Dependent error in covariates with Cox regression: Work in progress

Yijian (Eugene) Huang

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

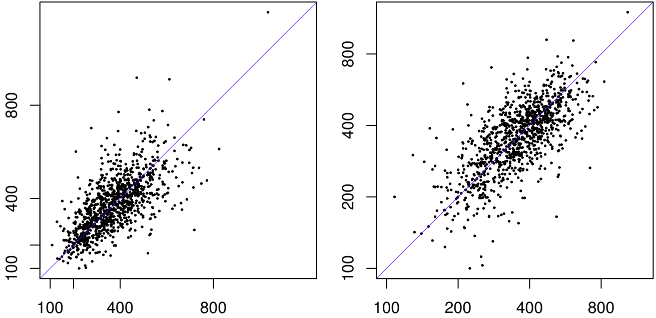
yhuang5@emory.edu

BIRS, Banff, August 2016

Supported by NIH grants P30AI050409 and P20HL113451

Classical MEM realistic? A CD4 example

$$W = X + \varepsilon \qquad \varepsilon \perp\!\!\!\perp X \text{ (and other r.v.'s)}$$



No, if W is CD4. Better, if W is log(CD4).

Often, ε is correlated with X, e.g., heteroscedastic error. And the dependence structure may be hard to model.

Problem of interest

Survival time S and censoring time C are observed thru

$$T = \min(S, C), \qquad \Delta = I(S \leq C).$$

With covariates $\mathbf{X}$, the PHM postulates

$$d\Lambda(t \mid \mathbf{X}) = \exp(\beta^\top \mathbf{X})\, d\Lambda_0(t), \qquad S \perp\!\!\!\perp C \mid \mathbf{X}.$$

$\mathbf{X} \equiv (X, \mathbf{Z}^\top)^\top$:

- Scalar X cannot be accurately measured, but thru W and an IV U.
- $\mathbf{Z}$ is accurately measured.

Data:n iid replicates of $\{T, \Delta, W, U, \mathbf{Z}\}$.

Goal: functional modeling to tackle heteroscedastic ε.

Related literature

- Approximate estimation: regression calibration (Prentice, 1982; Wang et al., 1997; Xie et al., 2001), Li & Ryan (2004)
- Likelihood-based approaches of Hu et al. (1998): restrictive assumptions on censoring mechanism and even X distribution
- Functional modeling
  - w/i error assumption: corrected score (Nakamura, 1992), conditional score (Tsiatis & Davidian, 2001)
  - no / minimal error assumption: nonparametric correction (Huang & Wang, 2000, 2006; Song & Wang, 2014), Hu & Lin (2004)

Proposed heteroscedastic MEM

$W = X + \varepsilon$ with $\varepsilon =$ homogeneous $\eta +$ heterogeneous ξ.

$$\eta \perp\!\!\!\perp \{T, \Delta, \mathbf{X}, \xi\} \qquad \xi \perp\!\!\!\perp \{T, \Delta\} \mid \mathbf{X} \qquad \xi \sim -\xi \mid \mathbf{X}$$

- $\xi \equiv 0 \Longrightarrow$ Classical MEM
- certain structure on ξ necessary for identifiability: symmetry seems reasonable and mild
- alternative formulation possible; e.g., in a special case, multiplicative-additive error:

$$W = X\zeta + \eta \qquad \eta \perp\!\!\!\perp \{T, \Delta, \mathbf{X}, \zeta\}$$
$$\zeta \perp\!\!\!\perp \{T, \Delta\} \mid \mathbf{X} \qquad \zeta - 1 \sim -(\zeta - 1) \mid \mathbf{X}$$

IV may be (marginally) correlated with ε: $U \perp\!\!\!\perp \{T, \Delta, W\} \mid \mathbf{X}$

Motivating identity

Notation: $N(t) = \Delta I(T \leq t)$, $Y(t) = I(T \geq t)$, $(\beta_1, \boldsymbol{\beta}_2^\top)^\top \equiv \boldsymbol{\beta}$.

