Semiparametric Regression Analysis of Interval-Censored Data

Danyu Lin, Ph.D.

Dennis Gillings Distinguished Professor

Department of Biostatistics

University of North Carolina

Chapel Hill, NC 27599-7420

email: lin@bios.unc.edu

website: http://dlin.web.unc.edu/

April 7, 2022

Stata Biostatistics and Epidemiology Virtual Symposium

OUTLINE

Analysis of Right-Censored Data

Analysis of Interval-Censored Data

Analysis of Multivariate Interval-Censored Data

- Multiple events
- Clustered data

Analysis of Right-Censored Data

COX PROPORTIONAL HAZARDS MODEL:

$$\lambda(t|X) \equiv \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(t \leq T < t + \Delta t | T \geq t, X)$$

$$= \lambda_0(t) e^{\beta' X(t)}$$

- T = failure time
- X = (possibly time-dependent) covariates
- $\lambda_0(t) \equiv \lambda(t|Z = 0)$ = arbitrary baseline hazard function
- β = unknown regression parameters
- $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$
- $S(t|X) = \Pr(T > t|X) = \exp\{-\int_0^t e^{\beta' X(s)} d\Lambda_0(s)\}$

RIGHT-CENSORED DATA: $(\widetilde{T}_i, \Delta_i, X_i)$

- C_i = censoring time
- $\widetilde{T}_i = \min(T_i, C_i)$
- $\Delta_i = I(T_i \leq C_i)$
- $I(\cdot)$ = indicator function

NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION (NPMLE)

Likelihood:

$$L(\beta, \Lambda_0) \propto f(\widetilde{T}_i|X_i)^{\Delta_i} S(\widetilde{T}_i|X_i)^{1-\Delta_i} = \lambda(\widetilde{T}_i|X_i)^{\Delta_i} S(\widetilde{T}_i|X_i)$$

$$= \prod_{i=1}^{n} \left\{ e^{\beta' X_i(\widetilde{T}_i)} \lambda_0(\widetilde{T}_i) \right\}^{\Delta_i} \exp\left\{ - \int_0^{\widetilde{T}_i} e^{\beta' X_i(t)} d\Lambda_0(t) \right\}$$

$$\widetilde{L}(\beta, \Lambda_0) = \prod_{i=1}^{n} \left\{ e^{\beta' X_i(\widetilde{T}_i)} \lambda_i \right\}^{\Delta_i} \exp\left\{ - \sum_{j:\widetilde{T}_j \le \widetilde{T}_i} e^{\beta' X_i(\widetilde{T}_j)} \lambda_j \right\}$$

For fixed β, $\widetilde{L}(\beta, \Lambda_0)$ is maximized at

$$\lambda_i = \frac{\Delta_i}{\sum_{j=1}^{n} I(\widetilde{T}_j \ge \widetilde{T}_i) e^{\beta' X_j(\widetilde{T}_i)}}, \quad i = 1, \cdots, n$$

Profile likelihood (partial likelihood) for β:

$$PL(\beta) = \sup_{\Lambda_0} \widetilde{L}(\beta, \Lambda_0) \propto \prod_{i=1}^{n} \left\{ \frac{e^{\beta' X_i(\widetilde{T}_i)}}{\sum_{j=1}^{n} I(\widetilde{T}_j \geq \widetilde{T}_i) e^{\beta' X_j(\widetilde{T}_i)}} \right\}^{\Delta_i}$$

Score function:

$$U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \Delta_i \left\{ X_i(\widetilde{T}_i) - \frac{\sum_{j=1}^{n} I(\widetilde{T}_j \geq \widetilde{T}_i) e^{\beta' X_j(\widetilde{T}_i)} X_j(\widetilde{T}_i)}{\sum_{j=1}^{n} I(\widetilde{T}_j \geq \widetilde{T}_i) e^{\beta' X_j(\widetilde{T}_i)}} \right\}$$

Information matrix: $\mathcal{I}(\beta) = -\partial^2 \log L(\beta) / \partial \beta^2$

MPLE $\widehat{\beta}$: $\{U(\beta) = 0\}$

Breslow Estimator:

