Analysis of Survival Data with Biased Sampling

Ana Elisa Lopez-Miranda

Summer 2025

Contents

1 Introduction

Survival analysis is a branch of statistics that investigates time to an event. Applications for survival analysis extend across the humanities, sciences, and social sciences. As scientists explore cures, insurance agents explore claims, and economists explore firm bankruptcy, there needs to be a statistical method to explore these questions. Such answers help inform doctors, insurance companies, and families.

A common application of survival analysis is trying to predict how long Alzheimer's patients live after being diagnosed with the disease. Researchers may wish to see what covariates affect survival time. Critical information about age, sex, and location can be found by using methods in survival analysis.

In a prevalent cohort study, subjects are recruited at a fixed time given they already have the condition and are still at risk. For example, for a study that involves seeing when students start to smoke in high school, all high schoolers would be in the prevalent cohort data. However, this technique is inherently biased. In the case of studying a disease, there is a tendency to pick those who live longer. Because only individuals who survive past a certain time can be observed, the data are left-truncated. This causes an overestimation. However, there are techniques to try to reduce the overestimation and make more accurate estimates. The composite partial likelihood method (Huang and Qin 2012) allows one to use maximum likelihood methods on a pseudo-likelihood function. If one assumes onset follows a stationary Poisson process then left truncated data can be assumed to have a uniform distribution (Asgharian, Wolfson, and Zhang 2006). The probability of survival time is proportional to its length and has length-biased distribution. The maximum partial likelihood estimator (Wang, Brookmeyer, and Jewell 1993) is consistent but not efficient (Qin and Shen 2010). Methods based on weighted risk sets cannot be used for covariate-dependent censoring (Huang and Qin 2012). The composite partial likelihood estimator is efficient and can also be used with covariate-dependent censored data.

The rest of the project is structured as follows. Section 2 reviews basic concepts such as parametric distributions and models. It also reviews censoring, truncation, and likelihood functions. Section 3 reviews proportional hazards models. Section 4 explains biased sampling. Section 5 has results. It also includes a simulation comparing the partial likelihood and composite partial likelihood methods. It concludes with data analysis on Channing House (Cutler and Ederer 1958), accessed via the asaur package in R (Kleinbaum and Klein 2012). Section 6 has an overview and conclusion.

2 Review of Basic Concepts

The following subsections use Klein and Moeschberger (2003) as a reference.

Survival Function

A cumulative distribution function $F(t)$ is defined as

$$F(t) = Pr(T \leq t) \tag{1}$$

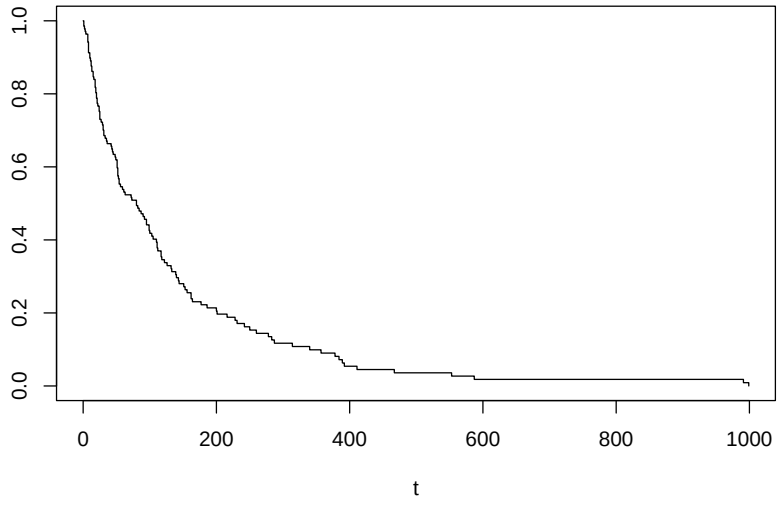
where T is a random variable representing time to an event. In relation to time, this is the accumulation of all the probabilities that have occurred up until time t.

The survival function $S(t)$ is defined as the complement of the cumulative distribution function.

$$S(t) = Pr(T > t) = \int_t^{\infty} f(u)\, du = 1 - Pr(T \leq t) \tag{2}$$

The survival function is the probability of an individual surviving to time t. As such, it makes sense it would be the complement of the probability that the event has occurred up till time t. $S(t) = 1$ when $t = 0$. Conversely, $S(t) = 0$ when $t \to \infty$. It is a non-increasing function.

