

Supplementary Material: Pseudo Partial Likelihood Method for Proportional Hazards Models When Time Origin Is Missing for Control Group With Applications to SARS-CoV-2 Seroprevalence Study

YUNRO CHUNG^{1,2}, VEL MURUGAN², KASSU MEHARI BEYENE³, AND DING-GENG CHEN^{1,4}

¹College of Health Solutions, Arizona State University, U.S.A.

²Virginia G. Piper Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, U.S.A.

³Barrow Neurological Institute, U.S.A.

⁴Department of Statistics, University of Pretoria, South Africa

A Proofs

The following assumptions were made:

Assumption 1. The baseline hazard function $\lambda_0(t)$ is bounded over $[0, \tau]$ by some constant λ_U .

Assumption 2. Z is bounded. The parameter $\theta = (\beta, \eta)$ lies in a compact set $\Theta = \mathbb{B} \times \mathbb{H}$ that includes an open neighborhood of θ .

Assumption 3. The limiting value of the matrix $H(\theta_0, \eta)$ is positive definite for all $\eta \in \mathbb{H}$.

Assumption 4. p and m are finite.

Assumptions 1–3 are standard in survival data analysis. Assumption 4 implies that the number of parameters, including the number of control samples, does not grow as $n \rightarrow \infty$.

Proof of Theorem 1. For any $\eta \in \mathbb{H}$, the approximation error between $I_{ij,s}(\eta)$ and $I_{ij}(\eta)$ is

$$\begin{aligned} 0 \leq \epsilon_{ij,s}(\eta) &\equiv |I_{ij,s}(\eta) - I_{ij}(\eta)| = \max\{|\Phi_s(C_{ij1}) - 0|, |\Phi_s(C_{ij2}) - 1|\} \\ &\leq \max\{\Phi_s(C_{ij1}), 1 - \Phi_s(C_{ij2})\} \end{aligned}$$

for some $-\infty < C_{ij1} < C_{ij2} < \infty$ and $i = 1, \dots, n$ and $j = 1, \dots, n$, with $\epsilon_{ij,s}(\eta) = 0$ if $i \in \mathbb{T}$ and $j \in \mathbb{T}$. Thus,

$$0 \leq \epsilon_s \equiv \max_{i,j} \left\{ \sup_{\eta \in \mathbb{H}} |\epsilon_{ij,s}(\eta)| \right\} \leq \Phi_s(C)$$

for some finite constant C . By the triangle inequality,

$$\begin{aligned} 0 \leq \sup_{\theta \in \Theta} |\ell_n^P(\theta) - \ell_{n,s}^P(\theta)| &\leq \sup_{\theta \in \Theta} \left\{ \sum_{i=1}^n \left| \log \left(\frac{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j}}{\sum_{j=1}^n I_{ij}(\eta) e^{\beta^\top Z_j}} \right) \right| \right\} \\ &\leq e^{U-L} \sup_{\eta \in \mathbb{H}} \left\{ \sum_{i=1}^n \left| \log \left(\frac{\sum_{j=1}^n I_{ij,s}(\eta)}{\sum_{j=1}^n I_{ij}(\eta)} \right) \right| \right\} \\ &\leq e^{U-L} \sup_{\eta \in \mathbb{H}} \left\{ \sum_{i=1}^n \left| \log \left(\frac{n^{-1} \sum_{j=1}^n |\epsilon_{ij,s}(\eta)| + n^{-1} \sum_{j=1}^n |I_{ij}(\eta)|}{n^{-1} \sum_{j=1}^n I_{ij}(\eta)} \right) \right| \right\} \\ &\leq e^{U-L} n \log \left(\left| \frac{\Phi_s(C) + \bar{a}_n}{\bar{b}_n} \right| \right) \end{aligned} \tag{A1}$$

for some $-\infty < L < U < \infty$, where

$$\bar{a}_n = \max_i \left\{ \sup_{\eta \in \mathbb{H}} \left\{ n^{-1} \sum_{j=1}^n I_{ij}(\eta) \right\} \right\}$$

and

$$\bar{b}_n = \min_i \left\{ \inf_{\eta \in \mathbb{H}} \left\{ n^{-1} \sum_{j=1}^n I_{ij}(\eta) \right\} \right\}.$$

By the law of large number, \bar{a}_n and \bar{b}_n converge to some constants a and b as $n \rightarrow \infty$, respectively, with $0 < a, b < 1$. Thus, by taking $s = C\bar{\Phi}(\exp(-n^{-r}))$ for $r > 1$, (A1) is bounded above by zero as $n \rightarrow \infty$, where $\bar{\Phi}$ is the inverse of Φ . This implies $\ell_{n,s}^P(\theta)$ uniformly converges to $\ell_n^P(\theta)$ over Θ as $n \rightarrow \infty$. \square

Proof of Theorem 2. Suppose that the algorithm, stated in Subsection 2.4, converges at the r th step and $\eta_n^{(r)} = (\eta_{n1}^{(r)}, \eta_{n2}^{(r)}, \dots, \eta_{nm}^{(r)})$ is in the $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_m)$ neighborhood of $\eta_0 = (\eta_{01}, \eta_{02}, \dots, \eta_{0m})$ in the sense that

$$I_{ij}(\eta_n^{(r)}) = \begin{cases} I(\eta_{nj}^{(r)} \geq \eta_{ni}^{(r)}) = I(\eta_{0j} + \zeta_j \geq \eta_{0i} + \zeta_i) = I(\eta_{0j} \geq \eta_{0i}) & \text{if } (i, j) \in (\mathbb{C}, \mathbb{C}), \\ I(Y_j \geq \eta_{ni}^{(r)}) = I(Y_j \geq \eta_{0i} + \zeta_i) = I(Y_j \geq \eta_{0i}) & \text{if } (i, j) \in (\mathbb{C}, \mathbb{T}), \\ I(\eta_{nj}^{(r)} \geq Y_i) = I(\eta_{0j} + \zeta_j \geq Y_i) = I(\eta_{0j} \geq Y_i) & \text{if } (i, j) \in (\mathbb{T}, \mathbb{C}). \end{cases} \quad (\text{A2})$$

