Supplementary materials: Analysis of Cohort Studies with Multivariate, Partially Observed, Disease Classification Data

By Nilanjan Chatterjee
Division of Cancer Epidemiology and Genetics,
National Cancer Institute, NIH, DHHS. Rockville, MD 20852, USA.
chattern@mail.nih.gov

Samiran Sinha
Texas A&M University, College Station, TX 77843, USA.
sinha@stat.tamu.edu

W. Ryan Diver and Heather Spencer Feigelson
Department of Epidemiology and Surveillance Research,
American Cancer Society, Atlanta, GA 30303, USA.

Derivation of derivatives of the estimating function

Define

$$\begin{aligned}
S^{(2)}_{\theta\theta}(V_i, y^{o_r}) &= \frac{1}{n}\sum_{j=1}^{n} I(V_j \geq V_i)\mathcal{E}_{y_i^{o_r}}(\mathcal{B}_Y\mathcal{B}_Y^\top | X_j) \otimes X_j X_j^\top \omega_{y_i^{o_r}}(X_j), \\
S^{(2)}_{\xi\xi}(V_i, u) &= \frac{1}{n}\sum_{j=1}^{n} I(V_j \geq V_i)\mathcal{E}_{u}(\mathcal{A}_Y\mathcal{A}_Y^\top | X_j)\omega_{u}(X_j), \\
S^{(2)}_{\theta\xi}(V_i, y^{o_r}) &= \frac{1}{n}\sum_{j=1}^{n} I(V_j \geq V_i)\mathcal{E}_{y_i^{o_r}}(\mathcal{B}_Y\mathcal{A}_Y^\top | X_j) \otimes X_j \omega_{y_i^{o_r}}(X_j), \\
S^{(2)}_{\xi\theta}(V_i, u) &= \frac{1}{n}\sum_{j=1}^{n} I(V_j \geq V_i)\mathcal{E}_{u}(\mathcal{A}_Y\mathcal{B}_Y^\top | X_j) \otimes X_j \omega_{u}(X_j), \\
S^{(1)}_{\xi}(V_i, y^{o_r}) &= \frac{1}{n}\sum_{j=1}^{n} I(V_j \geq V_i)\mathcal{E}_{y_i^{o_r}}(\mathcal{A}_Y | X_j)\omega_{y_i^{o_r}}(X_j), \\
S^{(1)}_{\theta}(V_i, u) &= \frac{1}{n}\sum_{j=1}^{n} I(V_j \geq V_i)\mathcal{E}_{u}(\mathcal{B}_Y | X_j) \otimes X_j \omega_{u}(X_j),
\end{aligned}$$

and let $s^{(2)}_{\theta\theta}(V_i, y^{o_r})$, $s^{(2)}_{\xi\xi}(V_i, u)$,$s^{(2)}_{\theta\xi}(V_i, y^{o_r})$,$s^{(2)}_{\xi\theta}(V_i, u)$, $s^{(1)}_{\xi}(V_i, y^{o_r})$ and $s^{(1)}_{\theta}(V_i, u)$ denote the corresponding population expectations. Now we can write