Figure 1: Survival Function



Hazard Function

The hazard function is the conditional rate of probability that given an individual has not experienced the

event, that they experience the event in the next instant of time. If t is continuous, $h(t)$ can be shown to be

$$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

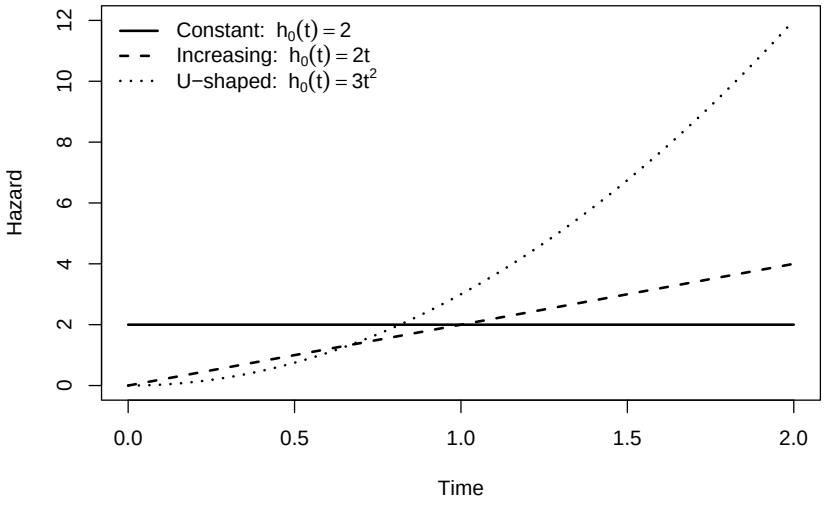
That is, the ratio of the density function over the survival function.

The cumulative hazard function, $H(t)$, is defined as

$$H(t) = \int_0^t h(u)\,du = -\log(S(t)) \tag{4}$$

Hazard functions must be non-negative.

Figure 2: Hazard Functions



2.1 Parametric Distributions

The following are common distributions used in survival analysis along with their hazard functions.

Exponential

For exponential, the density function is $f(t;\lambda) = \lambda \exp[-\lambda t]$. The survival function is $S(t) = \exp[-\lambda t]$ where both $\lambda, t > 0$. The hazard function is constant $h(t) = \lambda$. Since $H(t) = -\log S(t)$ then $H(t) = \lambda t$. Thus, when $H(t)$ is graphed against t, it is a line through the origin with a slope of λ. Further, there is a lack of memory with this parametric model as $E(T - t|T > t) = E(T) = \frac{1}{\lambda}$ which is simply a constant. The current probability does not depend on the past. This is why the hazard function is constant. The mean and standard deviation are $\frac{1}{\lambda}$.

Figure 3: Exponential Density Function (rate=1)

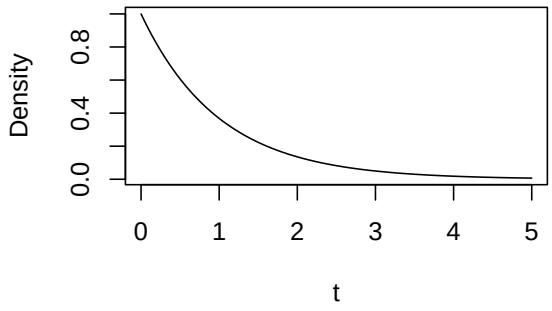
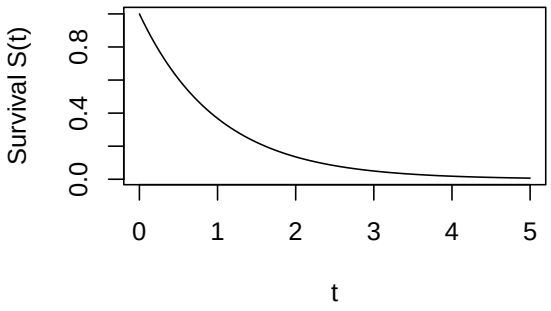


Figure 4: Exponential Survival Function (rate=1)



Weibull

For Weibull, the density function is $f(t;\alpha,\lambda) = \alpha\lambda t^{\alpha-1}\exp[-\lambda t^{\alpha}]$ where $\lambda,\alpha > 0$. Here λ is a scale parameter and α is a shape parameter. The survival function is $S(t;\alpha,\lambda) = \exp[-\lambda t^{\alpha}]$ where both $\lambda, t > 0$. Weibull distributions have hazard functions without a fixed shape: $h(t;\alpha,\lambda) = \alpha\lambda t^{\alpha-1}$. When $\alpha > 1$, the function is increasing. When $\alpha < 1$, the function is decreasing. When $\alpha = 1$, the function is constant. α is thus characterized as the shape parameter.

