Efficiency Gains by Exploiting Recurrences in Event-Time Analysis

Edsel A. Peña

(E-Mail: pena@stat.sc.edu)

Department of Statistics, University of South Carolina

Columbia, SC 29208

Some portions joint work with **A. Adekpedjou** and **J. Quiton**

Research Support from US NIH Grants

Presented August 2008 at JSM, Denver, CO

Outline of Talk

- Historical background.
- Random censorship model; noninformative censoring; informative censoring; KG Model.
- Migratory motor complex data set.
- Recurrent events.
- First-event analysis and marginal models.
- Full modeling and estimators in basic model.
- Informative monitoring with recurrent event.
- Efficiency issues with recurrent events under generalized KG model.
- Concluding remarks.

Random Censorship Model (RCM)

$$T_1, T_2, \ldots, T_n \stackrel{IID}{\sim} F$$

$$C_1, C_2, \ldots, C_n \stackrel{IID}{\sim} G$$

F and G not related

$$\{T_i\} \perp \{C_i\}$$

Random Observables:

$$(Z_1, \delta_1), (Z_2, \delta_2), \ldots, (Z_n, \delta_n)$$

$$Z_i = T_i \wedge C_i \quad \text{and} \quad \delta_i = I\{T_i \leq C_i\}$$

Goal: To make inference on the distribution F or the hazard $\Lambda = \int dF/F_-$.

Parametric Inference

$$F \in \mathcal{F} = \{F(\cdot;\theta) : \ \theta \in \Theta \subset \Re^p\}$$

$$\bar{F} = 1 - F$$

$$L(\theta) = L(\theta|(\mathbf{Z},\delta)) \propto \prod_i f(z_i;\theta)^{\delta_i}\bar{F}(z_i;\theta)^{1-\delta_i}$$

$$\hat{\theta} = \arg\min_{\theta} L(\theta) \quad \text{(ML Estimator)}$$

Properties of the MLE $\hat{\theta}$ well-known, e.g., consistency (though is usually biased), asymptotic normality, etc.

$$\hat{\theta} \sim AN\left(\theta, \frac{1}{n}[\mathcal{I}_1(\theta, G)]^{-1}\right)$$

Nonparametric Inference

$F \in \mathcal{F} =$ space of continuous distributions

$$N(s) = \sum_i I\{Z_i \le s; \delta_i = 1\} \quad \text{and} \quad Y(s) = \sum_i I\{Z_i \ge s\}$$

NAE: $\hat{\Lambda} = \int \frac{dN}{Y}$ and **PLE:** $\hat{\bar{F}} = \prod \left[1 - \frac{dN}{Y}\right]$

Properties of $\hat{\Lambda}$ and $\hat{F}$ well-known, e.g., biased; consistent; and when normalized, weakly convergent to Gaussian processes.

$$\text{Avar}(\hat{\bar{F}}(t)) = \frac{1}{n}\bar{F}(t)^2 \int_0^t \frac{dF(s)}{\bar{F}(s)^2 \bar{G}(s)}$$

An Informative RCM

- Koziol-Green (KG) Model (1976):

$$\exists \beta \geq 0, \quad \bar{G}(t) = \bar{F}(t)^{\beta}$$

- Lehmann-type alternatives
- Proportional hazards:

