta = 1 \,|\, T > t, \, C > t, \, Z, \, V(s) : \; s \le t)$$

of essentially immediate death.

GENERAL APPROACH is to model

$$\lim_{\delta \to 0+} \; \frac{1}{\delta} \, E(\, N(t+\delta) - N(t) \mid Z, (Y(s), \, V(s), s \le t) \,)$$

Intensity Model says causation affected only by current detailed state

$$h_{T|Y,Z,V}(t) \; = \; Y(t) \, g(Z, \, V(t-))$$

Contrast with time-series, reliability problems in which dependence is on $V(s), \; s < t)$

Conditional Death Hazards

In general, recall **hazard intensity**

$$h_X(t) \;\equiv\; \lim_{\delta\to 0}\,\frac{1}{\delta}\,P(X\in(t,t+\delta)\,|\,X>t) \;=\; \frac{f_X(t)}{S_X(t)}$$

Then

$$h_X(t) = -\,\frac{d}{dt}\,\ln S_X(t) \quad\Longrightarrow\quad S_X(t) = \exp\Big(-\int_0^t h_X(s)\,ds\Big)$$

In presence of time-varying information:

$$h_{X|W}(t) \;=\; \lim_{\delta\to 0}\,\frac{1}{\delta}\,P(X\in(t,t{+}\delta)\,|\,X>t,\,(W(s),\;s\le t))$$

If $W(\cdot)$ process is *not influenced* by X-occurrence, and the influence of W on X is only prospective, can extend the formula:

$$P(X>t\,|\,(W(s),s\ge 0)) \;=\; \exp\Big(-\int_0^t h_{X|W}(s)\,ds\Big)$$

Counting Processes & Martingales

Formalization of Intensity Model :

Def'n: a **counting process** $(N(t),\, t \geq 0)$ is a nondecreasing right-continuous integer-valued process s.t.

$$a.s. \quad \{\Delta N(t)\}_{t\geq 0} \;=\; \{0,\, 1\}$$

Say it is **compensated** by left-continuous increasing process $A(t)$ which is a function of (measurable wrt the *filtration of* σ-algebras generated up to times t by) baseline variables Z and left-continuous processes $(Y(s),\, V(s),\, s \leq t)$ if $N - A$ is a **martingale**, i.e., for all $\delta > 0,\, t \geq 0$

$$E(N(t+\delta) - N(t) - A(t+\delta) + A(t)\,|Z,\, (Y(s), V(s))_{s\leq t}) \;=\; 0$$

Interpretation: for small δ

$N(t+\delta) - N(t)$ = indicator of event in $(t, t+\delta)$

$A(t+\delta) - A(t) \approx$ probability known before t

So $N(t+\delta) - N(t) - (A(t+\delta) - A(t))$ is *Observed* minus (conditional) *Expected Count* on $(t, t+\delta)$ from vantage point of just before t.

Intensity Model: $A(t) \;=\; \int_0^t Y(s)\, g(Z,\, V(s))\, ds$

See p.10 of File for Picture

Multiplicative Intensity Model

Cox (1972), Aalen (1978) introduced the class of models

$$E(N(t+dt) - N(t) \,|\, Z,\, (Y(s), V(s) :\, s < t))$$
$$= \; Y(t-)\, e^{\beta' Z + \gamma' V(t-)}\, \lambda(t)\, dt$$

Idea: parameters (β, γ) to be fitted describe effect on prognosis of individual subjects, while the (infinite-dimensional) **nuisance hazard function** $\lambda(t)$ describes the general background population. Exponent usable as *prognostic index*.

Research Agenda:

- *Applied:* Use and validation of these models in new data settings (e.g., time-dependent covariates for patient histories).
- *Generalization:* to non-multiplicative forms with regression part, e.g. arising from *frailties* or non-observable covariates.
- *Theoretical:* large-sample theory of *efficient* estimators for *semiparametric* models like these with ∞-dim nuisance parameters NEXT LECTURE.