Figure 5: Weibull Density Function (shape=2, scale=1)

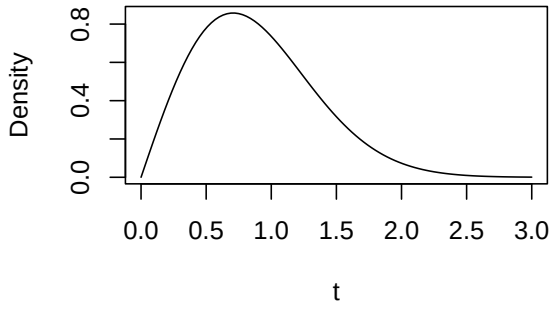
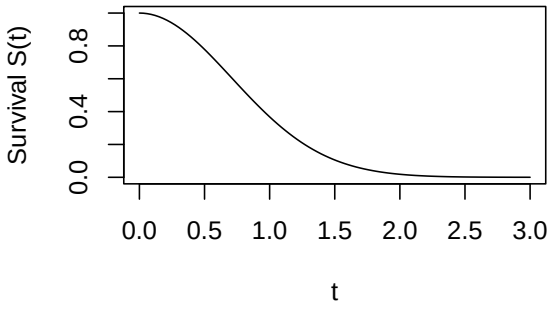


Figure 6: Weibull Survival Function (shape=2, scale=1)



Gamma

For gamma, the density function is

$$f(t;\alpha,\lambda) = \frac{\lambda^{\alpha}\, t^{\alpha-1}\, e^{-\lambda t}}{\Gamma(\alpha)} \tag{5}$$

The survival function is

$$S(t;\alpha,\lambda) = 1 - \frac{\gamma(\alpha,\lambda t)}{\Gamma(\alpha)} \tag{6}$$

The hazard function is the ratio between the density function and the survival function. In gamma distribu-

tions, λ is a scale parameter and α is a shape parameter. As α approaches infinity, the gamma distribution approximates a normal distribution. It also approximates the chi-square distribution when $v = 2\alpha$ and $\lambda = \frac{1}{2}$. The mean and variance are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$.

Figure 7: Gamma(2, 1) Density Function

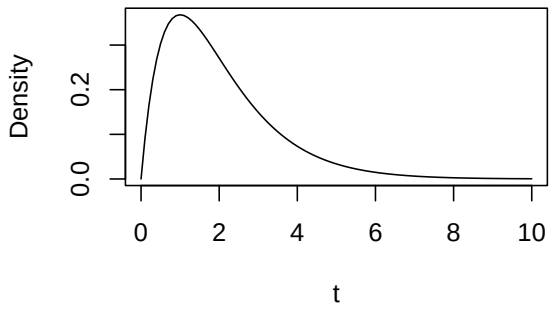
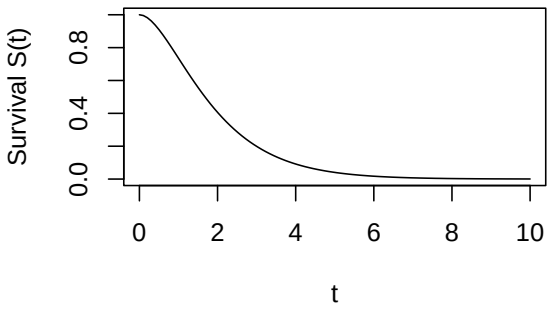


Figure 8: Survival Function



2.2 Censoring

Censoring occurs when only partial information is known. That is, either the beginning or end is unknown for some individual.

Right-censoring occurs when all that is known is that the event has not occurred for the individual when the study has ended. Consider the example of a study that wishes to measure the time-to-event of high school students choosing to smoke. In the example, this would be all the individuals in the study that did not choose to smoke throughout the length of the study. However, they chose to smoke after the study concluded. It is known that $T > t$, but it is unknown the exact t. Right-censoring is described as $C_r = (T, \delta)$ where T represents the random variable time and $\delta = 1$ if the event occurred and $\delta = 0$ if the individual was censored.