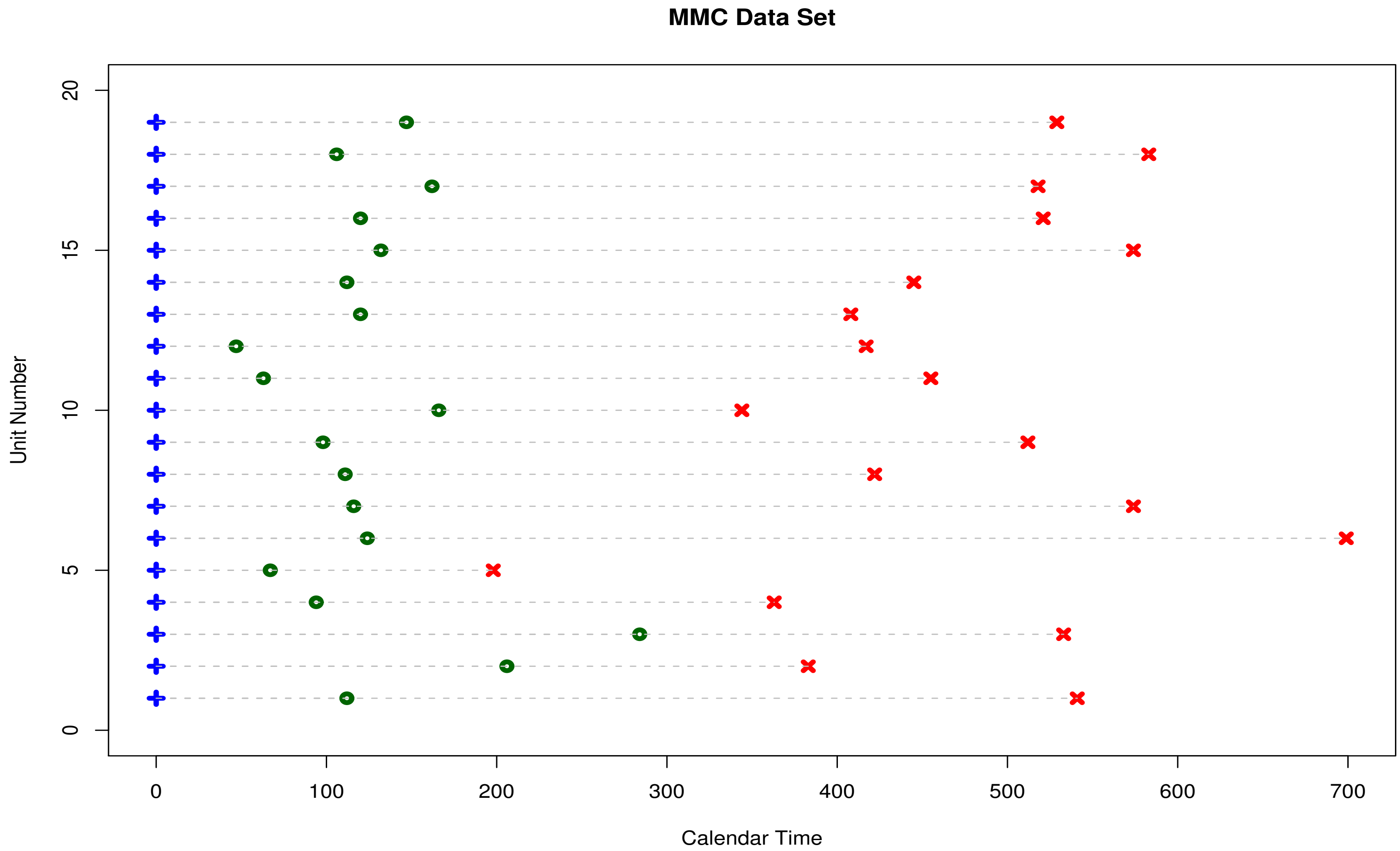
$$\Lambda_G = \beta \Lambda_F$$

- $Z_i \sim \bar{F}^{\beta+1}$
- $\delta_i \sim \text{BER}(1/(\beta+1))$
- Most importantly:

