

# Web-based Supplementary Materials for “Cox Regression in Nested Case-Control Studies with Auxiliary Covariates”

by Liu, Lu and Tseng

## Web Appendix A: Proof of Proposition 1

The following regularity conditions are assumed for our results.

- (a). the baseline hazard function  $\lambda_0(t)$  is bounded away from 0 and infinity on  $[0, \tau]$ ;
- (b).  $\text{pr}\{Y(t) = 1\} \geq 0$  for  $\forall t \in [0, \tau]$ ;
- (c). covariate process  $Z(t)$  and  $X(t)$  are bounded on  $[0, \tau]$ ;
- (d). there exists a neighborhood  $\mathcal{G}$  of  $\gamma_*$  such that for  $k = 0, 1, 2$ ,

$$\sup_{t \in [0, \tau], \gamma \in \mathcal{G}} \|S^{(k)}(t; \gamma) - s^{(k)}(t; \gamma)\| \rightarrow 0 \text{ and } \sup_{t \in [0, \tau]} \|S^{(k)}(t) - s^{(k)}(t)\| \rightarrow 0,$$

in probability;  $s^{(k)}(t; \gamma)/s^{(0)}(t; \gamma)$ ,  $k = 1, 2$  are bounded on  $\mathcal{G} \times [0, \tau]$ ;

- (e).  $\Gamma$ ,  $I$  and  $A$  are positive definite,

Assumptions (a)-(c) are analogous to Conditions 2-4 in Goldstein and Langholz(1992). Assumption (d) is the same as in Lin and Wei (1989). Assumption (e) is needed for the consistency and asymptotic normality of  $\hat{\beta}$ ,  $\hat{\gamma}$  and  $\tilde{\gamma}$  as in Goldstein and Langholz (1992), Xiang and Langholz (2003) and Lin and Wei (1989), respectively.

*Proof of Proposition 1.*

Under regularity conditions, Taylor expansion of  $U_Z(\hat{\beta})$  around  $\beta_0$  yields  $n^{1/2}(\hat{\beta} - \beta_0) = \Gamma^{-1}n^{-1/2}U_Z(\beta_0) + o_p(1)$  (Goldstein and Langholz, 1992). Similarly, we have  $n^{1/2}(\hat{\gamma} - \gamma_*) = I^{-1}n^{-1/2}U_X(\gamma_*) + o_p(1)$  (Xiang and Langholz, 2003) and  $n^{1/2}(\tilde{\gamma} - \gamma_*) = I^{-1}n^{-1/2}\tilde{U}(\gamma_*) + o_p(1)$  (Lin and Wei, 1989). Thus, it suffice to study the joint distribution of  $U_Z(\beta_0)$ ,  $U_X(\gamma_*)$  and  $\tilde{U}(\gamma_*)$ . First, it is easy to see that

$$\Delta = \text{cov}\{n^{1/2}(\hat{\beta} - \beta_0), n^{1/2}(\hat{\gamma} - \tilde{\gamma})'\}$$

$$\begin{aligned}
&= \text{cov}\{n^{1/2}(\hat{\beta} - \beta_0), n^{1/2}(\hat{\gamma} - \gamma_*)'\} - \text{cov}\{n^{1/2}(\hat{\beta} - \beta_0), n^{1/2}(\tilde{\gamma} - \gamma_*)'\} \\
&= \Gamma^{-1} \left[ n^{-1} \text{cov}\{U_Z(\beta_0), U'_X(\gamma_*)\} \right] I^{-1} - \Gamma^{-1} \left[ n^{-1} \text{cov}\{U_Z(\beta_0), \tilde{U}'(\gamma_*)\} \right] A^{-1}, \quad (\text{A.1})
\end{aligned}$$

and

$$\begin{aligned}
\Omega &= \text{Var}\{n^{1/2}(\hat{\gamma} - \tilde{\gamma})\} \\
&= \text{Var}\{n^{1/2}(\hat{\gamma} - \gamma_*)\} + \text{Var}\{n^{1/2}(\tilde{\gamma} - \gamma_*)\} - 2\text{cov}\{n^{1/2}(\hat{\gamma} - \gamma_*), n^{1/2}(\tilde{\gamma} - \gamma_*)'\} \\
&= I^{-1}VI^{-1} + A^{-1}BA^{-1} - 2I^{-1} \left[ n^{-1} \text{cov}\{U_X(\gamma_*), \tilde{U}'(\gamma_*)\} \right] A^{-1}. \quad (\text{A.2})
\end{aligned}$$