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{I(\widetilde{T}_i \leq t)\Delta_i}{\sum_{j=1}^{n} I(\widetilde{T}_j \geq \widetilde{T}_i) e^{\widehat{\beta}' X_j(\widetilde{T}_i)}}$$

$$\widehat{S}(t|X) = \exp\left\{ -\int_0^t e^{\widehat{\beta}' X(s)} d\widehat{\Lambda}_0(s) \right\}$$

ASYMPTOTIC PROPERTIES:

$$\widehat{\beta} \sim N(\beta, \mathcal{I}^{-1}(\widehat{\beta}))$$

$$\sup_t |\widehat{\Lambda}_0(t) - \Lambda_0(t)| \xrightarrow{a.s.} 0$$

- $\widehat{S}(t|x) = \exp\left\{- \int_0^t e^{\widehat{\beta}' x(s)} d\widehat{\Lambda}_0(s)\right\}$

SOFTWARE:

- Stata stcox
- SAS PHREG
- R coxph

Analysis of Interval-Censored Data

INTRODUCTION

Interval Censoring: Failure occurs within a time interval

Medical Research: Periodic monitoring of asymptomatic diseases

- HIV infection
- SARS-Cov-2 infection
- tumor occurrence
- diabetes onset

Theoretical/Computational Issues: No exact failure time

NPMLE

- asymptotic theory
- EM-type algorithm
- software

METHODS

Notation

T = failure time

X = (potentially time-dependent) covariates

$\lambda(t|X)$ = hazard function of T conditional on X

Cox PH Model

$$\lambda(t|X) = \lambda_0(t)e^{\beta' X(t)}$$

- β = regression parameters
- $\lambda_0(\cdot)$ = arbitrary baseline hazard function
- $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$

Data: $(L_i, R_i, X_i) \quad (i = 1, \ldots, n)$

Likelihood

$$\prod_{i=1}^{n} \left[\exp\left\{ -\int_0^{L_i} e^{\beta' X_i(s)} d\Lambda_0(s) \right\} - \exp\left\{ -\int_0^{R_i} e^{\beta' X_i(s)} d\Lambda_0(s) \right\} \right]$$

NPMLE

$$\prod_{i=1}^{n} \left[\exp\left\{ -\sum_{t_k \le L_i} \lambda_k e^{\beta' X_i(t_k)} \right\} - I(R_i < \infty) \exp\left\{ -\sum_{t_k \le R_i} \lambda_k e^{\beta' X_i(t_k)} \right\} \right]$$

$$= \prod_{i=1}^{n} \exp\left(-\sum_{t_k \le L_i} \lambda_k e^{\beta' X_{ik}} \right) \left\{ 1 - \exp\left(-\sum_{L_i < t_k \le R_i} \lambda_k e^{\beta' X_{ik}} \right) \right\}^{I(R_i < \infty)}$$

- $t_1 < \cdots < t_m = \{L_i > 0, R_i < \infty; i = 1, \ldots, n\}$
- $\lambda_k =$ jump size of Λ at t_k
- $X_{ik} = X_i(t_k)$

Implementation

- Direct maximization
  - non-concave likelihood
  - no analytic expression for λ_k
  - many λ_k are zero
- EM algorithm
  - latent Poisson variables with same observed-data likelihood
  - analytic expression for λ_k
  - partial-likelihood like estimating equation for β
  - observed-data likelihood increases at each iteration

EM Algorithm

Latent variables: $W_{ik} \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_k e^{\beta' X_{ik}})$ $(i = 1, \ldots, n; k = 1, \ldots, m)$

Observed data: $(L_i, R_i, X_i, A_i = 0, B_i > 0)$ $(i = 1, \ldots, n)$

- $A_i = \sum_{t_k \le L_i} W_{ik}$
- $B_i = I(R_i < \infty) \sum_{L_i < t_k \le R_i} W_{ik}$

Observed-data likelihood

$$\prod_{i=1}^{n} \left\{ \prod_{t_k \le L_i} \Pr(W_{ik} = 0) \right\} \left\{ 1 - \Pr\left(\sum_{L_i < t_k \le R_i} W_{ik} = 0 \right) \right\}^{I(R_i < \infty)}$$