$$Z_i \perp \delta_i$$

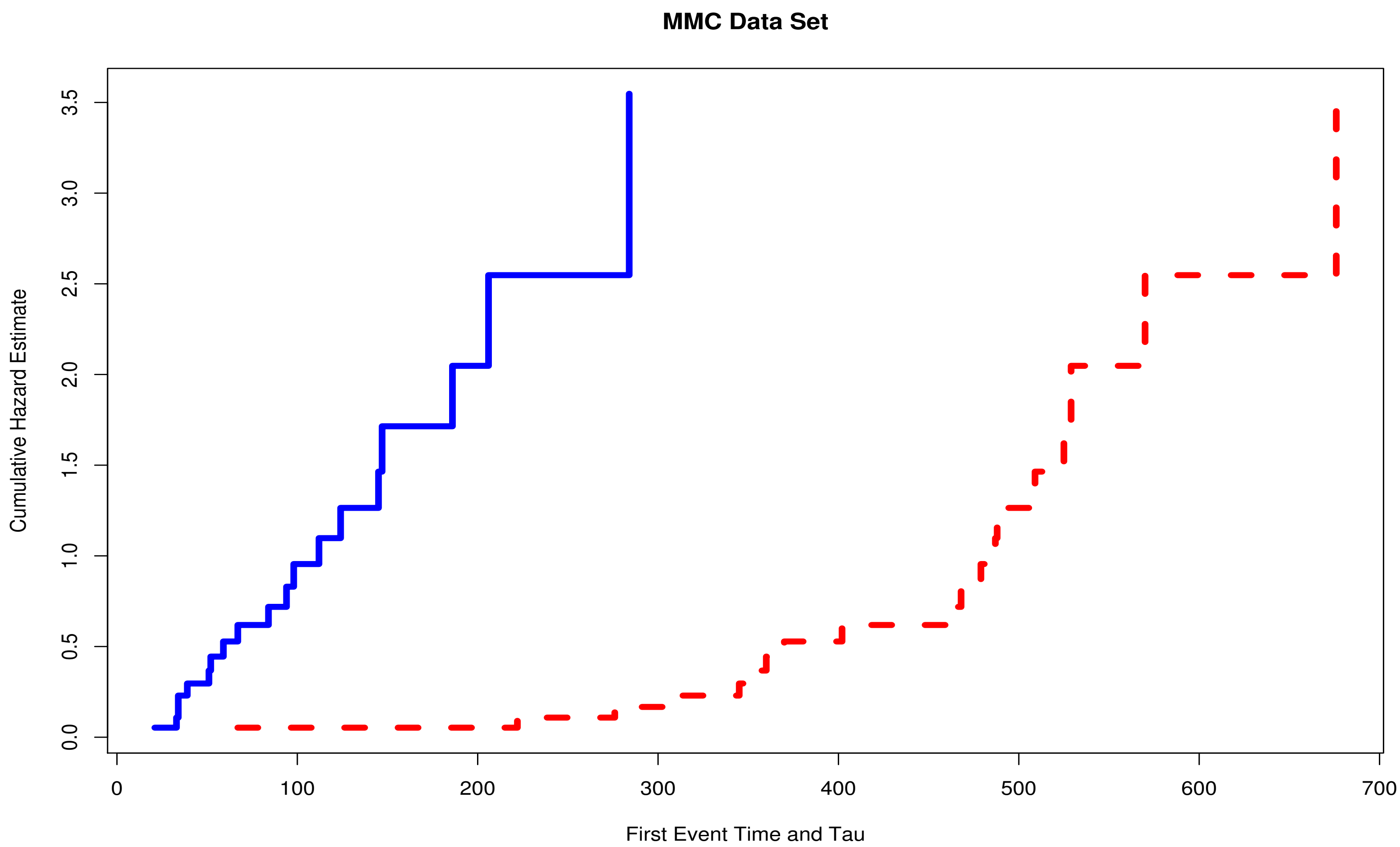
MMC Data: First Event Times and Tau's

Migratory Motor Complex (MMC) data set from Aalen and Husebye, Stat Med, 1991.



MMC Data: Hazard Estimates

KG model holds if and only if $\Lambda_\tau \propto \Lambda_{T_1}$



KG Model's Utility

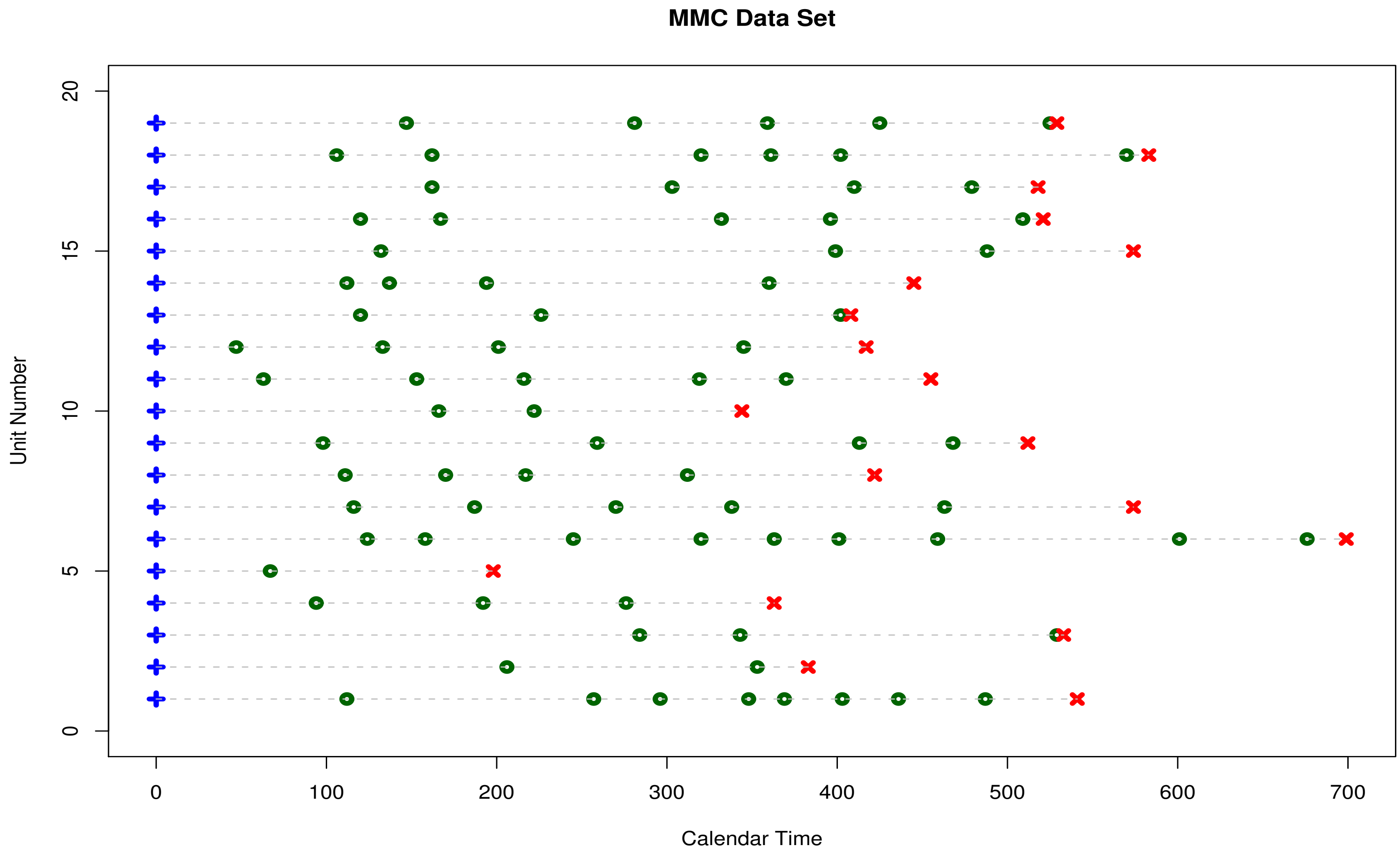
- Chen, Hollander and Langberg (JASA, 1982): exploited independence between Z_i and δ_i to obtain the exact bias, variance, and MSE functions of PLE. Comparisons with asymptotic results.
- Cheng and Lin (1987): exploited semiparametric nature of KG model to obtain a more efficient estimator of F compared to the PLE.
- Hollander and Peña (1988): obtained better confidence bands for F under KG model.
- Csorgo & Faraway: KG model not practically viable, but, just like yeast (!), it allows examination of exact properties of procedures and assessment of efficiency losses/gains.