Martingale under the PHM:

$$\exp(-\beta_1 X/2)\left\{N(t) - \int_0^t Y(s)\exp(\boldsymbol{\beta}^\top \mathbf{X})\, d\Lambda_0(s)\right\}$$

Under the MEM, $\forall t \geq 0$,

$$\mathbb{E}\left\{\exp(-\beta_1 W/2)N(t) - \int_0^t Y(s)\exp(\beta_1 W/2 + \boldsymbol{\beta}_2^\top \mathbf{Z})\, d\Omega_0(s) \mid \mathbf{X}, U\right\} = 0$$

with $\Omega_0(t) = \Lambda_0(t)\mathbb{E}\{\exp(-\beta_1\eta/2)\}/\mathbb{E}\{\exp(\beta_1\eta/2)\}$.

Proposed estimating equations

Write $\mathbf{U} \equiv (U, \mathbf{Z}^\top)^\top$.

$$\begin{aligned} \mathbb{E}_n \left[\mathbf{U} \left\{\exp(-b_1 W/2)\Delta - \exp(b_1 W/2 + \mathbf{b}_2^\top \mathbf{Z})\Omega(T)\right\}\right] &= 0 \\ \mathbb{E}_n \left\{\exp(-b_1 W/2)N(t) - \int_0^t Y(s)\exp(b_1 W/2 + \mathbf{b}_2^\top \mathbf{Z})\, d\Omega(s)\right\} &= 0 \\ & \forall t \geq 0 \end{aligned}$$

Profiling out $\Omega(\cdot) \Longrightarrow$

$$\mathbb{E}_n \int_0^\infty \exp[-b_1 W/2] \left[\mathbf{U} - \frac{\mathbb{E}_n \left\{Y(t)\mathbf{U}\exp\left(b_1 W/2 + \mathbf{b}_2^\top \mathbf{Z}\right)\right\}}{\mathbb{E}_n \left\{Y(t)\exp\left(b_1 W/2 + \mathbf{b}_2^\top \mathbf{Z}\right)\right\}}\right] dN(t) = 0$$

Comments

If $\xi = 0$, the EF is a corrected version of

$$\mathbb{E}_n \int_0^\infty \exp[-b_1 X/2] \left[\mathbf{U} - \frac{\mathbb{E}_n \left\{ Y(t) \mathbf{U} \exp\left(b_1 X/2 + \mathbf{b}_2^\top \mathbf{Z} \right) \right\}}{\mathbb{E}_n \left\{ Y(t) \exp\left(b_1 X/2 + \mathbf{b}_2^\top \mathbf{Z} \right) \right\}} \right] dN(t).$$

NOT in general.

Unlike CorrS / ConS / NC, this EF doesn't reduce to a (weighted) partial score when $W = U = X$.

$\Lambda_0(\cdot)$ is not fully identifiable, unless, say, η is zero-symmetric. If so, $\Lambda_0(\cdot) = \Omega_0(\cdot)$ may be estimated similar to Huang & Wang (2000).

Simulations

2 estimators for comparison:

- Naive (NV): Cox regression with X replaced with W
- Adapted NC:

$$\mathbb{E}_n \int_0^\infty \left[\mathbf{U} - \frac{\mathbb{E}_n \left\{ Y(t) \mathbf{U} \exp(\mathbf{b}^\top \mathbf{W}) \right\}}{\mathbb{E}_n \left\{ Y(t) \exp(\mathbf{b}^\top \mathbf{W}) \right\}} \right] dN(t) = 0$$

Skewness measure for normality adequacy:

$$\frac{Q(0.975) - Q(0.5)}{Q(0.5) - Q(0.025)}$$

Single-covariate model

- $\Lambda_0(t) = t$
- $X \sim N(0,1)$, $\beta = 1$
- censoring rate: 25%

$$C \sim \begin{cases} U[0,4.3] & X < 0 \\ U[0,8.6] & X \geq 0 \end{cases}$$

- $\xi \mid X \sim N$ with mean 0 and var $\propto \Phi(X)$ / $\xi = 0$
- $\eta = 0$ / $\eta \sim N$ / $\eta \sim$ scaled Beta(2,1)
- moderate / substantial error: $\sigma_{\eta+\xi} = 1/2$ or 1
- $U \sim N(0,1)$ s.t. $(X, U) \sim BN$, $\rho = 0.8$
- sample size: 300 or 600
- 1000 replicates