Complete-data log-likelihood

$$\sum_{i=1}^{n} \sum_{k=1}^{m} I(t_k \le R_i^*) \left\{ W_{ik} \log(\lambda_k e^{\beta' X_{ik}}) - \lambda_k e^{\beta' X_{ik}} - \log W_{ik}! \right\}$$

- $R_i^* = I(R_i < \infty) R_i + I(R_i = \infty) L_i$

E-step

$$\widehat{E}(W_{ik}) = \begin{cases} 0 & \text{if } t_k \leq L_i \\ \frac{\lambda_k e^{\beta' X_{ik}}}{1-\exp\left(-\sum_{L_i<t_l\leq R_i} \lambda_l e^{\beta' X_{il}}\right)} & \text{if } L_i < t_k \leq R_i < \infty \end{cases}$$

M-step

$$\sum_{i=1}^{n}\sum_{k=1}^{m} I(R_i^* \geq t_k)\widehat{E}(W_{ik}) \left\{ X_{ik} - \frac{\sum_{j=1}^{n} I(R_j^* \geq t_k) e^{\beta' X_{jk}} X_{jk}}{\sum_{j=1}^{n} I(R_j^* \geq t_k) e^{\beta' X_{jk}}} \right\} = 0$$

$$\lambda_k = \frac{\sum_{i=1}^{n} I(R_i^* \geq t_k)\widehat{E}(W_{ik})}{\sum_{j=1}^{n} I(R_j^* \geq t_k) e^{\beta' X_{jk}}} \quad (k = 1, \ldots, m)$$

Exact failure times: $T_i = L_i = R_i$

$$\widehat{E}(W_{ik}) = \begin{cases} 1 & \text{if } T_i = t_k \\ 0 & \text{if } T_i \neq t_k \end{cases}$$

$$\sum_{i=1}^{n} \sum_{k=1}^{m} I(T_i = t_k) \left\{ X_{ik} - \frac{\sum_{j=1}^{n} I(R_j^* \geq t_k) e^{\beta' X_{jk}} X_{jk}}{\sum_{j=1}^{n} I(R_j^* \geq t_k) e^{\beta' X_{jk}}} \right\} = 0$$

$$\lambda_k = \frac{\sum_{i=1}^{n} I(T_i = t_k)}{\sum_{j=1}^{n} I(R_j^* \geq t_k) e^{\beta' X_{jk}}} \quad (k = 1, \ldots, m)$$

Unified algorithm for right- and interval-censored data

Partially interval-censored data

Asymptotic Properties

Consistency

$$\|\widehat{\beta} - \beta\| + \sup_t |\widehat{\Lambda}_0(t) - \Lambda_0(t)| \xrightarrow{a.s.} 0$$

Asymptotic distribution

$$n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0, \Sigma)$$

- Σ = semiparametric efficiency bound
- Σ can be consistently estimated by the information matrix of the profile log-likelihood for β

Profile Log-likelihood

$$pl(\beta) = \sum_{i=1}^{n} \log \left\{ \exp\left(- \sum_{t_k \le L_i} \widetilde{\lambda}_k e^{\beta' X_{ik}} \right) - I(R_i < \infty) \exp\left(- \sum_{t_k \le R_i} \widetilde{\lambda}_k e^{\beta' X_{ik}} \right) \right\}$$

- $\widetilde{\lambda}_k$ $(k = 1, \ldots, m)$ are obtained from EM algorithm with fixed β

Covariance Matrix Estimator of $\widehat{\beta}$:

$$-\left\{D_h^2 pl(\widehat{\beta})\right\}^{-1} \approx \left\{\sum_{i=1}^{n} D_h pl_i(\widehat{\beta}) D_h pl_i(\widehat{\beta})'\right\}^{-1}$$

- $pl_i(\beta) = i$th term of $pl(\beta)$
- $D_h f(\beta) = \left(\frac{f(\beta+he_j)-f(\beta)}{h}\right)_{j=1,\ldots,p}$
- $D_h^2 f(\beta) = \left[\frac{f(\beta)-f(\beta+he_j)-f(\beta+he_k)+f(\beta+he_j+he_k)}{h^2}\right]_{j,k=1,\ldots,p}$
- $e_j = j$th canonical vector in $\mathcal{R}^p$
- $h =$ perturbation constant in the order of $n^{-1/2}$