Let  $\mathcal{F}_i(t)$  be the  $\sigma$ -algebra of the failure, censoring and true and auxiliary covariate information for subject  $i$  on  $[0, t]$ . Goldstein and Langholz (1992) showed that the score function of Thomas' estimator can be rewritten as

$$\begin{aligned}
n^{-1/2}U_Z(t; \beta_0) &= n^{-1/2} \sum_{i=1}^n \int_0^t \{Z_i(u) - E_{Z, \tilde{R}_i}(u; \beta_0)\} dM_i(u) + o_p(1), \\
&= n^{-1/2}U_{Z,M}(t, \beta_0) + o_p(1)
\end{aligned}$$

where  $M_i(u) = N_i(u) - \int_0^u Y_i(s) d\Lambda_i(s) = N_i(u) - \int_0^u Y_i(s) e^{\beta_0' Z_i(s)} \lambda_0(s) ds$  is an  $\mathcal{F}_i(u)$ -martingale. Note that  $U_{Z,M}(t, \beta_0)$  is an integral of martingale and thus itself is also a martingale. It has been shown that  $n^{-1/2}U_Z(\tau, \beta_0) \rightarrow N(0, \Gamma^{-1})$  (Goldstein and Langholz, 1992, where  $\tau$  was set to be 1 for convenience).

Next, we rewrite the score function  $U_X(t; \gamma_*)$  under the working model as

$$\begin{aligned}
U_X(t; \gamma_*) &= \sum_{i=1}^n \int_0^t \{X_i(u) - E_{X, \tilde{R}_i}(u, \gamma_*)\} dM_i(u) \\
&\quad + \sum_{i=1}^n \int_0^t \{X_i(u) - E_{X, \tilde{R}_i}(u, \gamma_*)\} Y_i(u) d\Lambda_i(u) \\
&= U_{X,M}(t; \gamma_*) + U_{X,P}(t; \gamma_*), \quad (\text{A.3})
\end{aligned}$$

where  $U_{X,M}(t; \gamma_*)$  is an  $\mathcal{F}(t)$ -martingale and  $U_{X,P}(t; \gamma_*)$  is an  $\mathcal{F}(t)$ -predictable process. Thus,  $U_{X,M}(t; \gamma_*)$  and  $U_{X,P}(t; \gamma_*)$  are  $\mathcal{F}$ -conditionally uncorrelated and so are  $U_{Z,M}(t; \beta_0)$  and  $U_{X,P}(t; \gamma_*)$ . Furthermore, by the martingale central limit theorem (Andersen et al, 1993), Cramer-Wold device (Serfling, 1980), and the martingale independence condition of

$\langle M_i, M_j \rangle(t) = 0$  for  $i \neq j$ ,  $n^{-1/2}\{U'_Z(t; \beta_0), U'_X(t; \gamma_*)\}'$  converges to a Gaussian process.