$$\begin{aligned}
\frac{\partial S_\theta}{\partial \theta^T} &= \sum_r \sum_{\Delta_i=1, R_i=r} \left[\mathcal{V}_{y_i^{o_r}}(\mathcal{B}_Y | X_i) \otimes X_i X_i^\top - \frac{S^{(2)}_{\theta\theta}(V_i, y^{o_r})}{S^{(0)}(V_i, y^{o_r})} + \frac{S^{(1)}_{\theta}(V_i, y^{o_r})}{S^{(0)}(V_i, y^{o_r})} \left\{ \frac{S^{(1)}_{\theta}(V_i, y^{o_r})}{S^{(0)}(V_i, y^{o_r})} \right\}^\top \right], \\
\frac{\partial S_\theta}{\partial \xi^T} &= \sum_r \sum_{\Delta_i=1, R_i=r} \left[\mathcal{C}_{y_i^{o_r}}(\mathcal{B}_Y, \mathcal{A}_Y | X_i) \otimes X_i - \frac{S^{(2)}_{\theta\xi}(V_i, y^{o_r})}{S^{(0)}(V_i, y^{o_r})} + \frac{S^{(1)}_{\theta}(V_i, y^{o_r})}{S^{(0)}(V_i, y^{o_r})} \left\{ \frac{S^{(1)}_{\xi}(V_i, y^{o_r})}{S^{(0)}(V_i, y^{o_r})} \right\}^\top \right], \\
\frac{\partial S_\xi}{\partial \xi^T} &= \sum_r \sum_{\Delta_i=1, R_i=r} \left[\mathcal{V}_{y_i^{o_r}}(\mathcal{A}_Y | X_i) - \frac{S^{(2)}_{\xi\xi}(V_i, u)}{S^{(0)}(V_i, u)} + \frac{S^{(1)}_{\xi}(V_i, u)}{S^{(0)}(V_i, u)} \left\{ \frac{S^{(1)}_{\xi}(V_i, u)}{S^{(0)}_{\xi}(V_i, u)} \right\}^\top \right], \\
\frac{\partial S_\xi}{\partial \theta^T} &= \sum_r \sum_{\Delta_i=1, R_i=r} \left[\mathcal{C}_{y_i^{o_r}}(\mathcal{A}_Y, \mathcal{B}_Y | X_i) \otimes X_i - \frac{S^{(2)}_{\theta\xi}(V_i, u)}{S^{(0)}(V_i, u)} + \frac{S^{(1)}_{\xi}(V_i, u)}{S^{(0)}(V_i, u)} \left\{ \frac{S^{(1)}_{\theta}(V_i, u)}{S^{(0)}(V_i, u)} \right\}^\top \right],
\end{aligned}$$

where $\mathcal{V}_{y_i^{o_r}}$ and $\mathcal{C}_{y_i^{o_r}}$ denote variances and covariances with respect to the conditional distribution $Q_{y_o^r}^{y_m^r}$. Thus, the components of the matrix $\mathcal{I} = \lim_{n\to\infty}(1/n)\partial T_n/\partial\eta$ can now be obtained by replacing S by s throughout the above equations and then taking the expectations of the corresponding i.i.d. sums.

Proof of the asymptotic unbiasedness of the estimating equation under general missing at random assumption

In this section, we prove asymptotic unbiasedness of the estimating equations $S_\theta = 0$ and $S_\xi = 0$ under the general missing at random mechanism specified by equation (6).

First, it is easy to see that the asymptotic limit of $(1/n)S_\theta^{(r)}$ can be written in general form as

$$E_{R,V,\Delta,Y_o^r}\left[I(R=r)\Delta\left\{\frac{\partial}{\partial\theta}\log h_{Y^{o_r}}(V|X)-\frac{s^{(1)}(V,Y^{o_r})}{s^{(0)}(V,Y^{o_r})}\right\}\right],$$

where

$$s^{(1)}(V,Y^{o_r})=E_{V',X'}I(V'\geq V)\left\{\partial\log h_{Y^{o_r}}(V'|X')/\partial\theta\right\}h_{Y^{o_r}}(V'|X'),$$

$s^{(0)}(V,Y^{o_r})=E_{V',X'}I(V'\geq V)h_{Y^{o_r}}(V'|X')$, and $\partial\log h_{Y^{o_r}}(V|X)/\partial\theta=\mathcal{E}_{Y^{o_r}}(\mathcal{B}_y|V,X)\otimes X$. Now, we can write

