
Analysis for recurrent event data with informative censoring

Chiung-Yu Huang

Biostatistics Research Branch

National Institute of Allergy and Infectious Diseases

IMS New Researchers Conference

July 30, 2009

Recurrent Event Data

Examples:

- Recurrent opportunistic infections of HIV patients
- Repeated hospitalizations, emergency room visits

Advantages:

- More informative about underlying medical conditions
- Assess disease progression

Modeling Recurrent Event Data

Two time scales:

- Time since the last event

Focus on distribution of the gap time between consecutive events.

Huang, Luo and Follmann (In preparation)

- Time since entering the study

Conventionally modeled as realizations of counting processes.

Huang, Qin and Wang (Bmcs, to appear)

Gap Time Models

Gap Time Data

Tricky data structure

- Gap times are ordered and correlated
- Number of recurrent events is informative.
- Length bias arises due to intercept sampling
- Induced informative censoring

Prentice, Williams, and Peterson (Bmka 1981), Wang and Chang (JASA 1999), Lin, Sun and Wei (Bmka 1999), Huang (JRSSB 2002)

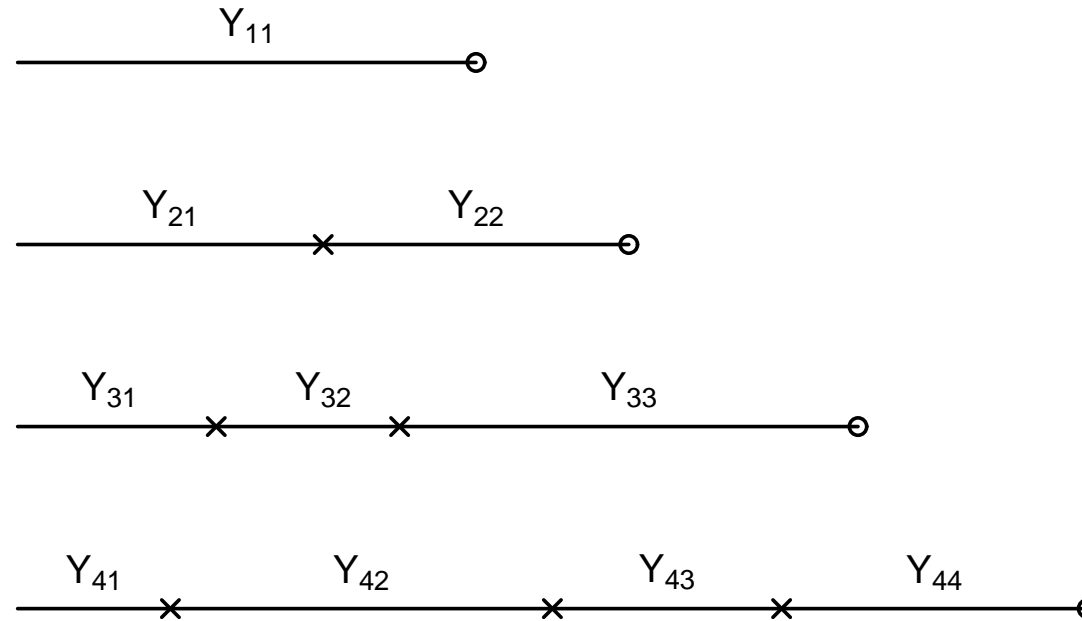


Recurrent Gap Time Data

Huang, Luo and Follmann (2009)