Heteroscedastic error: $\xi \mid X \sim$ normal

size		F	B	D	E	C	S	F	B	D	E	C	S
				$\sigma_{\eta+\xi} = 1/2$						$\sigma_{\eta+\xi} = 1$			
							$\eta = 0$						
300	NV	0.0	−306	71	68	1.9	1.14	0.0	−599	53	50	0.0	1.10
	NC	0.1	−46	127	126	90.6	1.29	1.2	−87	1187	288	65.2	2.08
	PP	0.0	12	134	139	96.9	1.17	0.0	25	167	171	96.1	1.37
600	NV	0.0	−310	51	49	0.0	1.10	0.0	−601	37	36	0.0	1.07
	NC	0.0	−55	92	88	86.1	1.21	0.2	−163	117	104	50.2	1.64
	PP	0.0	7	97	98	94.9	1.06	0.0	15	120	118	94.9	1.19
							$\eta \sim$ scaled Beta(2,1)						
300	NV	0.0	−282	71	70	3.3	1.23	0.0	−587	53	51	0.0	1.21
	NC	0.0	−12	134	130	94.6	1.44	1.1	−45	553	265	86.1	2.00
	PP	0.0	17	141	141	95.5	1.17	0.0	31	173	175	95.8	1.48
600	NV	0.0	−287	50	50	0.0	1.02	0.0	−589	37	37	0.0	1.04
	NC	0.0	−25	92	90	91.8	1.25	0.2	−88	118	108	77.5	1.48
	PP	0.0	6	99	99	94.3	1.10	0.0	13	121	120	94.3	1.23

Homogeneous error: $\xi = 0$

size		F	B	D	E	C	S	F	B	D	E	C	S
				$\sigma_\eta = 1/2$						$\sigma_\eta = 1$			
							$\eta \sim$ normal						
300	NV	0.0	−271	74	72	4.7	1.16	0.0	−593	56	53	0.0	1.09
	NC	0.0	16	140	139	95.2	1.29	3.3	66	392	262	94.2	2.13
	PP	0.0	12	139	142	96.1	1.22	0.2	25	179	179	96.1	1.46
600	NV	0.0	−274	52	51	0.1	1.05	0.0	−595	39	38	0.0	1.04
	NC	0.0	7	100	96	94.7	1.22	1.5	45	286	184	94.2	1.93
	PP	0.0	7	100	100	95.3	1.06	0.0	15	126	123	94.7	1.21
							$\eta \sim$ scaled Beta(2,1)						
300	NV	0.0	−248	71	71	8.2	1.02	0.0	−561	54	54	0.0	1.00
	NC	0.0	24	141	132	94.5	1.23	0.0	43	188	167	93.9	1.70
	PP	0.0	24	151	143	94.4	1.44	0.0	41	195	180	94.6	1.78
600	NV	0.0	−252	50	50	0.0	1.12	0.0	−563	38	38	0.0	1.11
	NC	0.0	12	95	93	94.8	1.07	0.0	21	117	113	95.0	1.24
	PP	0.0	11	103	100	95.5	1.19	0.0	18	127	122	95.5	1.30