$$\begin{aligned}
C^{(r)} &\equiv E\Delta I(R=r)\partial\log h_{Y^{o_r}}(V|X)/\partial\theta\\
&= E_X E_{\Delta,V,Y^{o_r}|X}\Delta\pi^{(r)}(V,X)\partial\log h_{Y^{o_r}}(V|X)/\partial\theta \quad \text{(assuming missing-at-random)}\\
&= E_X\int\pi^{(r)}(v,X)\left\{\partial\log h_{y^{o_r}}(v|X)/\partial\theta\right\}\text{pr}(\Delta=1,v,y^{o_r}|X)dvd\mu(y^{o_r})\\
&= E_X\int\pi^{r}(v,X)\left\{\partial\log h_{y^{o_r}}(v|X)/\partial\theta\right\}h_{y^{o_r}}(v|X)E(V\geq v|X)dvd\mu(y^{o_r})\\
&= \int_{v,y^{o_r}}E_{V,X}\left[I(V\geq v)\pi^{(r)}(v,X)\left\{\partial\log h_{y^{o_r}}(v|X)/\partial\theta\right\}h(v,y^{o_r}|X)\right]dvd\mu(y^{o_r}).
\end{aligned}$$

Further, we write

$$\begin{aligned}
D^{(r)} &\equiv E\left\{\Delta I(R=r)\frac{s^{(1)}(V,Y^{o_r})}{s^{(0)}(V,Y^{o_r})}\right\}\\
&= E_{\Delta,V,Y^{o_r}}\frac{\Delta s^{(1)}(V,Y^{o_r})\text{pr}(R=r|\Delta=1,V,Y^{o_r})}{\text{pr}(\Delta=1,V,Y^{o_r})}
\end{aligned}$$

where the last equality follows because

$$\begin{aligned}
\text{pr}(\Delta=1,V,Y^{o_r}) &= E_{X'}h_{Y^{o_r}}(v|X')E\left\{I(V'\geq V)|X'\right\}\\
&= E_{V',X'}h_{Y^{o_r}}(V'|X')I(V'\geq V)=s^{(0)}(V,Y^{o_r}).
\end{aligned}$$

Moreover, under missing at random assumption (6), we can write

$$\begin{aligned}
\text{pr}(R=r|\Delta=1,V,Y^{o_r}) &= \int\pi^{(r)}(V,x)\text{pr}(x|\Delta=1,V,Y^{o_r})dx\\
&= \frac{E_{X'}\pi^{(r)}(V,X')\text{pr}(\Delta=1,V,Y^{o_r}|X')}{\text{pr}(\Delta=1,V,Y^{O_r})}\\
&= \frac{E_{V',X'}I(V'\geq V)\pi^{(r)}(V,X')h_{Y^{o_r}}(V|X')}{s^{(0)}(V,Y^{o_r})}.
\end{aligned}$$

Thus, we can write

$$D^{(r)}=\int\frac{s^{(1)}(v,y^{o_r})}{s^{(0)}(v,y^{o_r})}E_{V,X}\left\{I(V\geq v)\pi^{r}(v,X)h_{y^{o_r}}(v|X)\right\}dvd\mu(y^{o_r}).$$

Now, we note that, if $\pi^{(r)}(T, X) \equiv \pi^{(r)}(T)$, then we have

$$C^{(r)} = D^{(r)} = \int \pi^{(r)}(v) s^{(1)}(v, y^{o_r}) dv d\mu(y^{o_r}),$$

implying asymptotic unbiasedness of $S_\theta^{(r)}$ for each specific r.

When $\pi^{(r)}(T, X)$ depends on X, in general $C^{(r)} \neq D^{(r)}$, however, $\sum_r C^{(r)} = \sum^r D^{(r)}$. To see this, note that by rearrangement of integrals and expectation, we can write

$$C^{(r)} = \int_v E_{V,X} \left[I(V \geq v) \pi^{(r)}(v, X) \left\{ \partial h_u(v|X)/\partial\theta \right\} \right] dv,$$

where

$$\partial h_u(v|X)/\partial\theta = \int_{y^{o_r}} \left\{ \partial h_{y^{o_r}}(v|X)/\partial\theta \right\} d\mu(y^{o_r}) = \int_y \left\{ \partial \log h_y(v|X)/\partial\theta \right\} h_y(v|X) d\mu(y).$$