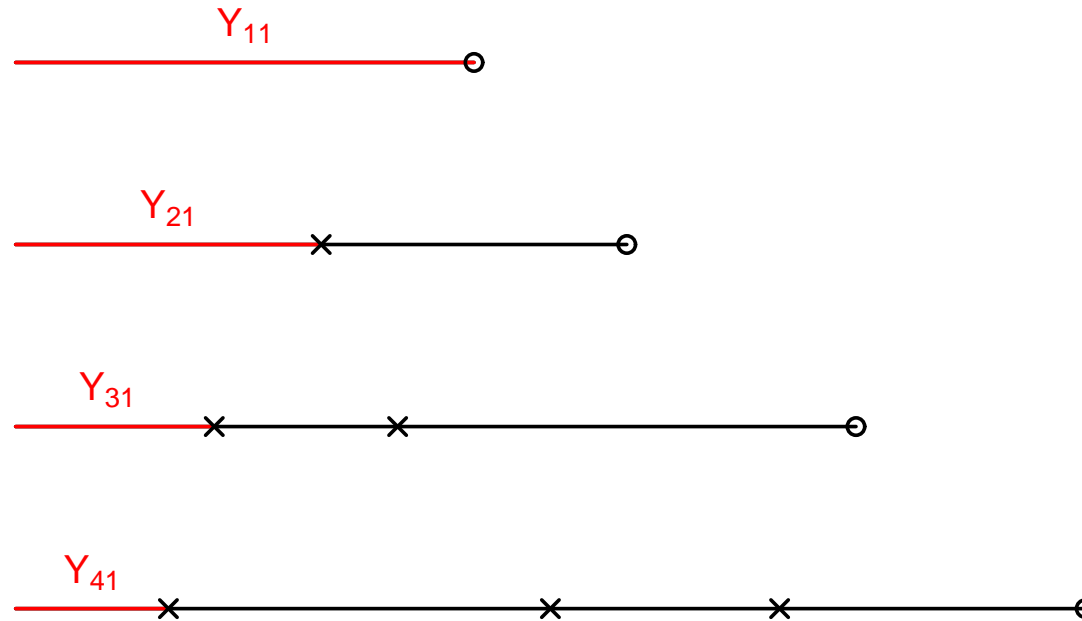
- Recurrent events of the same type; covariates are time-independent.
- Data are assumed to be a heterogeneous mix of independent renewal processes subject to random censoring.
- Uncensored gap times are exchangeable.

Recurrent Gap Time Data



Estimate the marginal distribution of gap times.

Time-to-First-Event Analysis

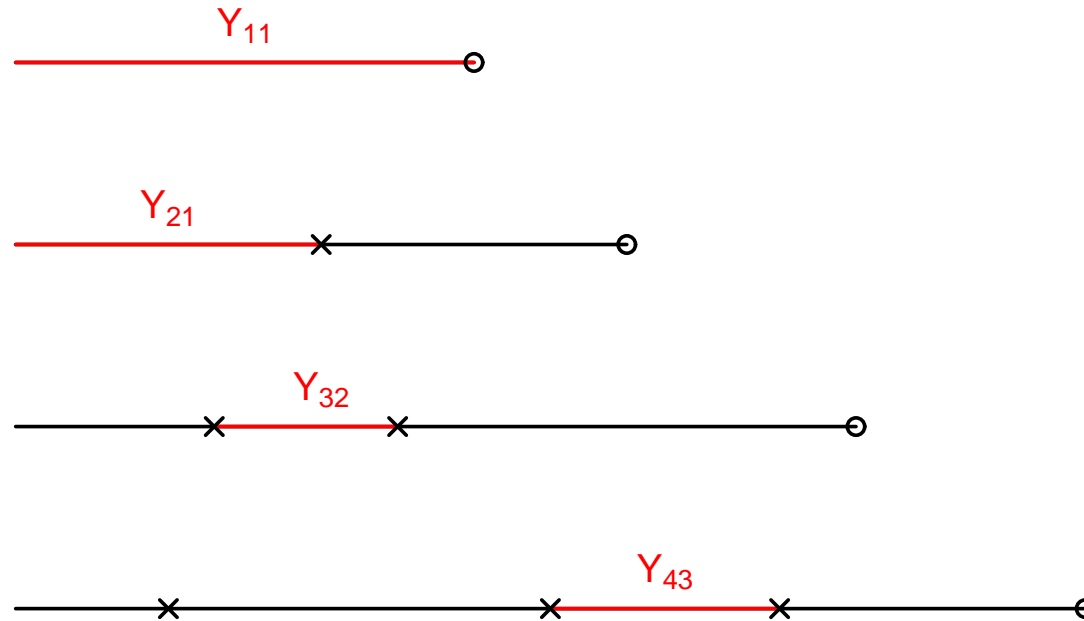


Time to first event:

$$(Y_{11}, \Delta_1 = 0), (Y_{21}, \Delta_2 = 1), (Y_{31}, \Delta_3 = 1), (Y_{41}, \Delta_4 = 1), \dots$$

Apply Kaplan-Meier estimator, Cox PHM, AFT model, etc.

Within-Cluster Resampling Method



One gap time is randomly selected from each subject to form a subsample of independent observations

$$(Y_{11}, \Delta_1 = 0), (Y_{21}, \Delta_2 = 1), (Y_{32}, \Delta_3 = 1), (Y_{43}, \Delta_4 = 1), \dots$$

Apply Kaplan-Meier estimator, Cox PHM, AFT model, etc.

Within-Cluster Resampling Method

- For each reseampling:
 - Y_{i1} is always chosen if subject i has only one censored gap time
 - Only uncensored gap times are selected if subject has ≥ 1 events
 - The subsample consists of independent observations. Standard software can be readily applied.
- The WCR estimator can be obtained through averaging over the estimates obtained from the resampled data.

$$\hat{\beta}^{wcr} = B^{-1} \sum_{b=1}^B \hat{\beta}_b$$

$$\hat{\Sigma}^{wcr} = \frac{1}{B} \sum_{b=1}^B \hat{\Sigma}_b - \frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}_b - \hat{\beta}^{wcr})^{\otimes 2},$$

Hoffman, Sen and Weinberg (Bmka 2001)

Within-Cluster Resampling Method

- WCR estimators are approximations to existing methods
 - Kaplan-Meier + WCR \Rightarrow Wang and Chang (JASA 1999)
 - Partial likelihood method + WCR \Rightarrow Huang and Chen (LIDA 2003)
- In fact, WCR method can be used to extend risk-set methods for survival analysis to analyze recurrent gap time data.
 - Assign each observation in a risk set a weight that is proportional to the inverse of the "effective" cluster size.
 - AFT model; Weighted logrank test
 - Model checking techniques using "martingale-like" residuals.

Luo (JSM Session #425)

Counting Process Models

Statistical Models for Recurrent Event Data

- Counting Process Models
 - Time since entering the study
 - More traditional; more references

Andersen and Gill (Ann Stat 1982); Prentice, Williams and Peterson (Bmka 1981); Nelson (JQT 1988); Lawless and Nadeau (95, Technmcs); Lin, Wei, Yang and Ying (JRSSB 2000); Wang, Qin, and Chiang (2001, JASA)

Challenge of Informative Censoring

- Two types of informative censoring:
 - Informative dropouts due to serious side effects;
 - Terminal events (e.g. death).

- The collection of patients under observation is not a representative sample of the study population; hence risk-set methods are not valid

Notations

Goal: Make inference about the recurrent event process on a pre-specified time interval $[0, \tau]$.

- Y – censoring time (including failure events and dropouts)
- m – number of observed events
- t_1, t_2, \dots, t_m – event times
- $N(t)$ – number of recurrent events up to t
- W – a vector of time-independent covariates (race, sex, etc)
- $X(t)$ – a vector of time-dependent covariates (CD4 count, viral load, HIV status, etc)
- $\mathcal{X}(t) = \{X(u) : 0 \leq u \leq t\}$: covariate history up to t

The Proposed Model

There exists a r.v. Z with $E[Z|\mathcal{X}(\tau), W] = \mu_z$ such that

- Given $(Z, \mathcal{X}(\tau), W)$, $N(\cdot)$ is a NHPP with intensity

$$\lambda(t) = Z\lambda_0(t)e^{X(t)\beta+W\gamma}$$

where $\lambda_0(t)$ is an arbitrary function with $\Lambda_0(\tau) = \int_0^\tau \lambda_0(u)du = 1$

