Estimation under Cox Model with Biased Sampling Data

Ana Elisa Lopez-Miranda, Omerzahid Ali, Sirui Yu

Department of Mathematical and Computational Sciences, University of Toronto Mississauga

Abstract

Prevalent cohort sampling is commonly used to study the natural history of a disease when the disease is rare or it usually takes a long time to observe the failure event. It is known, however, that the collected sample in this situation is not representative of the target population which in turn leads to biased estimates of regression parameters. In this poster, we introduce a composite partial likelihood estimation method for estimating parameters in the Cox proportional hazards model with right-censored and biased sampling data. We also use the method to analyze the dataset of Channing House from a retirement centre in Palo Alto, California.

Objectives

We aim to develop, evaluate, and apply a method that corrects for biased sampling in Cox regression.

- **Method comparison** We compare composite partial likelihood (CPL) estimation method with the typical partial likelihood (PL) estimation method.
- **Simulation study** We run simulations on biased sampling with left truncation and right censoring to see how CPL and PL perform in terms of bias and confidence-interval accuracy.
- **Real-data application** We apply both methods to the Channing House cohort and compare the estimated effects of age and sex on survival.

These objectives guide both our simulation design and the Channing House case study, leading to practical recommendations.

Review

Let T^0 denote the true event time of an individual in the target population. The survival function is defined as

$$S(t) = P(T^0 > t),$$

representing the probability of surviving beyond time t.

The hazard function $h(t)$ quantifies the instantaneous failure rate at time t, conditional on survival up to t:

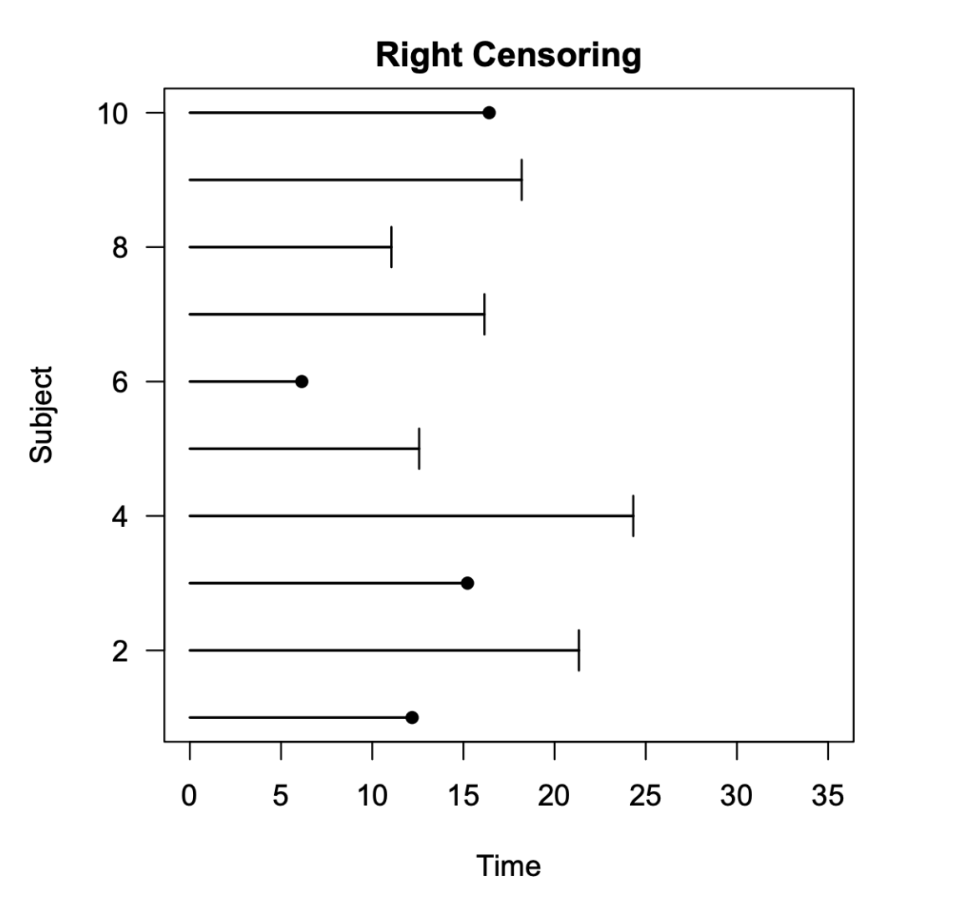
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T^0 < t + \Delta t | T^0 \ge t)}{\Delta t}$$