Left-censoring occurs when all that is known is that the event occurred for the individual before the study began. Considering the example of a study that wishes to measure the time-to-event of high school students choosing to smoke, this would be all the individuals that chose to start smoking before high school but were included in the study. It is known that $T < t$, but it is unknown the exact t. Left-censoring is described as $C_l = (T, \epsilon)$ where T represents the random variable time and $\epsilon = 1$ if the event occurred and $\epsilon = 0$ if the individual was censored.

Interval-censoring occurs when all that is known is that the event occurred within an interval of time. However, it is not known the exact time the event occurred. Considering the example of a study that wishes

to measure the time-to-event of high school students choosing to smoke, this would be all the individuals that started to smoke at some time t between date A and date B.

There is also random censoring where the end date of the study is not specified beforehand. Individuals drop out according to a random unknown distribution. The individuals should not be dropping out for reasons related to the event.

In all these cases, we only know the range of time T that the event occurred. We do not know the exact $T = t$ the event occurred. To exclude either group of individuals would create a biased sample. Thus, modeling such cases requires further techniques. Either a selective risk set or a conditional distribution must be used in constructing the likelihood.

2.3 Truncation

Truncation occurs when no information is known. The data for such individuals is not in the data set.

Right-truncation occurs when only individuals with $T \leq t$ are observed. Individuals with $T > t$ are not in the study. No information is known about the individuals. Considering the example of a study that wishes to measure the time-to-event of high school students choosing to smoke, this would be all the individuals who chose to smoke after the study concluded, but were never included in the study.

Left-truncation occurs when the event for the individual occurred before the study began. However, since the individual was not in the study, no information is known. Considering the example of a study that wishes to measure the time-to-event of high school students choosing to smoke, this would be all the individuals who chose to smoke before the study began and were never included in the study. When truncated, either a selective risk set or a conditional distribution must be used in constructing the likelihood.

2.4 Likelihood Function

A likelihood function is the probability of obtaining the observed data and is a function of some parameter.

$$L(\theta) = \prod_{i=1}^{n} f(t_i \mid \theta) \tag{7}$$

In order to maximize it, if the function is differentiable, take the derivative of the log, set to 0 and solve it. The likelihood function for right-censored data is

$$L(\theta) = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta i} S(t_i \mid \theta)^{1-\delta i} \tag{8}$$

where $f(t_i \mid \theta)$ is non-censored data ($\delta = 1$) and $S(t_i \mid \theta) = P(T > t_i \mid \theta)$ is censored data ($\delta = 0$).

Taking the log simplifies to

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log f(t_i \mid \theta) + \sum_{i=1}^{n} (1 - \delta_i) \log S(t_i \mid \theta) \tag{9}$$

If censoring was ignored, the mean would be an underestimation.

The likelihood function for left-censored data is

$$L(\theta) = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i} F(t_i \mid \theta)^{1-\delta_i} \tag{10}$$

where $f(t_i \mid \theta)$ is non-censored data ($\delta = 1$) and $F(t_i \mid \theta) = F(T \leq t_i \mid \theta)$ is the CDF.

Taking the log simplifies to

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log f(t_i \mid \theta) + \sum_{i=1}^{n} (1 - \delta_i) \log F(t_i \mid \theta) \tag{11}$$

The likelihood for left-truncated data is

$$L(\theta) = \prod_{i=1}^{n} \frac{f(t_i \mid \theta)}{S(u_i \mid \theta)} \tag{12}$$

where $S(u_i \mid \theta) = P(T > u_i \mid \theta)$.

The likelihood for right-truncated data is

$$L(\theta) = \prod_{i=1}^{n} \frac{f(t_i \mid \theta)}{F(v_i \mid \theta)} \tag{13}$$

3 Regression Models

3.1 Proportional Hazards Model

Proportional hazards models specify the hazard function directly. It models the dependency between the time to an event and covariates. This can be expressed as

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \exp(\delta_j \beta_j) \tag{14}$$

where the difference between the two covariates is δ_j.

If the hazard ratio does not depend on t, then there is a proportional hazard model. Exponential and Weibull regression pass this condition. Thus, the interpretation of regression coefficients in terms of covariates is the hazard ratio for δ-unit change $e^{\delta\beta}$.