Double-covariate model with heteroscedastic error

$X \perp\!\!\!\perp Z$

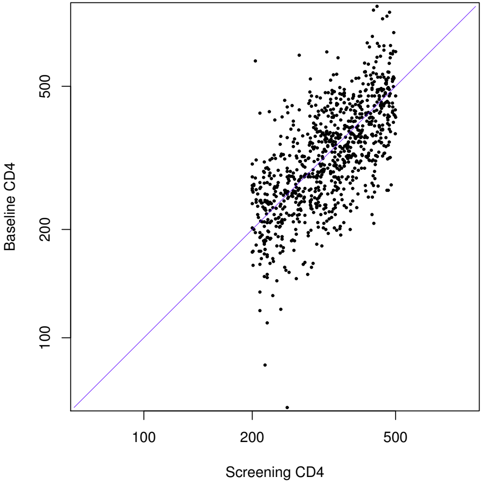
size		F	B	D	E	C	S	F	B	D	E	C	S
				$\sigma_\xi = 1/2$						$\sigma_\xi = 1$			
							$X \perp\!\!\!\perp Z$						
300	NV	0.0	−294	72	69	2.3	1.13	0.0	−590	54	51	0.0	1.14
			−81	87	83	80.8	1.15		−169	85	81	43.1	1.09
	NC	0.1	−27	150	154	91.6	1.29	1.4	−116	221	163	72.5	1.71
			25	119	123	94.2	1.35		57	178	153	94.5	1.49
	PP	0.0	25	140	142	95.4	1.17	0.2	48	193	184	95.7	1.52
			20	111	106	94.1	1.13		34	141	133	95.7	1.30
600	NV	0.0	−301	49	50	0.0	1.06	0.0	−595	37	37	0.0	1.08
			−88	59	59	66.4	0.99		−176	58	57	14.1	0.96
	NC	0.0	−44	90	89	88.4	1.17	0.5	−152	112	101	55.7	1.41
			11	76	74	95.0	1.23		28	112	101	94.4	1.37
	PP	0.0	13	98	99	96.3	1.22	0.0	23	122	120	95.5	1.35
			10	76	74	94.9	1.04		16	92	89	94.2	1.18

X and Z: $\rho = 0.5$

size		F	B	D	E	C	S	F	B	D	E	C	S
				$\sigma_\xi = 1/2$						$\sigma_\xi = 1$			
							X and Z: $\rho = 0.5$						
300	NV	0.0	−341	78	75	1.1	1.03	0.0	−646	56	54	0.0	1.00
			44	100	94	92.4	1.25		84	101	94	84.5	1.17
	NC	0.2	−23	179	164	92.7	1.52	2.1	−101	291	218	74.7	2.23
			20	122	118	93.7	1.13		53	223	175	92.6	1.02
	PP	0.0	26	165	164	95.7	1.20	0.5	49	218	219	96.5	1.62
			14	117	113	93.7	1.03		24	148	141	94.2	1.02
600	NV	0.0	−348	54	54	0.0	1.07	0.0	−651	39	39	0.0	1.10
			37	68	67	91.0	1.15		76	69	67	78.8	1.04
	NC	0.0	−42	107	106	89.7	1.27	0.6	−145	144	142	62.7	1.67
			10	81	79	94.4	1.13		23	120	114	92.4	1.08
	PP	0.0	13	110	115	96.2	1.22	0.0	25	141	144	96.4	1.46
			7	80	79	94.2	1.01		12	98	97	94.4	1.03

ACTG 175 study (Hammer et al. 1996)

- randomized trial: screening CD4 200 – 500, no AIDS at baseline
- comparing ZDV, ZDV+ddI, ZDV+ddC, ddC
- endpoint: AIDS or death
- analysis dataset of size 900: ≥ 1 yr prior ART; baseline CD4 prior to the study tx and within 5 days of randomization (not used for screening)
- 35-month median follow-up with 158 events
- variance ratio 0.53 of error vs. true log(CD4) (estimated from the subset of 277 patients with duplicates)
- error heteroscedasticity: sample variances of difference between baseline log(CD4) duplicates were 0.088, 0.089, and 0.068 for the sub-samples with screening CD4 in $[200, 300)$, $[300, 400)$, and $[400, 500]$


Baseline CD4
Screening CD4
100
200
500
100
200
500

	log(CD4)		ZDV+ddI		ZDV+ddC		ddI	
	Est	SE	Est	SE	Est	SE	Est	SE
NV	−1.579	0.220	−0.515	0.235	−0.200	0.211	−0.423	0.218
NC	−2.440	0.365	−0.626	0.257	−0.276	0.233	−0.441	0.224
PP	−2.762	0.457	−0.431	0.259	−0.182	0.231	−0.358	0.237

Remarks

- Work in progress: reasonably good numerical performance
- Generalizing the proposed EF so as to improve efficiency?
- Extension to accommodate time-dependent covariates with time-dependent measurement error?
- Specialization when a replicate is the IV?