Innovations & Statistics

Innovation means *new independent piece of information* $N(t+\delta) - N(t) - (A(t+\delta) - A(t))$.

In real-data situations, we obtain innovations from each of a large set $N_i(t)$ of counting processes. In some applications, multiple events (e.g. multiple recurrences of nonlethal tumors, polyps, etc. We search for non-chance *Observed – Expected* patterns among subsets of innovations (times, values i) defined through covariates Z_i, $V_i(t)$.

Limit Theorem: suppose that counting processes and associated predictors $(N_i(\cdot), Y_i(\cdot), V_i(\cdot), Z_i)$ are independent identically distributed across $i = 1, \ldots, n$ with intensities

$$A_i(t) \;=\; \int_0^t Y_i(s)\, g(Z_i,\, V_i(s))\, ds$$

Then for arbitrary sets B, C,

$$\int_0^t \sum_{i=1}^n I_{[Z_i \in B,\, V_i(s) \in C]}\, d(N_i - A_i)(s)$$

are asymptotically independent (for disjoint $B \times C$) normally-distributed variables with mean 0 and variance

$$\int_0^t \sum_{i=1}^n I_{[Z_i \in B,\, V_i(s) \in C]}\, dA_i(s)$$

Model-Building with 'Residual' Plots

If a specified intensity model is *correct*, then individual terms $N_i(t+\delta) - N_i(t)$ are coin-toss variables (from vantage point of $t-$), with heads-probability $A_i(t+\delta) - A_i(t)$ and variance

$$(A_i(t+\delta)-A_i(t))\,(1-A_i(t+\delta)+A_i(t)) \;\approx\; A_i(t+\delta)-A_i(t)$$

uncorrelated across different $i,\ t$. Plotting cum-sums of

Martingale Residuals $\qquad N_i(\infty) - A_i(\infty)$

and i ordered with respect to some external variable can indicate whether that variable belongs in the Intensity !

Did this with *Primary Biliary Cirrhosis* Data, 216-patient clinical trial in which BILIRUBIN is a very important predictive variable which was highly **unbalanced** across the treatment groups ! Successive (multiplicative intensity) models include: TRTGP only, then TRTGP+LOGBILI, then also ALBMN+CCHOL+CIRRH.

OBS	ID	DTH	TIME	TRTGP	LOGBIL	AGE	CIRRH	CCHOL	ALBMN
13	246	1	0.30	1	1.78	36.6	1	0	26
14	229	0	0.62	1	2.63	90.0	1	1	33
15	78	1	1.25	0	2.51	73.7	1	0	33

..

SEE pages 11, 12 of file for pictures

Measuring Distance Between Groups

Statistics used to measure treatment-effectiveness are generally of the form

$$\sum_{Gp1} (N_i(T_i) - E(N_i(T_i) \,|\, Z_i^*,\, T_i-)) \;/\; \sqrt{\sum_{Gp1} \mathrm{Var}(N_i(T_i) \,|\, Z_i^*,\, T_i-)}$$

including the most common, the **logrank** statistic in which Z_i^* is not present, sometimes in weighted version with $w(T_i)$ in numerator sum, $w^2(T_i)$ in denominator.

Logrank statistic for same survival in 2 groups. 2×2 Table for Death-time t, $\Delta N(t) = N(t+) - N(t-) > 0$.

Deaths		Totals
$\Delta N^{(1)}(t)$		$Y^{(1)}(t)$ $Gp.\ 1$ at-risk
		$Y^{(2)}(t)$ $Gp.\ 1$ at-risk
$\Delta N(t)$	$Y(t+)$	$Y(t)$ at-risk

$$E(\Delta N^{(1)}(t) \,|\, \text{totals},\ \Delta N(t) > 0) \;=\; \Delta N(t)\, \frac{Y^{(1)}(t)}{Y(t)}$$

Hypergeometric under same Survival in 2 Gps

Example. PBC data. Stat = .685 with no Z_i^* predictors, and 2.04 with all variables other than TRTGP.