Cox Model

The Cox model (Cox 1972) is a semi-parametric regression model. The baseline function, $\lambda_0(t)$, is the underlying risk of the event that is not dependent on the covariates. Sir David Cox said that the hazard function, $\lambda(t \mid X)$ where X are the covariates, could be approximated using the proportional hazard model and the baseline hazard function

$$\lambda_i(t \mid x) = \lambda_0(t) \exp(x_i^T \beta) \tag{15}$$

This model is semi-parametric because it includes both non-parametric term, baseline hazard function, and the parametric term.

Under the proportional hazards assumption, the baseline hazard, $\lambda_0(t)$, cancels out of the partial likelihood (Cox 1972). The Cox partial likelihood under the Cox proportional hazards model is as follows

$$L(\beta) = \prod_j \frac{\exp(\mathbf{x}_j^\top \beta)}{\sum_{k \in R(t_j)} \exp(\mathbf{x}_k^\top \beta)} \tag{16}$$

where x_j denotes the vector of covariates with jth failure times and $R(t)$ is the set of observations at risk at observation time t.

4 Biased Sampling

Biased sampling occurs when information is truncated from the data set. If the incidence rate is constant (called stationary case) then it is called length-biased sampling.

The full likelihood is the following.

$$\mathcal{L}_F = \prod_{i=1}^{n} \frac{f(Y_i \mid \mathbf{X}_i)^{\delta_i} \, S(Y_i \mid \mathbf{X}_i)^{1-\delta_i}}{\mu(\mathbf{X}_i)} \tag{17}$$

Direct maximization of $\mathcal{L}_F$ would lead to efficient estimators but it is too mathematically complex. Since length-biased sampling is a special case of truncated data, another method is to decompose the full likelihood into a combination of the truncation likelihood of Y conditioning on the truncation time A and the marginal

likelihood of A (Wang, Brookmeyer, and Jewell 1993).

$$\mathcal{L}_F = \mathcal{L}_T \times \mathcal{L}_M \tag{18}$$

$$\mathcal{L}_T = \prod_{i=1}^{n} \frac{f(Y_i \mid \mathbf{X}_i)^{\delta_i} \, S(Y_i \mid \mathbf{X}_i)^{1-\delta_i}}{S(A_i \mid \mathbf{X}_i)} \tag{19}$$

$$\mathcal{L}_M = \prod_{i=1}^{n} \frac{S(A_i \mid \mathbf{X}_i)}{\mu(\mathbf{X}_i)} \tag{20}$$

The truncation likelihood can be broken down into a combination of the partial likelihood and the residual likelihood (Wang et al., 1993).

$$\mathcal{L}_T = \mathcal{L}_P \times \mathcal{L}_R \tag{21}$$

where partial likelihood is the following and $R(t)$ is the set of observations at risk at observation time t

$$\mathcal{L}_P = \prod_{j=1}^{j} \left\{ \frac{\exp(x_j^\top \beta)}{\sum_{k \in R(t_j)} \exp(x_k^\top \beta)} \right\}^{\delta_i} \tag{22}$$

(Wang et al., 1993) have shown that the maximum partial likelihood estimator is as efficient as the maximum truncation likelihood estimator.

The composite partial likelihood function (Huang and Qin 2012) can be found by exploiting the idea that A, truncation time, and V, residual lifetime after enrollment ($V = T - A$), have an exchangeable joint density function in the absence of censoring. One can apply the composite conditional likelihood method (Arnold and Strauss 1988) as the conditional density function of V given A is the same as A given V. Thus, the truncation density of T conditioning on A is the same as the conditional density of T on V. Following the same argument as $\mathcal{L}_P$, the composite partial likelihood for complete data can be obtained as follows

$$\prod_{i=1}^{n} \left[\frac{2 \exp(\beta' X_i)}{\sum_{j=1}^{n} \exp(\beta' X_j) \{ I(A_j \le T_i \le T_j) + I(V_j \le T_i \le T_j) \}} \right]^2 \tag{23}$$

Likewise, the composite partial likelihood for censored and truncated data is as follows

$$\prod_{i=1}^{n} \frac{\{\lambda(Y_i) \exp(\beta' X_i)\}^{2\delta_i} \exp\{-(1+\delta_i)\Lambda(Y_i) \exp(\beta' X_i)\}}{\exp\{-\{\Lambda(A_i) + \delta_i \Lambda(\tilde{V}_i)\} \exp(\beta' X_i)\}} \tag{24}$$

The maximum composite partial likelihood estimator can be found to be consistent and asymptotically

normal (Huang and Qin 2012).