Recurrent Events

- admission to hospital due to chronic disease
- tumor re-occurrence
- migraine attacks
- alcohol or drug addiction
- commission of a criminal act by a deliquent minor!
- major disagreements between a couple
- non-life insurance claim
- drop of ≥ 200 points in DJIA during trading day
- publication of a paper

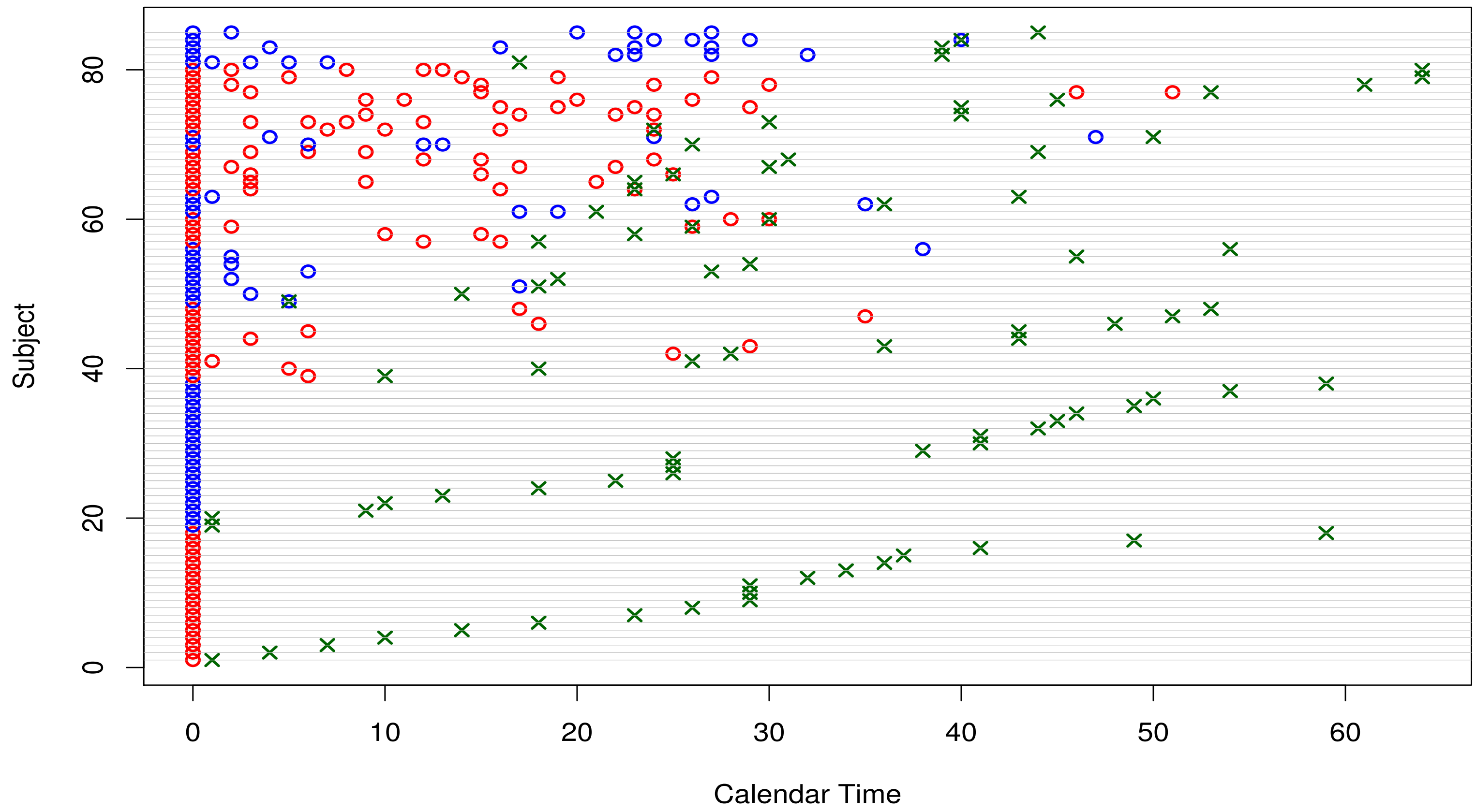
Full MMC Data: With Recurrences

$n = 19$ subjects; event = end of migratory motor complex cycle; random length of monitoring period per subject.