Now summing over r and using the constraint $\sum_r \pi^{(r)}(v, x) = 1$ we can write

$$C = \sum_r C^{(r)} = \int_v E_{V,X} \left[I(V \geq v) \left\{ \partial h_u(v|X)/\partial\theta \right\} \right] dv,$$

which again after rearrangement of integrals can be written as

$$C = \int_{v,y} s^{(1)}(v, y) dv d\mu(y).$$

Now define

$$\eta(v|X) = \int_{y^{o_r}} \frac{s^{(1)}(v, y^{o_r})}{s^{(0)}(v, y^{o_r})} h_{y^{o_r}}(v|X) d\mu(y^{o_r}) = \int_y \frac{s^{(1)}(v, y)}{s^{(0)}(v, y)} h_y(v|X) d\mu(y)$$

and note that we can write

$$D^{(r)} = \int_v E_{V,X} \left[I(V \geq v) \pi^r(v, X) \eta(v|X) \right] dv.$$

Thus, we have

$$\begin{aligned} D = \sum_r D^{(r)} &= \int_v E_{V,X} \left[I(V \geq v) \eta(v|X) \right] dv \\ &= \int_{v,y} \frac{s^{(1)}(v, y)}{s^{(0)}(v, y)} E_{V,X} I(V \geq v) h_y(v|X) dv d\mu(y) = \int_{v,y} s^{(1)}(v, y) dv d\mu(y), \end{aligned}$$

and thus the equality $C = D$ is proven. The proof of the asymptotic unbiasedness of S_ξ can be obtained following the similar steps with the observation that

$$\int_v s^{(1)}(v, u) = \int_{v,y} s^{(1)}(v, y).$$

Table 1: *Results of the simulation study, where the disease has 64 subtypes based on 3 disease traits each with 4 levels. The true value of* $\theta^{(0)} = 0.35$, $\theta_1^{(1)} = 0.15$, $\theta_2^{(1)} = 0$, *and* $\theta_3^{(1)} = 0.5$. *The working model for the baseline hazards is misspecified. Each of the disease traits is missing-completely-at-random with probability 0.20 or 0.30. Estvar and 95%CP stand for means of estimated variances and 95% coverage probability, respectively, over the different simulations.*

Method		$\theta^{(0)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta^{(0)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$
		n=5,000				n=10,000			
Full-cohort	Bias($\times 10^2$)	-0.76	-0.56	0.90	0.03	0.34	-0.29	-0.12	-0.02
	Var($\times 10^2$)	2.24	0.72	0.67	0.83	1.13	0.33	0.39	0.41
	Etvar($\times 10^2$)	2.26	0.72	0.66	0.82	1.24	0.37	0.38	0.43
	95% CP	94.2	94.8	94.8	94.4	96.0	96.0	95.4	95.0
		20% missing				20% missing			
Complete-case	Bias($\times 10^2$)	1.83	0.58	0.35	1.96	3.16	0.11	-0.15	1.82
	Var($\times 10^2$)	4.28	1.46	1.35	1.56	3.33	0.90	1.07	0.92
	Etvar($\times 10^2$)	4.68	1.49	1.36	1.69	3.72	1.08	1.21	1.28
	95% CP	95.6	95.4	96.2	94.8	92.8	96.0	95.0	94.6
Estimating-equation	Bias($\times 10^2$)	0.86	-0.91	0.66	-0.51	1.69	-0.59	-0.70	-0.36
	Var($\times 10^2$)	2.62	0.85	0.86	0.98	1.37	0.43	0.52	0.51
	Etvar($\times 10^2$)	2.64	0.88	0.81	0.99	1.47	0.46	0.47	0.52
	95% CP	94.6	95.4	94.4	94.4	95.6	95.0	93.6	94.8
		30% missing				30% missing			
Complete-case	Bias($\times 10^2$)	0.63	1.49	0.68	2.78	3.95	0.93	-0.37	3.17
	Var($\times 10^2$)	7.21	2.42	2.24	2.48	3.25	1.30	1.26	1.34
	Etvar($\times 10^2$)	7.26	2.32	2.11	2.63	3.43	1.18	1.12	1.30
	95% CP	94.2	95.6	93.8	95.2	95.6	94.4	95.2	94.6
Estimating-equation	Bias($\times 10^2$)	1.25	-1.14	0.75	-0.62	0.44	-0.16	-0.17	0.02
	Var($\times 10^2$)	2.96	0.97	0.99	1.12	1.30	0.49	0.50	0.57
	Etvar($\times 10^2$)	2.92	1.00	0.91	1.11	1.45	0.51	0.49	0.56
	95% CP	94.6	96.2	93.7	93.8	95.8	95.0	94.6	95.0