The covariance process is given by

$$\begin{aligned} n^{-1}\text{cov}\{U_Z(t; \beta_0), U'_X(t; \gamma_*)\} &= n^{-1}\text{cov}\{U_{Z,M}(t; \beta_0), U'_{X,M}(t; \gamma_*)\} + o_p(1) \\ &= n^{-1} \sum_{i=1}^n \int_0^t \{Z_i(u) - E_{Z, \tilde{R}_i}(u; \beta_0)\} \{X_i(u) - E_{X, \tilde{R}_i}(u; \gamma_*)\}' Y_i(u) d\Lambda_i(u) + o_p(1). \end{aligned} \quad (\text{A.4})$$

Then by the similar proof of Lemma 1 in the supplemental material of Xiang and Langholz (2003), we can show that (A.4) converges to

$$K_1(t) = \int_0^t P_Y(u) \mathbb{E} \left( m^{-1} \sum_{i \in r} \{Z_i(u) - E_{Z,r}(u; \beta_0)\} \{X_i(u) - E_{X,r}(u; \gamma_*)\}' \lambda_i(u) \mid Y_r(u) = 1 \right) du.$$

In a similar fashion, the score function for the full-cohort estimator under the working model can be rewritten as

$$\begin{aligned} \tilde{U}(t; \gamma_*) &= \sum_{i=1}^n \int_0^t \{X_i(u) - \bar{x}(u; \gamma_*)\} dM_i(u) + \sum_{i=1}^n \{X_i(u) - \bar{X}(u; \gamma_*)\} Y_i(u) d\Lambda_i(u) + o_p(n^{1/2}) \\ &= \tilde{U}_M(t; \gamma_*) + \tilde{U}_P(t; \gamma_*) + o_p(n^{1/2}). \end{aligned} \quad (\text{A.5})$$

Again,  $\tilde{U}_M(t; \gamma_*)$  is a martingale and  $\tilde{U}_P(t; \gamma_*)$  is an  $\mathcal{F}(t)$ -predictable process. Thus, we have  $n^{-1/2}\{U_Z(t; \beta_0)', \tilde{U}(t; \gamma_*)'\}'$  converge to a Gaussian process with the limiting covariance process given by

$$\begin{aligned} n^{-1}\text{cov}\{U_Z(t; \beta_0), \tilde{U}'(t; \gamma_*)\} &= n^{-1}\text{cov}\{U_{Z,M}(t; \beta_0), \tilde{U}'_M(t; \gamma_*)\} + o_p(1) \\ &= n^{-1} \sum_{i=1}^n \int_0^t \{Z_i(u) - E_{Z, \tilde{R}_i}(u; \beta_0)\} \{X_i(u) - \bar{x}(u; \gamma_*)\}' Y_i(u) d\Lambda_i(u) + o_p(1) \\ &\rightarrow \int_0^t P_Y(u) \mathbb{E} \left( m^{-1} \sum_{i \in r} \{Z_i(u) - E_{Z,r}(u; \beta_0)\} \{X_i(u) - \bar{x}(u; \gamma_*)\}' \lambda_i(u) \mid Y_r(u) = 1 \right) du \\ &= K_2(t). \end{aligned}$$

The convergence also follows Lemma 1 in Xiang and Langholz (2003). Let  $K_1 = K_1(\tau)$  and  $K_2 = K_2(\tau)$ . Plugging in (A.1), we have  $\Delta = \Gamma^{-1}K_1I^{-1} + \Gamma^{-1}K_2A^{-1}$ .

Finally, by (A.3) and (A.5),  $n^{-1}\text{cov}\{U_X(t; \gamma_*), \tilde{U}'(t; \gamma_*)\}$  can be decomposed into

$$n^{-1}\text{cov}\{U_{X,M}(t; \gamma_*), \tilde{U}'_M(t; \gamma_*)\} + n^{-1}\text{cov}\{U_{X,P}(t; \gamma_*), \tilde{U}'_P(t; \gamma_*)\}.$$

Follow similar arguments leading to  $K_1(t)$  and  $K_2(t)$ , it is easy to show that

$$n^{-1} \text{cov}\{U_{X,M}(t; \gamma_*), \tilde{U}'_M(t; \gamma_*)\} \rightarrow \Sigma_1(t),$$

where

$$\Sigma_1(t) = \int_0^t P_Y(u) \mathbb{E} \left( m^{-1} \sum_{i \in r} \{X_i(u) - E_{X,r}(u; \gamma_*)\} \{X_i(u) - \bar{x}(u; \gamma_*)\}' \lambda_i(u) \mid Y_r(u) = 1 \right) du.$$