- Given $(Z, \mathcal{X}(\tau), W)$, Y is independent of $N(\cdot)$

The Proposed Model

There exists a r.v. Z with $E[Z|\mathcal{X}(\tau), W] = \mu_z$ such that

- Given $(Z, \mathcal{X}(\tau), W)$, $N(\cdot)$ is a NHPP with intensity

$$\lambda(t) = Z\lambda_0(t)e^{X(t)\beta+W\gamma}$$

where $\lambda_0(t)$ is an arbitrary function with $\Lambda_0(\tau) = \int_0^\tau \lambda_0(u)du = 1$

- Given $(Z, \mathcal{X}(\tau), W)$, Y is independent of $N(\cdot)$
 - No parametric assumptions about Z
 - Y can be arbitrarily correlated with $N(\cdot)$ through $(Z, \mathcal{X}(\tau), W)$
 - The rate function is given by $\mu_z\lambda_0(t)e^{X(t)\beta+W\gamma}$
 - Poisson assumption can be relaxed
 - Reduces to Wang, Qin, and Chiang (JASA 2001) if $\beta \equiv 0$

Keys to the Inference Procedure

- Maximum likelihood estimation is not available because the distribution of Z is not specified

- Conditional likelihood method

Under Poisson assumption, $\{t_1, \dots, t_m\}$ conditioned on (Y, m) are order statistics of **iid** r.v.'s with density

$$\frac{\lambda(t)}{\int \lambda(u) I(0 \leq u \leq Y) du}$$

It has the form of a truncated density function!

Keys to the Inference Procedure (Conti.)

- Pairwise likelihood method (Liang and Qin JRSS B 2000)
- Consider joint density $f(y, x; \beta, \phi) = f(y | x, \beta)f(x; \phi)$
- Condition on the order statistic of paired outcomes (Y_1, Y_2) to eliminate nuisance parameter ϕ

$$\frac{f(y_1, x_1; \beta, \phi)f(y_2, x_2; \beta, \phi)}{f(y_1, x_1; \beta, \phi)f(y_2, x_2; \beta, \phi) + f(y_2, x_1; \beta, \phi)f(y_1, x_2; \beta, \phi)}$$
$$= \frac{f(y_1 | x_1, \beta)f(y_2 | x_2, \beta)}{f(y_1 | x_1, \beta)f(y_2 | x_2, \beta) + f(y_2 | x_1, \beta)f(y_1 | x_2, \beta)}$$

Overview of Proposed Estimation Procedure

- Use conditional method to eliminate Z and γ
- Apply pairwise likelihood method to estimate β
- Use a modified product-limit estimator to estimate $\lambda_0(t)$
- Estimate γ and μ_Z by solving estimating equations

Conditional Likelihood Under Proposed Model

- Observed event times $\{t_{i1}, \dots, t_{im_i}\}$ conditioned on $(Y_i, m_i, Z_i, \mathcal{X}_i(\tau), W_i)$ are order statistics of **iid** r.v. with density

$$\frac{Z_i \lambda_0(t) e^{W_i \gamma + X_i(t) \beta}}{\int_0^{Y_i} Z_i \lambda_0(u) e^{W_i \gamma + X_i(u) \beta} du} = \frac{\lambda_0(t) e^{X_i(t) \beta}}{\int_0^{Y_i} \lambda_0(u) e^{X_i(u) \beta} du}, \quad 0 \leq t \leq Y_i$$

- The conditional likelihood involves both λ_0 and β
- Maximization is complicated in both theory and computation

Farrington and Whitaker (JRSS C 2006)