5 Results

5.1 Simulation

Simulation studies were conducted to examine the performance of the methodology proposed by Huang and Qin (2012). We set sampling time to be 100 and simulated the onset of an event from a uniform distribution. The covariates X_1 and X_2 were independently generated from a normal distribution (continuous) and a Bernoulli distribution, respectively. The failure time was generated from three Cox models with different hazard functions: (Constant) $\lambda_0 = 2\exp(X_1 + X_2)$, (Increasing) $\lambda_0 = 2t\exp(X_1 + X_2)$, and (U-shaped) $\lambda_0 = 3t^2\exp(X_1 + X_2)$. The censoring rate was approximately generated to be 40%, 20%, and 0%. In each simulation, 1000 repetitions were done with first a sample size of N=400 and then a sample size of N=200. The composite partial likelihood method constructs a composite from exchangeable conditional likelihoods of (A, V) where A is the truncation time and V is the residual lifetime after enrollment. The composite partial likelihood method was compared to the partial likelihood method. The relative efficiency was calculated using the empirical variance of the maximum partial likelihood estimator divided by that of the maximum composite partial likelihood estimator.

Table 1: Partial Likelihood Method vs. Composite Partial Likelihood Method with N=400

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

Table 1 summarizes the empirical bias, empirical standard error, and estimated standard error. The simulation results show that both estimators are close to their estimands as seen from the small bias. The results also show that the variance increases as censoring rate increases as expected. Significantly, the bias for the composite partial likelihood method is always smaller than the bias for the partial likelihood method. Looking at the relative efficiency column, the composite partial likelihood method outperformed the partial likelihood method on all the entries showing it is more efficient.

Table 2: Partial Likelihood Method vs. Composite Partial Likelihood Method with N=200

Scenario	Cens	Coef	PL			CPL			
			Bias	SE	ESE	Bias	SE	ESE	RE
Constant	40%	$\hat{\beta}_1$	0.024	0.149	0.148	0.008	0.104	0.121	1.496
		$\hat{\beta}_2$	0.063	0.336	0.340	0.024	0.235	0.265	1.646
	20%	$\hat{\beta}_1$	0.014	0.131	0.133	0.005	0.092	0.109	1.489
		$\hat{\beta}_2$	0.047	0.315	0.330	0.019	0.221	0.264	1.563
	0%	$\hat{\beta}_1$	0.009	0.122	0.118	0.005	0.085	0.097	1.480
		$\hat{\beta}_2$	0.043	0.306	0.311	0.026	0.214	0.236	1.737
Increasing	40%	$\hat{\beta}_1$	0.015	0.133	0.135	0.010	0.093	0.118	1.309
		$\hat{\beta}_2$	0.026	0.264	0.274	0.013	0.186	0.235	1.359
	20%	$\hat{\beta}_1$	0.010	0.118	0.122	0.009	0.083	0.105	1.350
		$\hat{\beta}_2$	0.023	0.242	0.247	0.021	0.171	0.214	1.332
	0%	$\hat{\beta}_1$	0.011	0.108	0.106	0.008	0.076	0.093	1.299
		$\hat{\beta}_2$	0.030	0.227	0.231	0.025	0.160	0.207	1.245
U-shaped	40%	$\hat{\beta}_1$	0.018	0.129	0.136	0.013	0.091	0.121	1.263
		$\hat{\beta}_2$	0.028	0.255	0.257	0.024	0.180	0.224	1.316
	20%	$\hat{\beta}_1$	0.008	0.114	0.119	0.006	0.080	0.107	1.237
		$\hat{\beta}_2$	0.014	0.229	0.234	0.010	0.162	0.216	1.174
	0%	$\hat{\beta}_1$	0.009	0.104	0.105	0.008	0.073	0.095	1.222
		$\hat{\beta}_2$	0.027	0.211	0.212	0.024	0.149	0.195	1.182

Table 2 reports the same simulation but with a smaller sample size. The bias for all entries have increased due to the smaller sample size but the composite partial likelihood method consistently has smaller bias compared to the partial likelihood method. Again, looking at the RE column, the composite partial likelihood method was more efficient than the partial likelihood method in all situations. Thus, the composite partial likelihood method is more efficient than the partial likelihood method.