References

Probability Theory: general sources on martingales

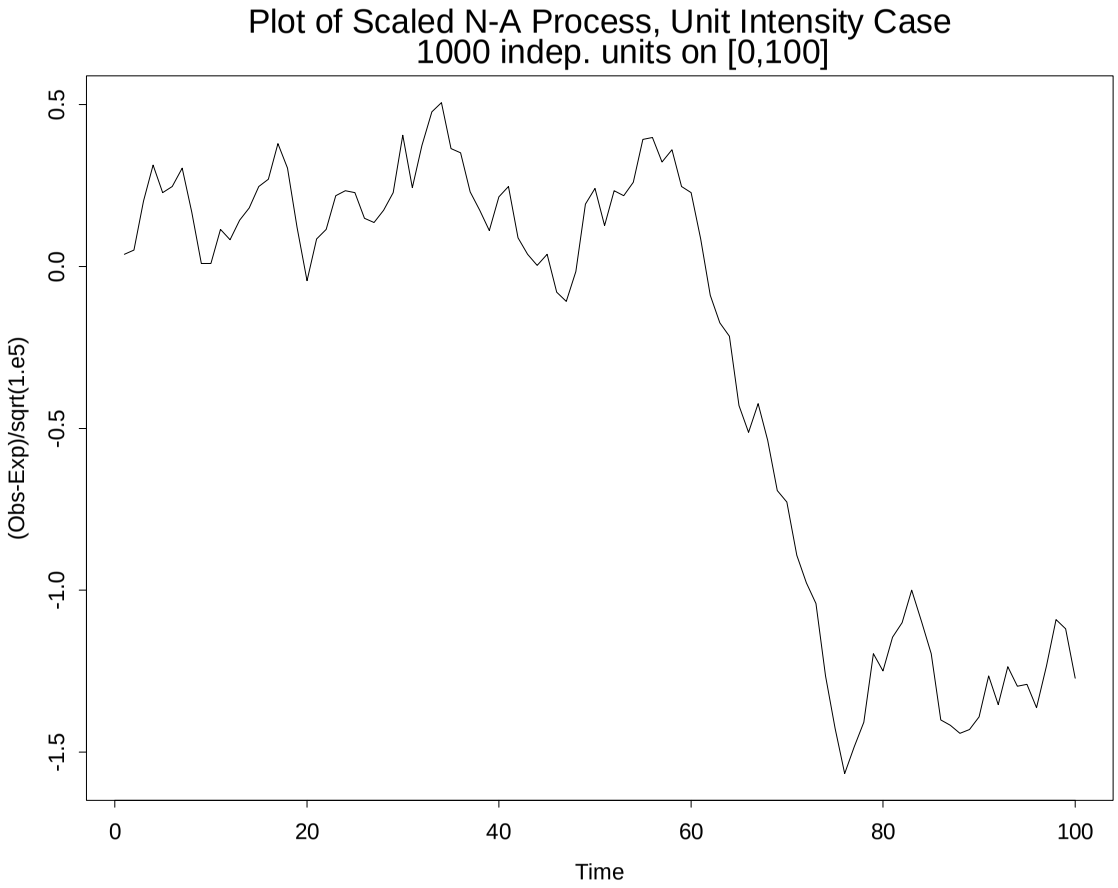
books by M. Loeve; P. Billingsley, ...
Brémaud, P. (1980) **Point Processes and Queues.**
Williams, D. (1991) **Probability with Martingales.**