Pairwise likelihood Method

- Reformulate the problem as estimating β from right truncated data $\{t_{ij}\}$ with truncation time Y_i and density function proportional to

$$\lambda_0(t)e^{X_i(t)\beta}$$

- Consider a comparable pair of event times (t_{ij}, t_{kl}) , where t_{ij} and t_{kl} belong to the same observation period $[0, Y_i \wedge Y_k]$

$$t_{ij} \leq Y_i \wedge Y_k \text{ and } t_{kl} \leq Y_i \wedge Y_k$$

- Conditioning on the order statistic of (t_{ij}, t_{kl}) , we have

$$\frac{\lambda_0(t_{ij})\lambda_0(t_{kl})e^{[X_i(t_{ij})+X_k(t_{kl})]\beta}}{\lambda_0(t_{ij})\lambda_0(t_{kl})e^{[X_i(t_{ij})+X_k(t_{kl})]\beta} + \lambda_0(t_{kl})\lambda_0(t_{ij})e^{[X_i(t_{kl})+X_k(t_{ij})]\beta}}$$

The pairwise likelihood depends on β but not λ_0 .

Estimation of β

- Let $\delta_{ijkl} = 1$ if (t_{ij}, t_{kl}) is a comparable pair, $\delta_{ijkl} = 0$ otherwise
- The pairwise likelihood is given by

$$\prod_{i < k} \prod_{j=1}^{m_i} \prod_{l=1}^{m_k} \left(\frac{e^{[X_i(t_{ij}) + X_k(t_{kl})]\beta}}{e^{[X_i(t_{ij}) + X_k(t_{kl})]\beta} + e^{[X_i(t_{kl}) + X_k(t_{ij})]\beta}} \right)^{\delta_{ijkl}}$$

- Maximize to obtain $\hat{\beta}$
- Asymptotic properties of $\hat{\beta}$ can be established by using the central limit theorem for U-statistics

Estimation of $\Lambda_0(t)$

- Recall that the conditional likelihood is computationally equivalent to the semiparametric likelihood of a set of independently right-truncated r.v.'s

$$\mathcal{L}_c \propto \prod_{i=1}^n \prod_{j=1}^{m_i} \frac{\lambda_0(t_{ij}) e^{X_i(t_{ij})\beta}}{\int_0^{Y_i} \lambda_0(u) e^{X_i(u)\beta} du}$$

- If $\beta = 0$ (WQC model), Λ_0 can be estimated by the product-limit estimator

$$\prod_{s_l > t} \left(1 - \frac{d_l}{R_l} \right)$$

- Product-limit estimator yields inconsistent estimate when $\beta \neq 0$

Estimation of $\Lambda_0(t)$

- Estimate Λ_0 with a modified product-limit estimator:

$$\hat{\Lambda}_0(t) = \prod_{s_l > t} \left(1 - \frac{d_l(\hat{\beta})}{R_l(\hat{\beta})} \right),$$

$$d_l(\beta) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{m_i} I(t_{ij} = s_l) e^{-X_i(t_{ij})\beta},$$

and

$$R_l(\beta) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{m_i} I(t_{ij} \leq s_l \leq Y_i) e^{-X_i(t_{ij})\beta}.$$

- Reduces to the usual product-limit estimator if $\beta = 0$

Estimation of γ and μ_Z

- Next, conditioned on $(Z_i, Y_i, W_i, \mathcal{X}_i(Y_i))$, the expected value of m_i is given by

$$E[m_i | Z_i, Y_i, W_i, \mathcal{X}_i(Y_i)] = \int_0^{Y_i} Z_i e^{X_i(u)\beta + W_i\gamma} d\Lambda_0(u)$$

Hence

$$E\left[\frac{m_i}{\int_0^{Y_i} e^{X_i(u)\beta} d\Lambda_0(u)} \mid Y_i, W_i, \mathcal{X}_i(Y_i)\right] = e^{\log \mu_Z + W_i\gamma}.$$

- γ and μ_Z can be estimated by estimation equations with (β, Λ_0) replaced by $(\hat{\beta}, \hat{\Lambda}_0)$
-