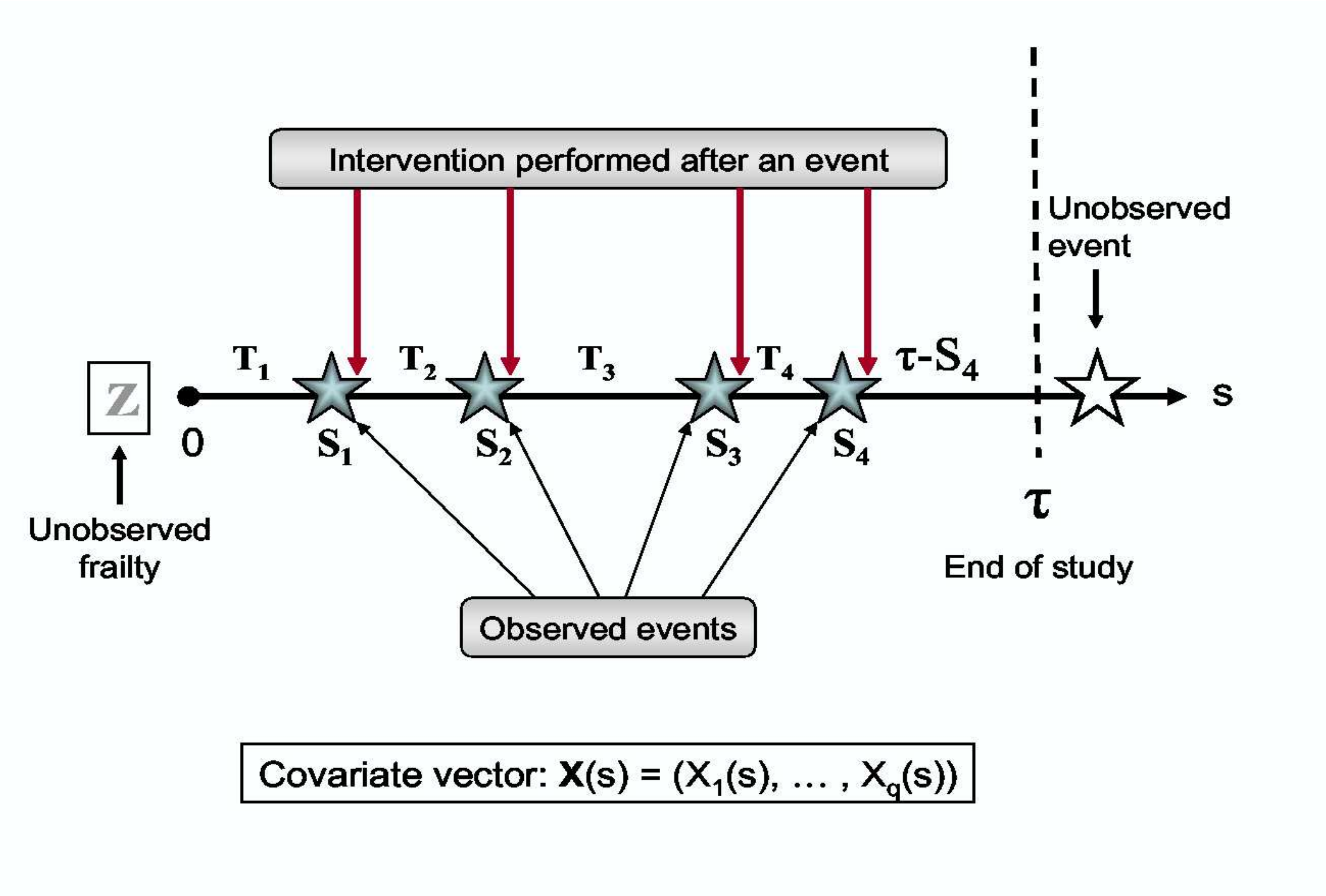


Another One: Bladder Data Set

A ‘famous’ data set used by Wei, Lin, and Weissfeld (1989). 85 subjects; two treatments (control & thiotepa); occurrence of bladder cancer; two covariates.



Data Accrual: One Subject



Some Aspects in Recurrent Data

- random monitoring length (τ).
- random # of events (K) and sum-quota constraint:

$$K = \max\left\{k : \sum_{j=1}^{k} T_j \leq \tau\right\} \text{ with } \sum_{j=1}^{K} T_j \leq \tau < \sum_{j=1}^{K+1} T_j$$

- **Basic Observable:** $(K, \tau, T_1, T_2, \ldots, T_K, \tau - S_K)$
- always a right-censored observation.
- dependent censoring.
- informative censoring.
- effects of covariates, frailties, interventions after each event, and accumulation of events.