Since $\ell_n(\beta) = \ell_n^P(\beta, \eta_0)$, given $\eta_n^{(r)}$,

$$\begin{aligned} \sup_{\beta \in B} |\ell_{n,s}^P(\beta, \eta_n^{(r)}) - \ell_n(\beta)| &= \sup_{\beta \in B} |\ell_{n,s}^P(\beta, \eta_n^{(r)}) - \ell_n^P(\beta, \eta_0)| \\ &\leq \sup_{\beta \in B} |\ell_{n,s}^P(\beta, \eta_n^{(r)}) - \ell_n^P(\beta, \eta_n^{(r)})| + \sup_{\beta \in B} |\ell_n^P(\beta, \eta_n^{(r)}) - \ell_n^P(\beta, \eta_0)| \end{aligned}$$

by the triangle inequality. The first term on the right-hand side of the above equation is bounded above by zero as $n \rightarrow \infty$ by Theorem 1, and the second term is greater than or equal to

$$\sup_{\beta \in B} \left\{ \sum_{i=1}^n \left| \log \left(\frac{\sum_{j=1}^n I_{ij}(\eta_0) e^{\beta^\top Z_j}}{\sum_{j=1}^n I_{ij}(\eta_n^{(t)}) e^{\beta^\top Z_j}} \right) \right| \right\} \leq e^{U-L} \left\{ \sum_{i=1}^n \left| \log \left(\frac{\sum_{j=1}^n I_{ij}(\eta_0)}{\sum_{j=1}^n I_{ij}(\eta_n^{(t)})} \right) \right| \right\} = 0$$

for some $-\infty < L < U < \infty$, where the last equality holds by (A2). This implies that $\beta_n^{(r)}$ is asymptotically equivalent to $\bar{\beta}_n$ as $n \rightarrow \infty$. Under regularity conditions, the consistency and asymptotic normality of $\bar{\beta}_n$ were established by Fleming and Harrington (2013), and the same properties hold for $\beta_n^{(r)}$. \square

B Extra Simulations

We performed additional simulations under the same scenarios, as stated in Section 3, except that p was set to 0.7; thus approximately 30% of the samples were control samples with missing survival times. Compared to previous simulation studies with $p = 0.9$, the numbers of subjects in the control and treatment groups were more balanced. This further reduced the bias and RMSE, but the increased number of parameters increased computational costs, as shown in Table B1.

References

Fleming TR, Harrington DP (2013). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Table B1: Extra simulation results under with $p = 0.7$: Bias/RMSE/coverage probability, multiplied by 100. Cens: Censoring rate, CC: Complete-case, Naive Cox: Cox model with random time origin; Pseudo Cox: Proposed pseudo Cox model. CPU time in sections for Pseudo Cox.

Scenario	Cens	n	Parameter	CC Cox	Naive Cox	Pcox	CPU
1	80%	250	β_1	-	59/75/67	-2/39/96	1
		500	β_1	-	86/93/17	-3/26/99	9
		1000	β_1	-	97/100/0	-2/19/94	41
	60%	250	β_1	-	7/28/94	-3/26/96	2
		500	β_1	-	30/39/59	-1/18/96	14
		1000	β_1	-	37/42/30	-3/14/92	147
	40%	250	β_1	-	-32/39/59	-3/21/96	4
		500	β_1	-	-11/25/70	-1/14/98	21
		1000	β_1	-	2/15/76	-4/11/93	291
	20%	250	β_1	-	-55/59/9	-4/18/95	7
		500	β_1	-	-39/43/23	-2/12/96	31
		1000	β_1	-	-28/32/22	-4/10/91	401
2	80%	250	β_1	-	86/101/40	-1/37/98	2
			β_2	1/30/96	-14/29/87	2/27/93	
		500	β_1	-	113/119/4	0/25/98	8
			β_2	1/21/96	-14/22/84	1/17/95	
		1000	β_1	-	130/133/0	1/21/95	69
			β_2	3/12/95	-16/20/68	4/11/98	
	60%	250	β_1	-	37/51/71	-3/27/93	4
			β_2	0/23/94	-24/31/68	1/19/93	
		500	β_1	-	60/67/22	0/16/99	17
			β_2	-1/12/100	-24/27/40	1/11/100	
		1000	β_1	-	71/74/2	1/13/95	258
			β_2	1/9/96	-27/29/13	2/8/94	
	40%	250	β_1	-	3/27/83	-3/22/94	8
			β_2	1/19/95	-33/37/41	0/14/95	
		500	β_1	-	26/36/56	1/14/98	33
			β_2	0/11/96	-32/34/13	2/10/98	
		1000	β_1	-	39/43/17	0/10/94	515
			β_2	0/8/97	-36/37/0	1/7/97	
	20%	250	β_1	-	-21/31/73	-2/18/96	8
			β_2	1/17/93	-39/42/19	0/13/96	
		500	β_1	-	-4/20/76	0/12/95	48
			β_2	1/11/95	-40/41/1	2/9/97	
		1000	β_1	-	9/19/62	0/9/93	851
			β_2	0/7/94	-43/43/0	0/6/97	