Data Example – ALIVE Study

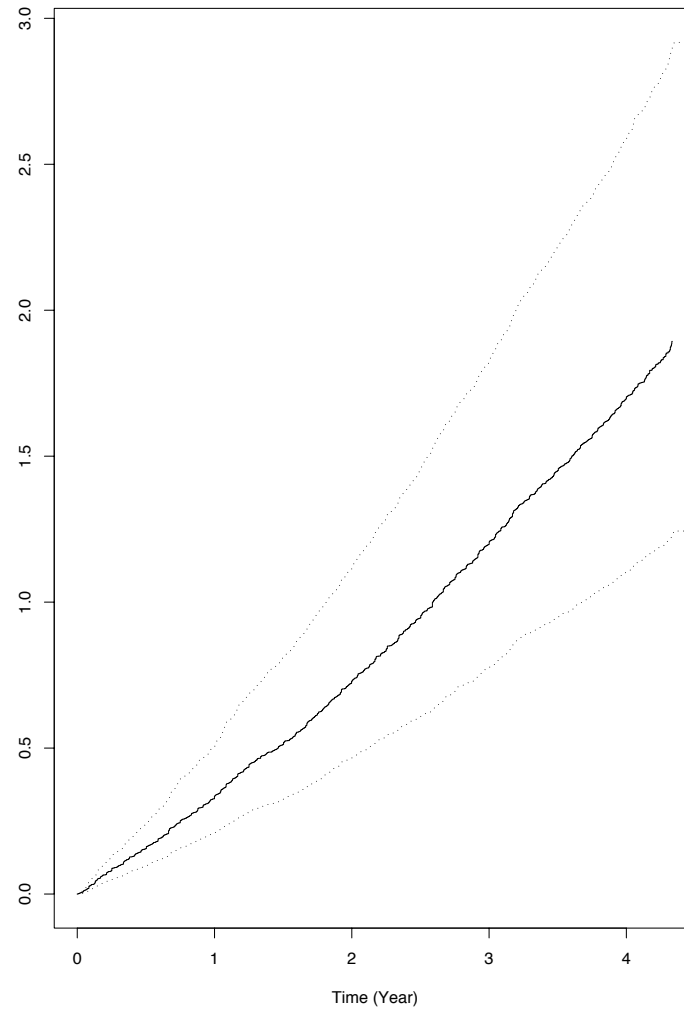
- 2,946 active injection drug users in Baltimore were recruited between Feb 1988 and March 1989. Info on HIV testing was collected every 6 months
- Consider hospital admission records from 1,896 drug users who had ≥ 1 follow-up visit between July 16, 1993 and Dec 31, 1997
 - 46% had ≥ 1 hospital admissions
 - No. of admissions ranges from 0-19, averaging 3.5/subject
 - 13% died before end of study – possibility of informative censoring
- Goal: Evaluate the risk of hospitalization over time

ALIVE study – Results

- Time-independent covariates: gender and race
- Time-dependent covariate: HIV status

	Coef	95% CI
HIV status	0.16	(-0.62, 0.99)
Male	-0.26	(-0.45,-0.08)
Black	-0.06	(-0.06, 0.19)

Estimated Rate Function



Discussions

- $N(\cdot)$ and Y can be correlated through observed covariates (\mathcal{X}, W) as well as unobserved latent variable Z
- Distributions of Z and Y are treated as nuisances
- A test statistic for testing $\beta = 0$ can be given by the score function derived from the pairwise pseudolikelihood

$$\frac{1}{\binom{n}{2}} \sum_{i < k} N_i(Y_{ik}) N_k(Y_{ik}) \int_0^{Y_{ik}} \{X_i(u) - X_k(u)\} \left\{ \frac{dN_i(u)}{N_i(u)} - \frac{dN_k(u)}{N_k(u)} \right\}$$

- does not require information about W
- can be used to check proportional intensity model assumption