Next,

$$\begin{aligned} n^{-1} \text{cov}\{U_{X,P}(t; \gamma_*), \tilde{U}'_P(t; \gamma_*)\} &= n^{-1} \text{cov} \left[ \sum_{i=1}^n \int_0^t \{X_i(u) - E_{X, \tilde{R}_i}(u; \gamma_*)\}' Y_i(u) \lambda_i(u) du, \right. \\ &\quad \left. \sum_{i=1}^n \int_0^t \{X_i(v) - \bar{X}(v; \gamma_*)\}' Y_i(v) \lambda_i(v) dv \right] \\ &= \int_0^t \int_0^t n \text{cov} \left[ n^{-1} \binom{|\mathcal{R}(u)|-1}{m-1}^{-1} \sum_{w \subset \mathcal{R}(u), |w|=m} F^{(1)}(u; w), \right. \\ &\quad \left. n^{-1} \sum_{i=1}^n \{X_i(v) - \bar{X}(v; \gamma_*)\}' Y_i(v) \lambda_i(v) \right] dudv \end{aligned} \quad (\text{A.6})$$

where  $F^{(1)}(u; w) = \sum_{i \in w} \{X_i(u) - E_{X,w}(u; \gamma_*)\} \lambda_i(u)$ , the set of  $\mathcal{R}(u)$  denotes the indices of at-risk subjects at time  $u$ , and for a set  $w$ , the size of the set is denoted by  $|w|$  throughout.

The last equation follows from the conditional selection probability

$$\mathbb{E}\{I(w = \tilde{R}_i) | \mathcal{F}(u)\} = \binom{|\mathcal{R}(u)|-1}{m-1}^{-1} Y_w(u) I(i \in w).$$

Moreover, let

$$S_n(u) = n^{-1} \binom{|\mathcal{R}(u)|-1}{m-1}^{-1} \sum_{w \subset \mathcal{R}(u), |w|=m} F^{(1)}(u; w).$$

Xiang and Langholz (2003) showed that  $S_n(u)$  is asymptotically equivalent to  $Q_n(u)$ , where

$$Q_n(u) = P_Y^{-m+1}(u) \binom{n}{m}^{-1} m^{-1} \sum_{w: |w|=m, w \in P_m^n} Y_w(u) F^{(1)}(u; w),$$

where  $P_m^n$  be the set of all subsets of size  $m$  from  $\{1, \dots, n\}$ . Thus, we can replace  $S_n(u)$

with  $Q_n(u)$  in (A.6) and rewrite  $n^{-1} \text{cov}\{U_{X,P}(t; \gamma_*), \tilde{U}'_P(t; \gamma_*)\}$  into

$$\int_0^t \int_0^t n \text{cov} \left[ Q_n(u), n^{-1} \sum_{i=1}^n \{X_i(v) - \bar{x}(v; \gamma_*)\}' Y_i(v) \lambda_i(v) \right] dudv \quad (\text{A.7})$$

$$- \int_0^t \int_0^t n \text{cov} \left[ Q_n(u), \{\bar{X}(v; \gamma_*) - \bar{x}(v; \gamma_*)\}' s^{(0)}(v) \right] dudv + o_p(1) \quad (\text{A.8})$$

First, in (A.7), it is easy to see that, for  $i \in \{1, \dots, n\}$  and  $w \in P_m^n$ , the covariance term  $\text{cov}\left[Y_w(u)F^{(1)}(u; w), \{X_i(v) - \bar{x}(v; \gamma_*)\}Y_i(v)\lambda_i(v)\right]$  is non-zero only when  $i \in w$ , and the probability of  $\text{pr}(i \in w) = m/n$ . Following the proof of Lemma 2 in Xiang and Langholz (2003), the integrand of (A.7) converges to

$$P_Y^{-m+1}(u)\text{cov}\left[F^{(1)}(u; r), \{X_1(v) - \bar{x}(v; \gamma_*)\}'Y_1(v)\lambda_1(v)\right]. \quad (\text{A.9})$$