5.2 Data Analysis

In this section, Channing House was analyzed by applying the proposed method. Channing House (Cutler and Ederer 1958) was accessed via the asaur package in R (Kleinbaum and Klein 2012). Channing House is a retirement centre in Palo Alto, California. The dataset contains entries from 1964 to 1975. During that time, 97 men and 365 women joined the centre. Their age in months was recorded on entry as well as their

age on exit (either leaving or death). A large number of entries were censored (38%) as many residents were still alive past 1975. During the study 130 women and 46 men died while at Channing House. The data is also left truncated as many residents entered at different ages, but were not observed before entry (Davison and Hinkley 1997). Each entry contains five columns: sex, entry, exit, time, and cens.

To use the data, first a stationarity test was applied. Recall, length-biased failure time $T = A + V$ where A is truncation time, or backward recurrence time, and V is residual lifetime after enrollment, or forward recurrence.

Figure 10: Stationarity Test for Channing House Data (Age 65+)

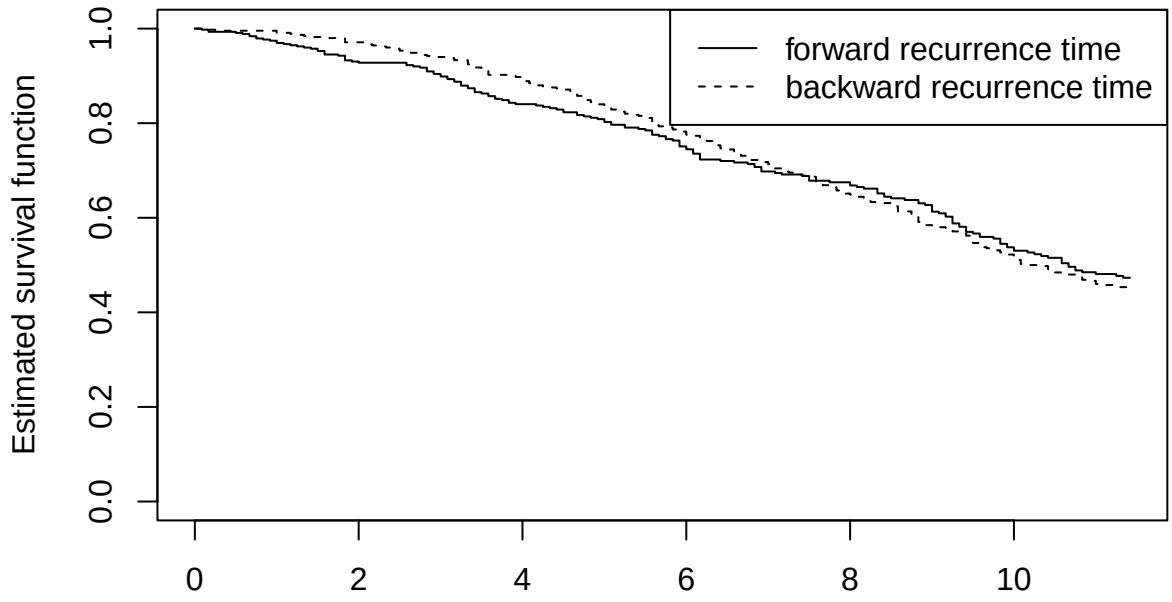


Figure 10 shows a stationarity check. Since the curves are visually close to each other, Channing House appears to be stationary. In addition, we can conduct an analytical test.

Table 3: Test Statistic and P-value for Stationarity

Test Statistic	P-value
0.261	0.794

Since Table 3 shows a p-value=0.794, we fail to reject the null hypothesis. Thus, the stationarity assumption for Channing House is valid.