Statistical Inference:

- Wald statistics based on $\widehat{\beta}$ and its covariance matrix estimator
- Likelihood ratio statistics based on profile log-likelihood

Software

- IntCens (http://dlin.web.unc.edu/software)
- Stata stintcox
- SAS ICPHREG

HIV STUDY

Bangkok Metropolitan Administration Study: cohort of 1,209 injecting drug users initially sero-negative for HIV-1

Study Period: 1995 $\sim$ 1998

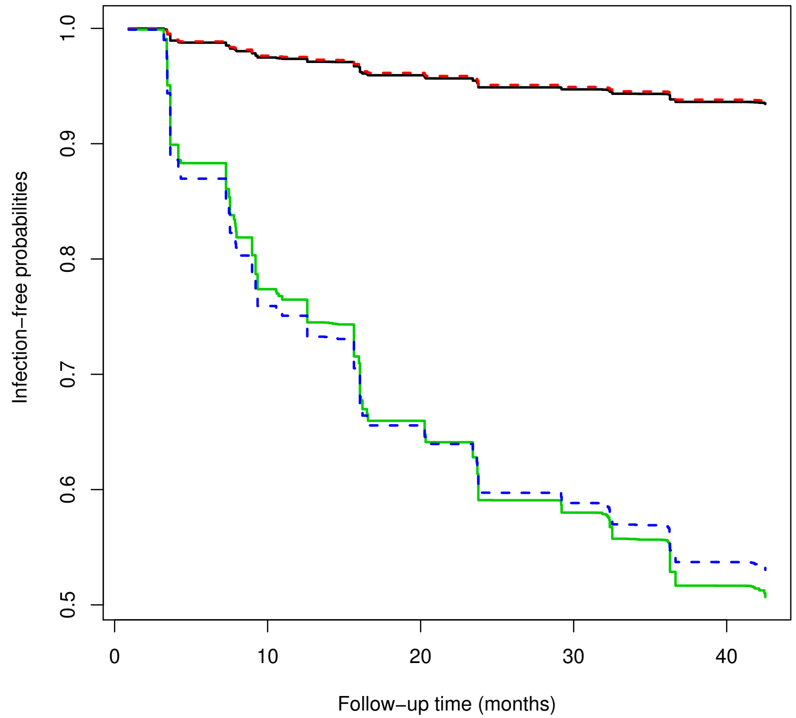
Blood Tests for HIV-1:

- at study enrollment
- approximately every 4 months thereafter

Data:

- 2,300 person-years of follow-up
- 133 HIV-1 sero-conversions
- risk factors

Risk Factor	Est	St error	p-value
Age	−0.028	0.012	0.021
Gender	0.424	0.270	0.117
Needle sharing	0.237	0.183	0.196
Drug injection	0.313	0.184	0.089
Imprisonment	0.502	0.211	0.017



Estimation of infection-free probabilities for a high-risk versus a low-risk subject

- solid curves $\sim$ proportional hazards
- dashed curves $\sim$ proportional odds

Analysis of Multivariate Interval-Censored Data

METHODS

Notation

n = number of clusters

n_i = number of subjects in the ith cluster

K = types of failures

T_{ijk} = kth failure time for the jth subject of the ith cluster

$X_{ijk}(\cdot)$ = (time-dependent) covariates

Marginal Cox Models

$$\lambda_{ijk}(t) = \lambda_{k0}(t)e^{\beta_k' X_{ijk}(t)}$$

- β_k = regression parameters
- $\lambda_{k0}(\cdot)$ = arbitrary baseline hazard function
- $\Lambda_{k0}(t) = \int_0^t \lambda_{k0}(s)ds$

Data: $(L_{ijk}, R_{ijk}, X_{ijk})$ $(i = 1, \ldots, n; j = 1, \ldots, n_i; k = 1, \ldots, K)$