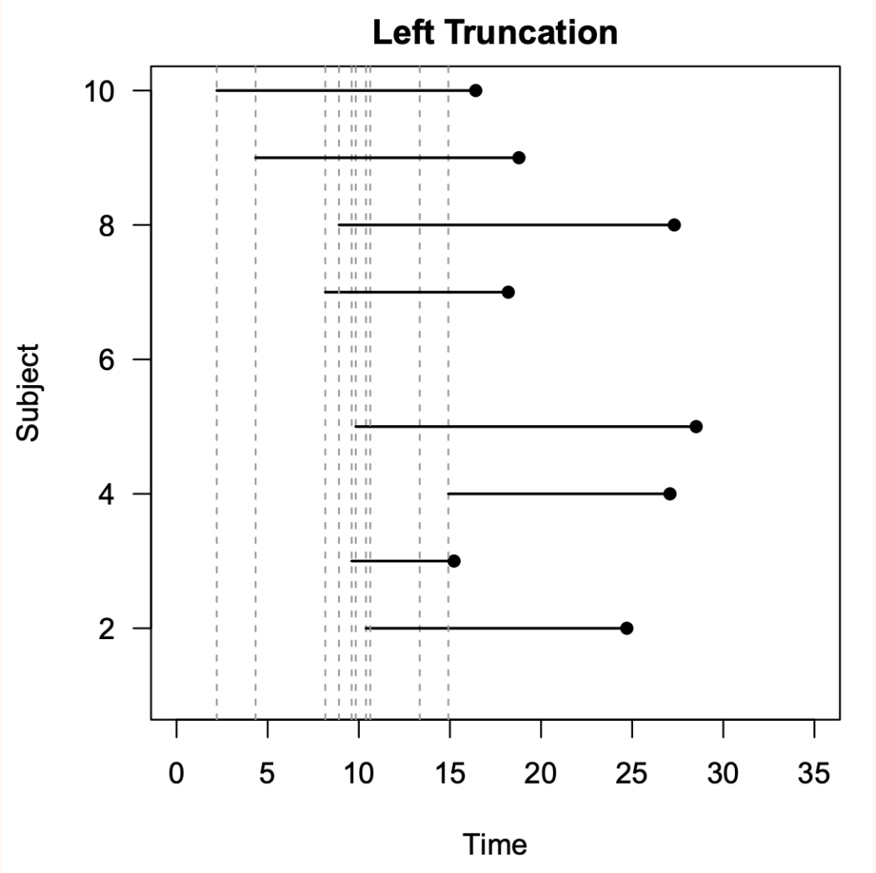
The Cox Proportional Hazards (PH) model describes how the risk of an event happening at time t depends on an individual's covariates X and does not assume any particular shape for the baseline hazard function over time.

$$h(t \mid X) = h_0(t) \exp(X^\top \beta),$$

where $h_0(t)$ denotes the baseline hazard function, which is unspecified and non-parametric.

Right-censoring occurs when all that is known is that the event has not occurred for the individual when the study has ended. Left-truncation occurs when the event for the individual occurred before the study began.





Methods

- **Composite partial likelihood estimation method** extends Cox regression to account for sampling-induced biases by strategically combining likelihood contributions from different data components.
- **Partial likelihood estimation method** was introduced by Sir David Cox (Cox, 1975) [2] which bypasses the need to model the baseline hazard using its semi-parametric properties to maximize the likelihood based on the order of even times and risk sets, leading to consistent and asymptotically normal estimates under mild conditions.

Simulation

The simulation design followed the framework of Huang and Qin (2012) [3], generating length-biased survival data under three hazard shapes–constant, increasing, and U-shaped–and three censoring levels corresponding to approximately 40%, 20%, and 0% censoring. The sample size was set at $N = 400$ for all configurations. For each hazard-censoring combinations, we estimated the regression coefficients using both PL and CPL and evaluated the performance across 1000 replications.

Scenario	Cens	Coef	PL			CPL			
			Bias	SE	ESE	Bias	SE	ESE	RE
Constant	40%	$\hat\beta_1$	0.013	0.102	0.103	0.004	0.072	0.085	1.468
		$\hat\beta_2$	0.030	0.230	0.237	0.010	0.162	0.187	1.606
	20%	$\hat\beta_1$	0.004	0.090	0.089	<0.001	0.064	0.072	1.528
		$\hat\beta_2$	0.014	0.216	0.212	0.007	0.152	0.171	1.537
	0%	$\hat\beta_1$	0.006	0.084	0.086	0.002	0.059	0.068	1.599
		$\hat\beta_2$	0.022	0.209	0.212	0.009	0.147	0.165	1.651
Increasing	40%	$\hat\beta_1$	0.005	0.091	0.088	0.004	0.064	0.078	1.273
		$\hat\beta_2$	0.009	0.183	0.185	0.006	0.129	0.163	1.288
	20%	$\hat\beta_1$	0.011	0.082	0.079	0.008	0.058	0.069	1.311
		$\hat\beta_2$	0.023	0.168	0.175	0.021	0.119	0.153	1.308
	0%	$\hat\beta_1$	0.006	0.075	0.076	0.004	0.053	0.068	1.249
		$\hat\beta_2$	0.009	0.158	0.154	0.004	0.111	0.132	1.361
U-shaped	40%	$\hat\beta_1$	0.007	0.089	0.092	0.004	0.063	0.081	1.290
		$\hat\beta_2$	0.006	0.177	0.176	0.003	0.125	0.159	1.225
	20%	$\hat\beta_1$	0.008	0.079	0.078	0.006	0.056	0.070	1.242
		$\hat\beta_2$	0.018	0.160	0.159	0.014	0.113	0.144	1.219
	0%	$\hat\beta_1$	0.008	0.072	0.073	0.006	0.051	0.066	1.223
		$\hat\beta_2$	0.010	0.147	0.145	0.007	0.104	0.134	1.171

Data Analysis

Channing House was analyzed by applying the proposed methods. During the study 130 women and 46 men died while at Channing House. The data is left truncated as many residents entered at different ages, but were not observed before entry. To use the data, first an analytical stationarity test was applied.