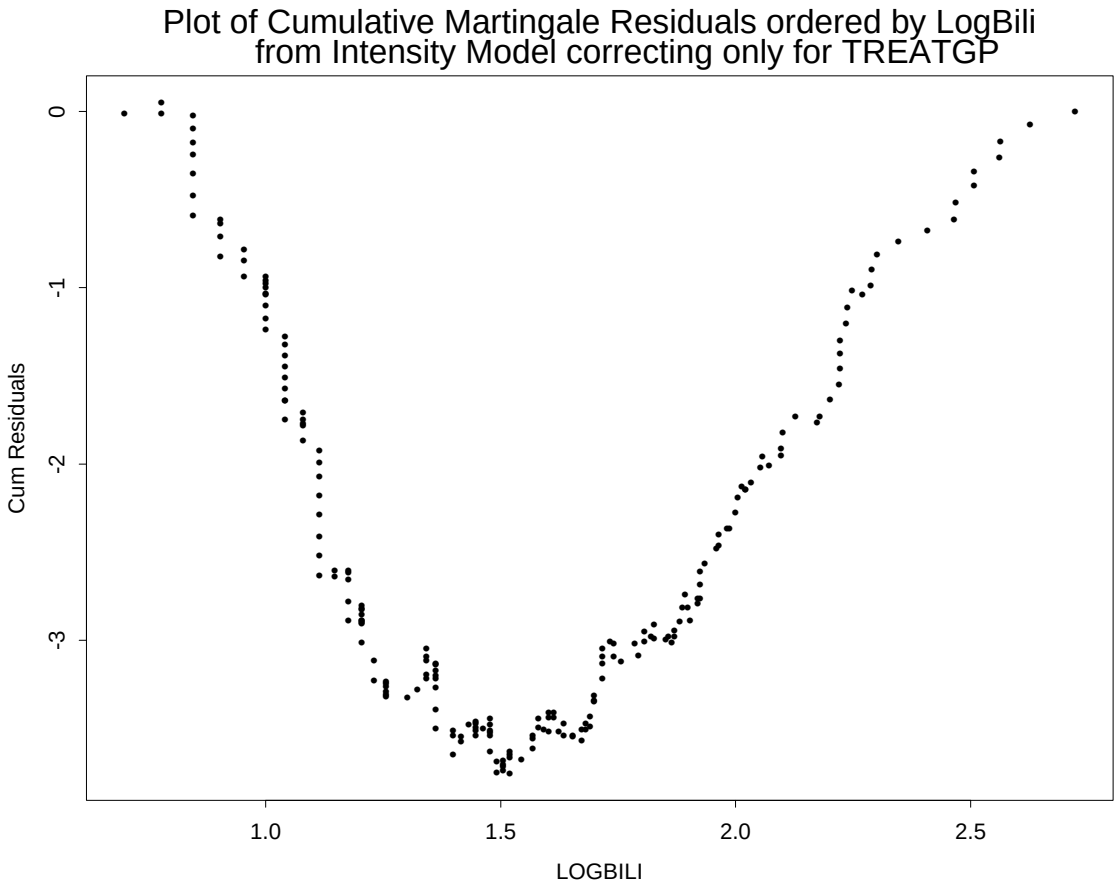
Statistics: material on survival-data regression models

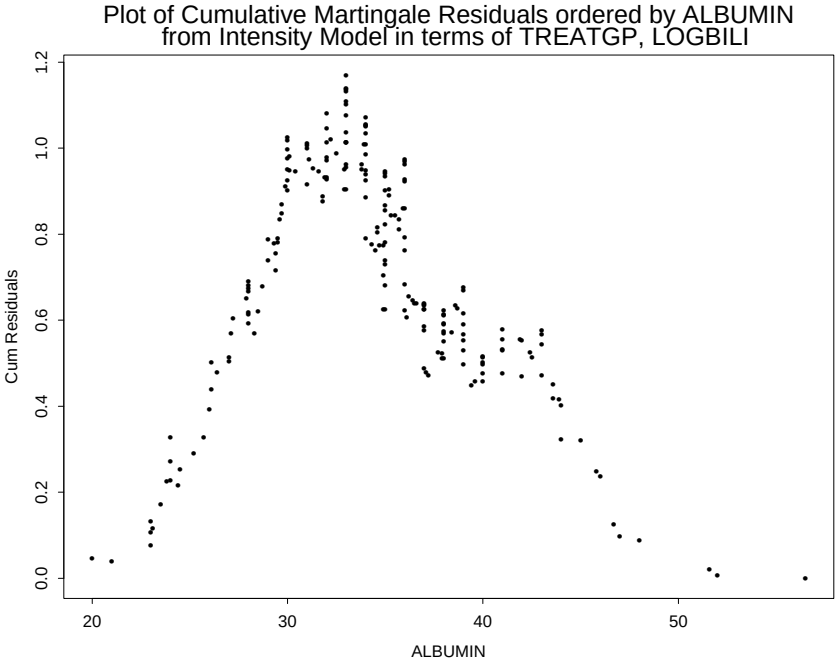
Aalen, O. (1978) *Ann. Statist.*
Andersen, Borgan, Gill, and Keiding (1993) **Statistical Models based on Counting Processes**
Cox, D.R. (1972) *Regression models and life tables*, *Jour. Roy. Statist. Soc.* **B**
R. Miller **Survival Analysis** (1980)