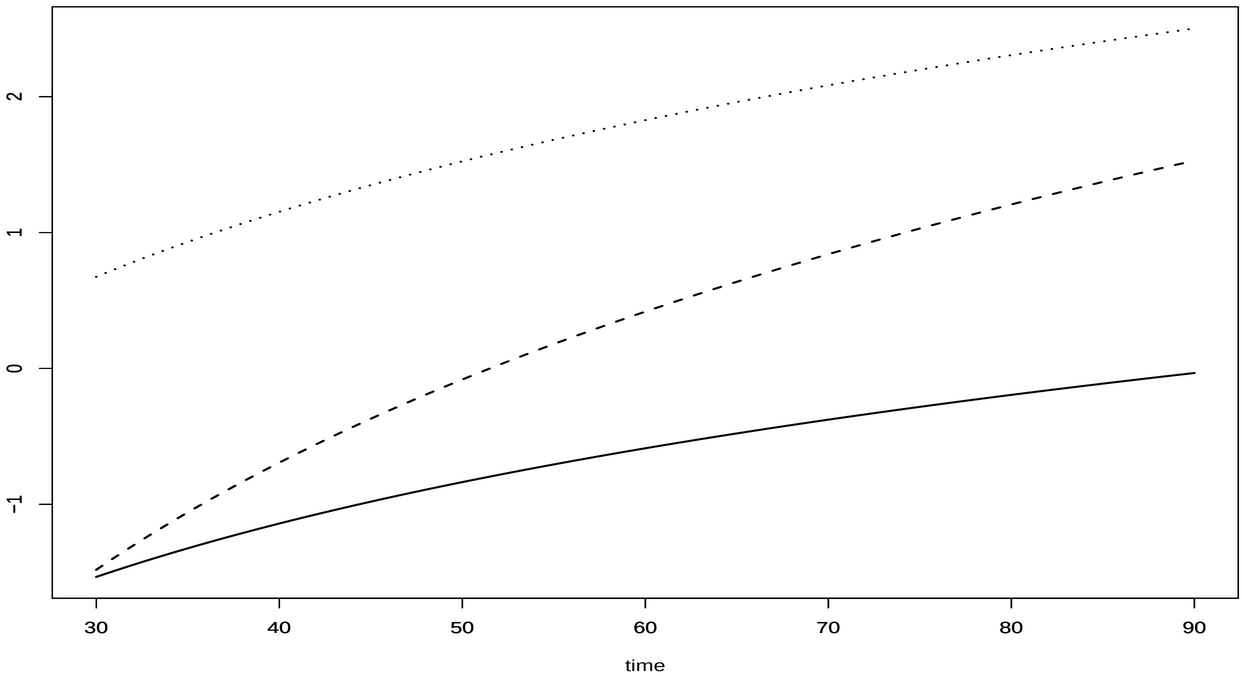

Figure 1: Plot of the $\log\{\lambda_{(y_1,y_2)}(t)/\lambda_{(1,1)}(t)\}$. The solid line (—), dashed line ($--$), and dotted line ($\cdots$) correspond to $(y_1, y_2) = (1, 2)$, $(2, 1)$ and $(2, 2)$, respectively.