Approaches: Recurrent Event Analysis

- First-event analysis. Inefficient. How much do we gain by utilizing the recurrences?
- Full modeling approach: Andersen and Gill (82) and in Peña and Hollander (04) and Peña, Slate and Gonzalez (07).
- Full modeling: harder to implement and requires intensity process specification.
- Marginal modeling approach: Wei, Lin, and Weissfeld (89).
- Conditional modeling approach: Prentice, Williams, and Peterson (81).
- Marginal and conditional approaches: quite popular, but some foundational questions.

Simplest Model: One Subject

- $T_1, T_2, \ldots \overset{IID}{\sim} F$
- corresponds to 'perfect interventions' after each event
- $\tau \sim G$
- F and G not related
- no covariates (X)
- no frailties (Z)
- F could be parametrically or nonparametrically specified.
- Simple model dealt with in Peña, Strawderman and Hollander (JASA, 01); nonparametric estimation of F.

Nonparametric Estimation of F

$$N(t) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} I\{T_{ij} \le t\}$$

$$Y(t) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K_i} I\{T_{ij} \ge t\} + I\{\tau_i - S_{iK_i} \ge t\} \right\}$$

$$\textbf{GNAE}: \quad \hat{\Lambda}(t) = \int_0^t \frac{dN(w)}{Y(w)}$$

$$\textbf{GPLE}: \quad \hat{\bar{\bar{F}}}(t) = \prod_{0}^{t} \left[1 - \frac{dN(w)}{Y(w)} \right]$$

Main Asymptotic Result

k**th Convolution:** $F^{\star(k)}(t) = \Pr\{\sum_{j=1}^{k} T_j \le t\}$

Renewal Function: $\rho(t) = \sum_{k=1}^{\infty} F^{\star(k)}(t)$

$$\nu(t) = \frac{1}{\bar{G}(t)} \int_t^{\infty} \rho(w-t)dG(w)$$

$$\sigma^2(t) = \bar{F}(t)^2 \int_0^t \frac{dF(w)}{\bar{F}(w)^2\bar{G}(w)[1+\nu(w)]}$$

Theorem (JASA, 01): $\sqrt{n}(\hat{\bar{F}}(t) - F(t)) \Rightarrow \mathrm{GP}(0, \sigma^2(t))$

Efficiency Gain: Recurrent vs Single Event

- Asymptotic relative efficiency of the estimator of F using recurrences relative to the estimator of F from a first-event analysis:

$$\text{ARE(RecuEst:SingEst)} = \left(\int_0^t \frac{dF}{\bar{F}^2\bar{G}}\right)\left(\int_0^t \frac{dF}{\bar{F}^2\bar{G}[1+\nu]}\right)^{-1}$$

- When $F = \text{EXP}(\theta)$ and $G = \text{EXP}(\lambda)$, the ARE expression reduces to:

$$\text{ARE(RecuEst:SingEst)} = 1 + \frac{\theta}{\lambda} = 1 + \frac{E(\tau)}{E(T)}$$

- Gain in efficiency is, in hindsight, somewhat intuitive!

Extending KG Model: Recurrent Setting

- Wanted a tractable model with monitoring time informative about F.
- Potential for a refined analysis of efficiency gains or losses.
- Idea: Why not simply build or generalize the KG model for the RCM.
- Generalized KG model for Recurrent Events:

$$\exists \beta \geq 0, \quad \bar{G}(t) = \bar{F}(t)^{\beta}$$

with β unknown, and where F is the common inter-event distribution.

In light of this GKG model we ask ...

- What is the degree of efficiency loss when the informative monitoring model structure is ignored?
- How much is the penalty when using a Single-event analysis relative to Recurrent-event analysis.
- In a study is it better to increase the number of subjects with the monitoring time distribution remaining the same, or is it better to lengthen the monitoring times but fixing the number of subjects?
- If one uses the nonparametric estimator in PSH (01), how does it compare to the estimator that exploits the informative monitoring structure?

Ignoring Informative Monitoring

- If GKG model holds, could only do better exploiting informative monitoring.
- When $F = \text{EXP}(\theta)$, no lose of efficiency by ignoring the generalized KG structure!
- In this case, estimators of θ concide:

$$\hat{\theta} = \tilde{\theta} = \frac{\sum_{i=1}^{n} K_i}{\sum_{i=1}^{n} \tau_i}$$

  the occurrence-exposure rate.
- Other F leads to loss when informative monitoring is ignored.

Single versus Recurrent Analyses

- $\check{\theta}$: estimator based only on, possibly right-censored, first event time. *Single-Event Analysis*
- $\hat{\theta}$: estimator derived under GKG and uses the event recurrences. *Recurrent-Event Analysis*
- General result is $\hat{\theta}$ is never inefficient compared to $\check{\theta}$.
- When $F = \text{EXP}(\theta)$,

$$\Delta\text{ARE}(\hat{\theta} : \check{\theta}) = \frac{1}{\beta}$$

- Interpretations & Implications.