Next, in (A.8), the Taylor expansion yields that

$$\bar{X}(v; \gamma_*) - \bar{x}(v; \gamma_*) = \frac{1}{s^{(0)}(v; \gamma_*)}n^{-1}\sum_{i=1}^n Y_i(v)e^{\gamma_*'X_i(v)}\{X_i(v) - \bar{x}(v; \gamma_*)\} + o_p(1).$$

Thus, by the same arguments leading to (A.9), the integrand of (A.8) converges to

$$P_Y^{-m+1}(u)\text{cov}\left[F^{(1)}(u; r), Y_1(v)e^{\gamma_*'X_1(v)}\{X_1(v) - \bar{x}(v; \gamma_*)\}'\right]\frac{s^{(0)}(v)}{s^{(0)}(v; \gamma_*)}.$$

This together with (A.9) yields

$$n^{-1}\text{cov}\{U_{X,P}(t; \gamma_*), \tilde{U}'_P(t; \gamma_*)\} \rightarrow \Sigma_2(t),$$

where  $\Sigma_2(t)$  equals

$$\int_0^t \int_0^t P_Y(u)\text{cov}\left[F^{(1)}(u; r), Y_1(v)\{X_1(v) - \bar{x}(v; \gamma_*)\}'\left\{\lambda_1(v) - \frac{e^{\gamma_*'X_1(v)}s^{(0)}(v)}{s^{(0)}(v; \gamma_*)}\right\}\right]_{Y_r(u)=1} dudv.$$

Let  $\Sigma_1 = \Sigma_1(\tau)$ ,  $\Sigma_2 = \Sigma_2(\tau)$ , and together with (A.2), we have that

$$\Omega = I^{-1}VI^{-1} + A^{-1}BA^{-1} - 2I^{-1}(\Sigma_1 + \Sigma_2)A^{-1}.$$

Then by the asymptotic representations of  $n^{1/2}(\hat{\beta} - \beta_0)$ ,  $n^{1/2}(\hat{\gamma} - \gamma_*)$  and  $n^{1/2}(\tilde{\gamma} - \gamma_*)$ , and the finiteness of the matrices  $\Gamma$ ,  $\Delta$  and  $\Omega$ , the results established in Proposition 1 easily follows from the Cramer-Wold device (Serfling, 1980).

## Web Appendix B: Simulation results under common diseases

Consider the scenario S2 described in Section 3.1 and the random censoring  $C \sim U(0, 5)$ . The value of the baseline hazard  $\lambda_0$  is chosen to have the disease incidence rate at 15% and 25%. We consider the number of controls in NCC studies to be 1, 2 or 3. Other simulation

settings are similar to Section 3.1 and the parameter values are specified in the table below. Under all settings, the proposed estimator using the bootstrap method yield satisfactory results with small bias and reasonable coverage probability. Its efficiency gain, comparing to Thomas' estimator and other competing estimators, is similar to what we observe in the rare-disease setting (Tables 1 and 2 in the paper). But it is evident that, when the disease is no longer rare, the rare-disease approximation works poorly producing unsatisfactory coverage probabilities.

[Table 1 about here.]