Test Statistic	P-value
0.261	0.794

Since it shows a p-value=0.794, we fail to reject the null hypothesis. Thus, the stationarity assumption for Channing House is valid. To measure the time to death of Channing house residents, a Cox proportional hazards model was fit. The first covariate and second covariate chosen were age upon entry and sex, respectively.

Coef	PL				CPL				
	Estimates	SE	P-values	HR	Estimates	SE	P-values	HR	RE
β_1	-0.042	0.025	0.0994	0.959	-0.150	0.015	0.000	0.861	2.778
β_2	0.344	0.174	0.048	1.41	0.336	0.123	0.006	1.40	2.001

Derivation of Composite Partial Likelihood Estimator

Maximizing the full likelihood $\mathcal{L}_F$ yields efficient estimators but is computationally infeasible. Since length-biased sampling is a special case of truncation, the likelihood can be decomposed into the truncation likelihood of Y given T and the marginal likelihood of A:

$$\mathcal{L}_F = \mathcal{L}_T \times \mathcal{L}_M$$

$$\mathcal{L}_T = \prod_{i=1}^{n} \frac{f(Y_i \mid \mathbf{X}_i)^{\delta_i} S(Y_i \mid \mathbf{X}_i)^{1-\delta_i}}{S(A_i \mid \mathbf{X}_i)}, \quad \mathcal{L}_M = \prod_{i=1}^{n} \frac{S(A_i \mid \mathbf{X}_i)}{\mu(\mathbf{X}_i)}$$

The truncation likelihood factorizes into the partial likelihood and a residual component (Wang et al., 1993) [4]:

$$\mathcal{L}_T = \mathcal{L}_P \times \mathcal{L}_R$$

where $R(t)$ denotes the risk set at time t and

$$\mathcal{L}_P = \prod_{j=1}^{j} \left\{ \frac{\exp(x_j^\top \beta)}{\sum_{k \in R(t_j)} \exp(x_k^\top \beta)} \right\}^{\delta_i}.$$

Wang et al. (1993) [4] show that the maximum partial likelihood estimator is as efficient as the maximum truncation likelihood estimator.

Huang (2012) [3] extends this approach by noting that A (truncation time) and V (residual lifetime $T - A$) have an exchangeable joint density in the absence of censoring. By the composite conditional likelihood method (Strauss, 1988) [1], the conditional density of V given A equals that of A given V, allowing construction of the composite partial likelihood:

$$\prod_{i=1}^{n} \frac{\{\lambda(Y_i)\exp(\beta' X_i)\}^{2\delta_i} \exp\{-(1+\delta_i)\Lambda(Y_i)\exp(\beta' X_i)\}}{\exp\{-\{\Lambda(A_i) + \delta_i \Lambda(\tilde{V}_i)\}\exp(\beta' X_i)\}}.$$

The resulting estimator is consistent and asymptotically normal (Huang, 2012) [3].

Conclusion

Prevalent cohort sampling can bias Cox regression because individuals with longer survival are disproportionately represented. Standard partial likelihood may therefore understate or distort covariate effects. In both our simulations and the Channing House study, the composite partial likelihood reduced this bias, yielding more accurate and efficient estimates with greater stability. However, CPL relies on the assumption that entry times are approximately stationary, so when many subjects enter systematically late, estimates may still be affected; careful checks of entry times and proportional hazards are therefore essential. Overall, CPL offers a practical improvement over PL for analyses based on prevalent cohorts.

References

[1] Barry C Arnold and David Strauss.
Bivariate distributions with exponential conditionals.
Journal of the American Statistical Association, 83(402):522–527, 1988.

[2] David R Cox.
Partial likelihood.
Biometrika, 62(2):269–276, 1975.

[3] Chiung-Yu Huang and Jing Qin.
Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia.
Journal of the American Statistical Association, 107(499):946–957, 2012.

[4] Mei-Cheng Wang, Ron Brookmeyer, and Nicholas P Jewell.
Statistical models for prevalent cohort data.
Biometrics, pages 1–11, 1993.

Acknowledgements

We thank Professor Omidali Aghababei Jazi for his supervision, guidance, and for making this project possible. This work was completed as part of the Research Opportunity Program at the University of Toronto Mississauga.

Undergraduate Research Showcase

a.lopez.miranda@mail.utoronto.ca, omer.ali@utoronto.ca, sirui.yu@mail.utoronto.ca