Pseudo-Likelihood

$$\prod_{i=1}^{n} \prod_{j=1}^{n_i} \prod_{k=1}^{K} \left[\exp\left\{ -\int_0^{L_{ijk}} e^{\beta' X_{ijk}(s)} d\Lambda_{k0}(s) \right\} - \exp\left\{ -\int_0^{R_{ijk}} e^{\beta' X_{ijk}(s)} d\Lambda_{k0}(s) \right\} \right]$$

Nonparametric Maximum Pseudo-Likelihood Estimation

- $0 < t_{k0} < t_{k1} < \cdots < t_{km_k} < \infty = \{L_{ijk} > 0, R_{ijk} < \infty; i = 1, \ldots, n; j = 1, \ldots, n_i\}$
- λ_{kq} = jump size of $\Lambda_{k0}(\cdot)$ at t_{kq}

EM Algorithm

Latent variables: $W_{ijkq} \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_{kq} e^{\beta_k' X_{ijkq}})$
$(i = 1, \ldots, n; j = 1, \ldots, n_i; k = 1, \ldots, K; q = 1, \ldots, m_k)$

- $X_{ijkq} = X_{ijk}(t_{kq})$

Observed data: $(L_{ijk}, R_{ijk}, X_{ijk}, A_{ijk} = 0, B_{ijk} > 0)$
$(i = 1, \ldots, n; j = 1, \ldots, n_i; k = 1, \ldots, K)$

- $A_{ijk} = \sum_{t_{kq} \le L_{ijk}} W_{ijkq}$
- $B_{ijk} = I(R_{ijk} < \infty) \sum_{L_{ijk} < t_{kq} \le R_{ijk}} W_{ijkq}$

E-step

$$\widehat{E}(W_{ijkq}) = I(L_{ijk} < t_{kq} \le R_{ijk} < \infty) \frac{\lambda_{kq} e^{\beta_k' X_{ijkq}}}{1 - \exp\{-\sum_{L_{ijk} < t_{kq'} \le R_{ijk}} \lambda_{kq'} e^{\beta_k' X_{ijkq'}}\}}$$

M-step

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{q=1}^{m_k} I(R_{ijk}^* \ge t_{kq}) \widehat{E}(W_{ijkq})$$

$$\times \left\{ X_{ijkq} - \frac{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(R_{i'j'k}^* \ge t_{kq}) e^{\beta_k' X_{i'j'kq}} X_{i'j'kq}}{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(R_{i'j'k}^* \ge t_{kq}) e^{\beta_k' X_{i'j'kq}}} \right\} = 0$$

- $R_{ijk}^* = I(R_{ijk} < \infty) R_{ijk} + I(R_{ijk} = \infty) L_{ijk}$

$$\lambda_{kq} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I(R_{ijk}^* \ge t_{kq}) \widehat{E}(W_{ijkq})}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I(R_{ijk}^* \ge t_{kq}) e^{\beta_k' X_{ijkq}}}$$

Asymptotic Properties

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \qquad \widehat{\beta} = \begin{bmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_K \end{bmatrix}$$

$$\|\widehat{\beta} - \beta\| + \sum_{k=1}^{K} \sup_t |\widehat{\Lambda}_{k0}(t) - \Lambda_{k0}(t)| \xrightarrow{\text{a.s.}} 0$$

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{D} N(0, \Omega)$$

Profile Pseudo-log-likelihood for β_k

$$pl_k(\beta_k) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log \left\{ \exp\left(- \sum_{t_{kq} \le L_{ijk}} \widetilde{\lambda}_{kq} e^{\beta_k' X_{ijkq}} \right) \right.$$

$$\left. - I\left(R_{ijk} < \infty\right) \exp\left(- \sum_{t_{kq} \le R_{ijk}} \widetilde{\lambda}_{kq} e^{\beta' X_{ijkq}} \right) \right\}$$

- $\widetilde{\lambda}_{kq}$ $(q = 1, \ldots, m_k)$ are obtained from EM with fixed β_k