**Table 1**  
*Simulation results with common disease and random censoring under S2*

Dis. Freq.	$\beta$	$\sigma_\varepsilon$	$m - 1$	Bias	SD	SE	CP	CP*	RE	RE <sup>1</sup>	RE <sup>2</sup>	RE <sup>3</sup>
15%	0	0.5	1	0.001	0.139	0.138	95.0	94.8	65.6	47.9	48.3	53.7
			2	0.006	0.126	0.125	95.2	91.9	80.6	71.5	70.0	77.0
			3	0.005	0.121	0.121	95.2	85.3	86.1	83.1	80.9	87.6
		0.2	1	0.003	0.120	0.120	94.8	92.8	88.4	47.9	48.3	53.7
			2	0.006	0.117	0.115	95.4	86.1	93.0	71.5	70.0	77.0
			3	0.006	0.115	0.114	95.4	77.0	95.3	83.1	80.9	87.6
	-0.5	0.5	1	0.000	0.150	0.147	96.2	94.6	60.8	46.1	46.2	51.6
			2	0.009	0.132	0.131	95.0	91.8	78.5	64.5	64.3	74.6
			3	0.000	0.131	0.126	94.6	88.6	79.9	73.2	74.0	85.1
		0.2	1	0.006	0.127	0.125	94.8	91.2	84.5	46.1	46.2	51.6
			2	0.009	0.122	0.119	93.4	89.2	92.4	64.5	64.3	74.6
			3	0.005	0.122	0.118	94.4	82.1	91.2	73.2	74.0	85.1
	-1	0.5	1	0.009	0.166	0.158	93.6	93.4	52.4	42.4	48.3	51.8
			2	0.009	0.143	0.139	95.2	92.9	70.3	59.7	61.4	70.3
			3	0.012	0.133	0.131	95.0	90.2	81.7	72.2	72.4	83.0
		0.2	1	0.017	0.138	0.129	93.2	92.8	75.3	42.4	48.3	51.8
			2	0.012	0.128	0.123	93.2	93.0	87.1	59.7	61.4	70.3
			3	0.015	0.123	0.120	93.4	89.4	94.7	72.2	72.4	83.0
25%	0	0.5	1	0.007	0.113	0.111	94.8	92.2	64.5	52.3	52.0	59.5
			2	0.010	0.100	0.101	95.8	90.2	82.0	65.7	64.9	75.1
			3	0.007	0.101	0.097	94.8	74.4	81.0	70.6	69.6	84.3
		0.2	1	0.006	0.099	0.096	94.4	90.3	84.1	52.3	52.0	59.5
			2	0.008	0.093	0.093	94.6	80.7	95.2	65.7	64.9	75.1
			3	0.006	0.094	0.092	95.2	67.1	93.6	70.6	69.6	84.3
	-0.5	0.5	1	0.002	0.118	0.115	95.6	95.0	60.6	46.1	47.1	57.4
			2	0.004	0.104	0.103	94.8	91.8	78.0	63.3	63.5	77.2
			3	0.006	0.101	0.099	93.8	79.4	81.5	75.4	74.7	85.9
		0.2	1	0.007	0.101	0.098	95.2	93.0	82.3	46.1	47.1	57.4
			2	0.007	0.096	0.094	93.8	87.8	91.9	63.3	63.5	77.2
			3	0.007	0.095	0.093	95.0	79.0	93.5	75.4	74.7	86.0
	-1	0.5	1	0.001	0.129	0.127	94.4	94.2	55.4	45.1	50.7	57.0
			2	0.005	0.116	0.111	93.5	92.5	68.3	59.6	62.3	75.9
			3	0.005	0.109	0.105	93.7	90.0	76.5	69.4	68.2	83.2
		0.2	1	-0.000	0.110	0.103	93.2	95.0	76.4	45.1	50.7	57.0
			2	0.008	0.104	0.099	94.3	93.3	85.4	59.6	62.3	75.9
			3	-0.001	0.101	0.097	93.5	93.9	89.3	69.4	69.2	83.2

SD: sample standard deviation of the proposed estimates from 500 runs; SE: average standard error estimates of 500 runs; CP: coverage probability of 95% Wald-type confidence interval using the bootstrap method; CP\*: coverage probability of 95% Wald-type confidence interval using the rare-disease asymptotic variance estimator; empirical relative efficiency of each estimator is calculated by the ratio of sample variances of the estimator with that of the full cohort maximum partial likelihood estimator under the true model; RE: relative efficiency of the proposed estimator; RE<sup>1</sup>: relative efficiency of Thomas' estimator; RE<sup>2</sup>: relative efficiency of Chen (2004) estimator; RE<sup>3</sup>: relative efficiency of Samuelsen (1997) estimator.