Simulated Efficiencies: Two-Parameter Weibull

n	θ_1	θ_2	β	MeanEvs	Eff$(\hat{\theta} : \tilde{\theta})$	Eff$(\hat{\theta} : \check{\theta})$
20	0.9	1	0.3	3.89	1.26	37.62
20	0.9	1	0.5	2.25	1.47	17.12
20	0.9	1	0.7	1.57	1.72	9.88
20	1.0	1	0.3	3.34	1.25	30.24
20	1.0	1	0.5	2.01	1.51	12.62
20	1.0	1	0.7	1.43	1.82	8.69
20	1.5	1	0.3	1.98	1.54	11.11
20	1.5	1	0.5	1.34	1.90	6.83
20	1.5	1	0.7	1.02	2.12	4.38

Nonparametric Estimator in PSH

- Parametric Model:

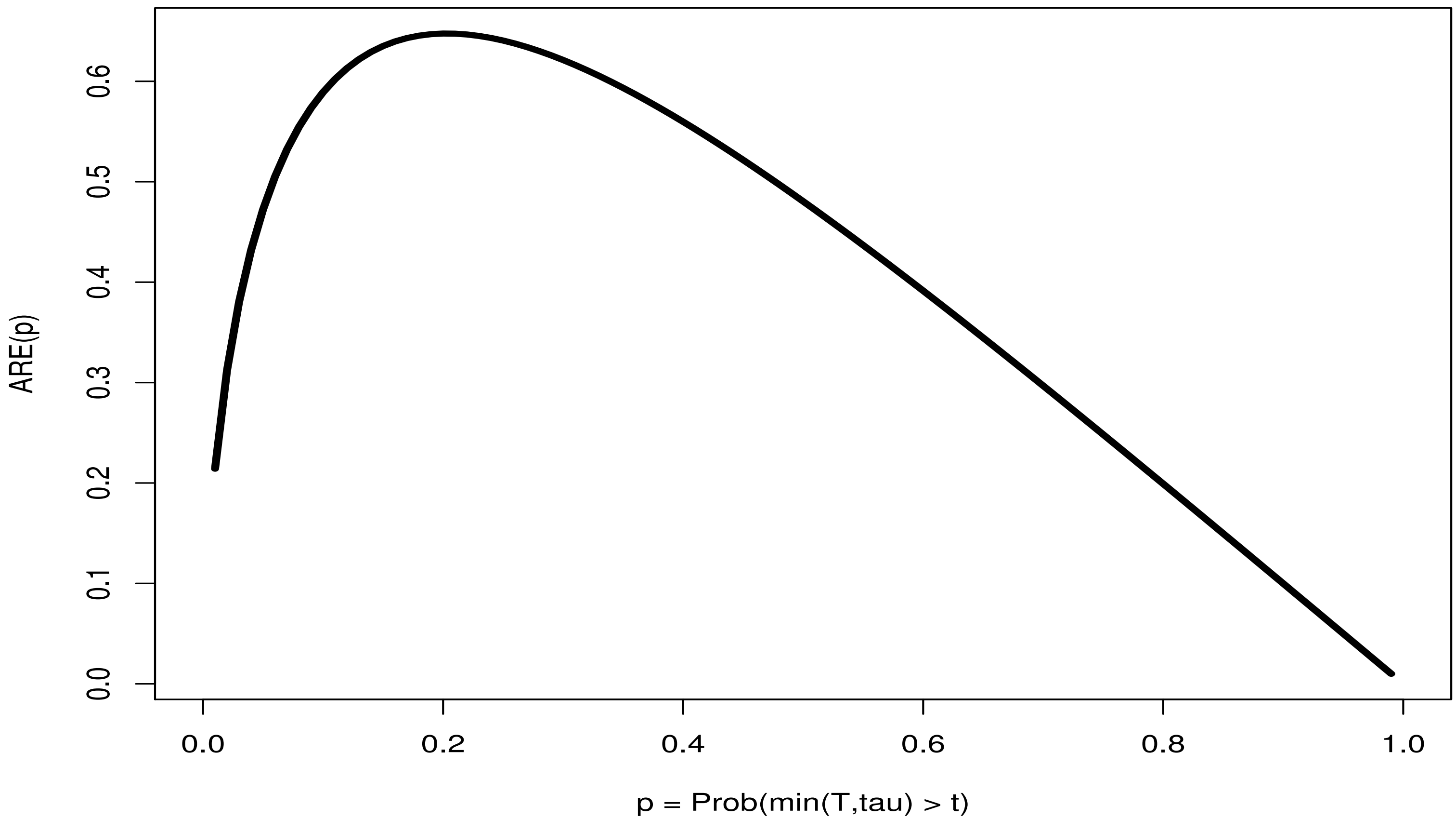
$$F \in \mathcal{F} = \{F(\cdot;\theta) : \ \theta \in \Theta\}$$

- $\hat{\bar{F}}(t)$ is the parametric-based estimator exploiting informative monitoring.
- $\tilde{\bar{F}}(t)$ is the nonparametric estimator from PSH (01). It does not exploit informative monitoring.
- When $\bar{F}(t;\theta) = \exp(-\theta t)$, an exact expression of ARE is obtained:

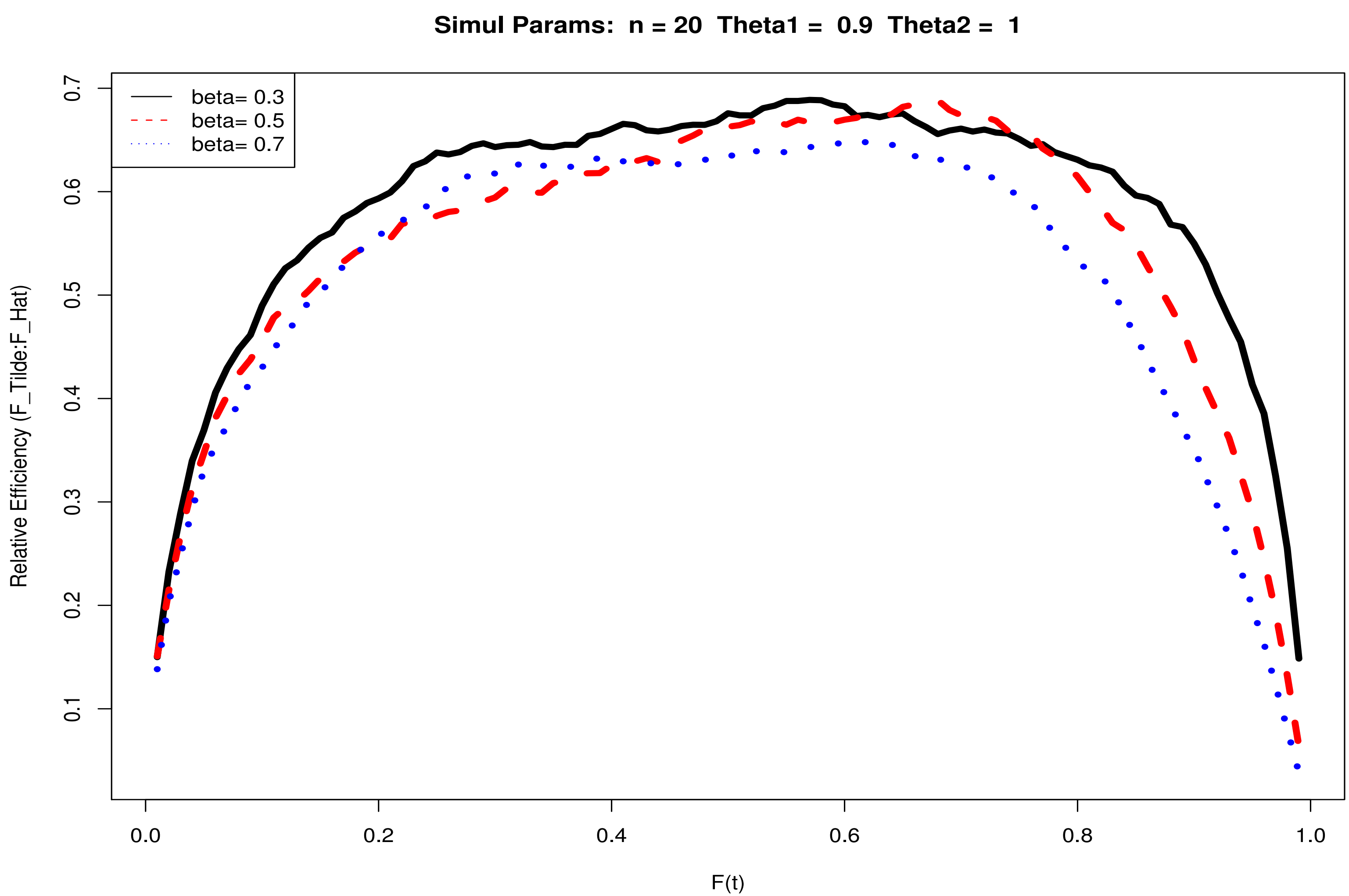
$$\text{ARE}(\tilde{F} : \hat{F}) = \frac{[(1+\beta)t]^2}{\exp[(1+\beta)t] - 1} = \frac{p(t;\beta)}{1 - p(t;\beta)}[-\log p(t;\beta))]^2$$

Efficiency Plot: Under F = EXP(θ)

Plot of ARE($\tilde{F} : \hat{F}$) as function of $p(t;\beta) = \Pr(T_1 \wedge \tau > t)$.



Simulation Results: Two-Parameter Weibull



Concluding Remarks

- Single-event and recurrent-event analysis reviewed, together with informative censoring models.
- KG model for the random censorship model revisited, and then extended to the recurrent event setting.
- Efficiency gains and losses examined when using single-event analysis and when informative monitoring structure is ignored.
- Behooves to gather recurrent event data, if feasible. Significant gain in efficiency could be achieved.
- Gain in efficiency translates to better decision making, especially in health-related matters.