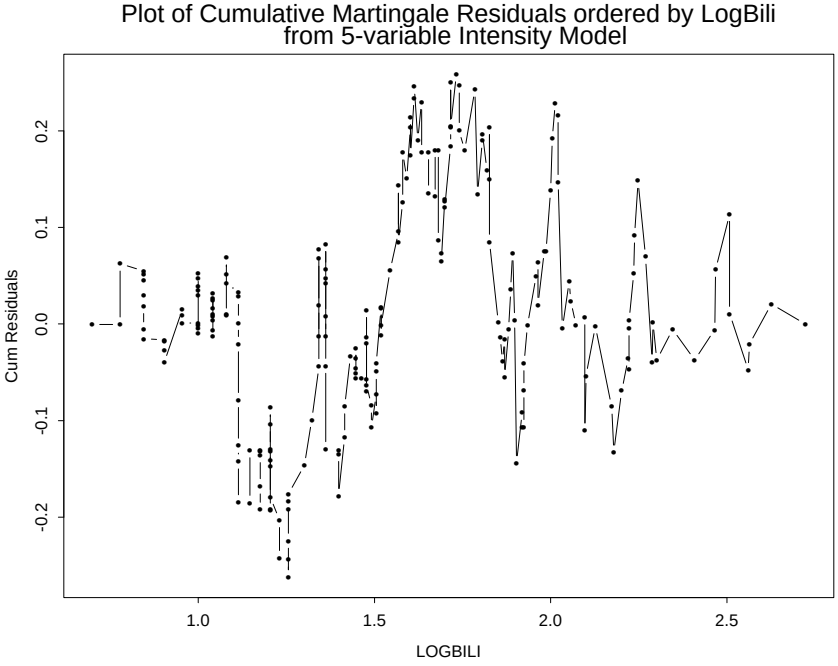
Time Series: approach via prospective regression models and martingales

Kedem, B. & Fokianos (2002) **Regression models for time series analysis.**
Slud, E. and Kedem, B. (1994) *Statistica Sinica*


Plot of Scaled N-A Process, Unit Intensity Case
1000 indep. units on [0,100]
(Obs-Exp)/sqrt(1.e5)
0.5
0.0
-0.5
-1.0
-1.5
0
20
40
60
80
100
Time


Plot of Cumulative Martingale Residuals ordered by LogBili
from Intensity Model correcting only for TREATGP
Cum Residuals
0
-1
-2
-3
1.0
1.5
2.0
2.5
LOGBILI


Plot of Cumulative Martingale Residuals ordered by ALBUMIN
from Intensity Model in terms of TREATGP, LOGBILI
Cum Residuals
0.0
0.2
0.4
0.6
0.8
1.0
1.2
20
30
40
50
ALBUMIN


Plot of Cumulative Martingale Residuals ordered by LogBili
from 5-variable Intensity Model
Cum Residuals
-0.2
-0.1
0.0
0.1
0.2
1.0
1.5
2.0
2.5
LOGBILI