Covariance matrix estimator between $\widehat{\beta}_k$ and $\widehat{\beta}_l$

$$V_{kl} = \left\{ D_h^2 pl_k(\widehat{\beta}_k) \right\}^{-1} \sum_{i=1}^{n} D_h pl_{ki}(\widehat{\beta}_k) D_h pl_{li}(\widehat{\beta}_l)^{\mathrm{T}} \left\{ D_h^2 pl_l(\widehat{\beta}_l) \right\}^{-1}$$

- $pl_{ki}(\beta_k)$ = contribution of the ith cluster to $pl_k(\beta_k)$

Statistical Inference:

$$L\widehat{\beta} \sim N(L\beta, LVL')$$

- linear combinations (e.g., a subset of parameters, difference of two parameters)

$$V = \begin{bmatrix} V_{11} & \cdots & V_{1K} \\ \vdots & \vdots & \vdots \\ V_{K1} & \cdots & V_{KK} \end{bmatrix}$$

ARIC STUDY

Atherosclerosis Risk in Communities Study (ARIC): cohort of 14,751 white and black individuals from 4 U.S. communities

Baseline examination: 1987–1989

Follow-up examinations: 3-year intervals

Final examination: 2011–2013

Diabetes

- fasting glucose $\geq$ 126 mg/dL
- non-fasting glucose $\geq$ 200 mg/dL
- self-reported physician diagnosis of diabetes
- use of diabetic medication

Hypertension

- systolic blood pressure $\geq$ 140
- diastolic blood pressure $\geq$ 90
- use of anti-hypertensive medication

Analysis set: 8,735 individuals without diabetes or hypertension

	Diabetes			Hypertension			Overall test		Difference		
Factor	Est	SE	*P*	Est	SE	*P*	Test	*P*	Est	SE	95% CI
Jackson	-.145	.149	.332	-.239	.077	.002	10.1	.006	.094	.162	(-.234, .413)
Minn.	-.389	.076	.000	-.100	.046	.031	29.2	.000	-.289	.085	(-.455, -.122)
Wash.	.115	.073	.114	.078	.048	.103	4.68	.096	.037	.083	(-.125, .199)
Age	-.014	.005	.007	.013	.003	.000	26.2	.000	-.027	.006	(-.038, -.016)
Male	-.062	.055	.265	-.238	.034	.000	49.3	.000	.176	.062	(.056, .297)
White	-.451	.160	.005	-.480	.081	.000	40.3	.000	.029	.172	(-.307, .366)
BMI	.075	.005	.000	.017	.004	.000	237	.000	.059	.006	(.047, .070)
Glucose	.096	.003	.000	.001	.002	.595	962	.000	.095	.004	(.088, .102)
SBP	.005	.003	.096	.058	.002	.000	914	.000	-.053	.003	(-.060, -.046)
DBP	.005	.004	.310	.011	.003	.000	17.5	.000	-.007	.005	(-.016, .003)

Est, estimate

SE, standard error

P, p-value

CI, confidence interval

REMARKS

Random-Effects Models for Multivariate Interval-Censored Data

Mixed Censoring

- interval censoring
- right censoring

Informative Drop-out

Panel Count Data

Competing Risks

REFERENCES

Zeng, D., Mao, L. & Lin, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, **103**, 253–271.

Xu, Y., Zeng, D. & Lin, D. Y. (2022). Marginal proportional hazards models for multivariate interval-censored data. *Biometrika*, in revision.

Mao, L., Lin, D. Y. & Zeng, D. (2017). Semiparametric regression analysis of interval-censored competing risks data. *Biometrics*, **73**, 857–865.

Zeng, D., Gao, F. & Lin, D. Y. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika*, **104**, 505–525.

Zeng, D & Lin, D. Y. (2020). Maximum likelihood estimation for semiparametric regression models with panel count data. *Biometrika*, **108**, 947–963.

Gao, F., Zeng, D. & Lin, D. Y. (2019). Semiparametric regression analysis of interval-censored data with informative dropout. *Biometrics*, **74**, 1213–1222.

Gao, F., Zeng, D., Couper, D. & Lin, D. Y. (2019). Semiparametric regression analysis of multiple right- and interval-censored events. *J. Amer. Statist. Assoc.*, **114**, 1232–1240.