Figure 11: Summary Statistics for Entry and Exit

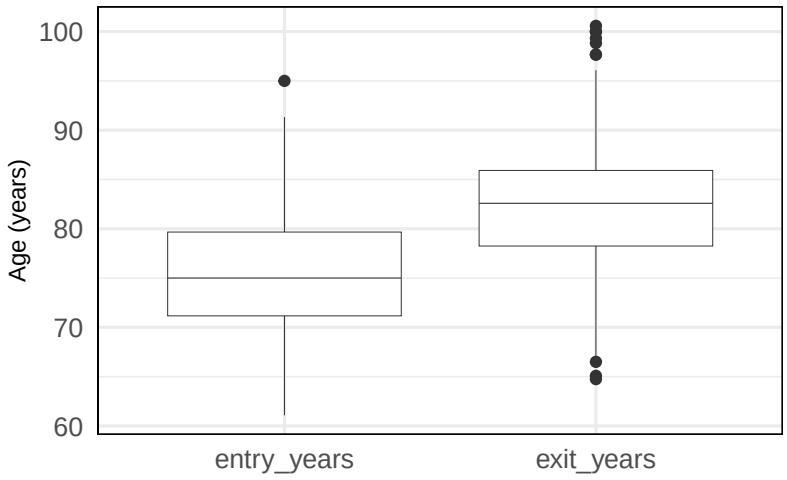


Figure 11 illustrates variation in age for those entering and exiting Channing House. The earliest age upon entry was 60 with the latest being an outlier at 95. The earliest age upon exit was an outlier at 65 and the latest age was an outlier at 100. The median age upon entry was 75 and the median age upon exit was 83.

To measure the time to death of Channing House residents, a Cox proportional hazards model was fit. Keeping consistent with the simulation, the first covariate chosen was continuous and was the age upon entry. The second covariate was binary and was the sex of the resident. The relative efficiency was calculated using the empirical variance of the maximum partial likelihood estimator divided by that of the composite partial likelihood estimator.

Table 4: The Partial Likelihood Method vs. The Composite Partial Likelihood Method with Channing House

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

Table 4 shows that the standard error for the composite partial likelihood method is smaller than the standard error for the partial likelihood method. Looking at the relative efficiency column, the composite partial likelihood method continues to be more efficient compared to the partial likelihood method for both covariates as expected. With $\alpha = 0.05$, the p-values indicate that β_2 for the partial likelihood method and β_1 and β_2 for the composite partial likelihood method are statistically significant. The hazard ratio indicates that for PL and age of entry, the risk of dying decreases by 4% per year. The hazard ratio for CPL and the age at entry indicates that with every year, the risk of dying decreases by 14% per year. Each method agrees

there is a reduction but differs on the amount. The hazard ratio for CPL and sex indicates that males have a 40% higher chance of dying than their female counterparts. Both methods agreed that the risk of death decreased with age and that men have a higher chance of mortality than women.

6 Conclusions

This project explored the methods done by Huang and Qin (2012) to show that the composite partial likelihood method is more efficient than the partial likelihood method under length-biased prevalent cohort sampling. The benefits of the composite partial likelihood method are avoiding cumbersome computations but benefiting from likelihood methods. Efficiency was improved by including additional information about the distribution of the truncated data. The method was then applied to a left-truncated data set, Channing House. The results in data analysis show that under the assumption that the left-truncated data comes from a uniform distribution the composite partial likelihood method is more efficient. The composite partial likelihood method can be generalized from uniform distributions to symmetric distributions. Future studies may be able to pursue it.

References

Arnold, Barry C, and David Strauss. 1988. "Bivariate Distributions with Exponential Conditionals." *Journal of the American Statistical Association* 83 (402): 522–27.

Asgharian, Masoud, David B Wolfson, and Xun Zhang. 2006. "Checking Stationarity of the Incidence Rate Using Prevalent Cohort Survival Data." *Statistics in Medicine* 25 (10): 1751–67.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187–202.

Cutler, Sidney J, and Fred Ederer. 1958. "Maximum Utilization of the Life Table Method in Analyzing Survival." *Journal of Chronic Diseases* 8 (6): 699–712.

Davison, Anthony C., and David V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press. https://CRAN.R-project.org/package=boot.

Huang, Chiung-Yu, and Jing Qin. 2012. "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): 946–57.

Klein, John P., and Melvin L. Moeschberger. 2003. *Survival Analysis Techniques for Censored and Truncated Data*. Second.

Kleinbaum, David G., and Mitchel Klein. 2012. *Asaur: Data Sets for "Applied Survival Analysis Using r"*.

https://CRAN.R-project.org/package=asaur.

Qin, Jing, and Yu Shen. 2010. "Statistical Methods for Analyzing Right-Censored Length-Biased Data Under Cox Model." *Biometrics* 66 (2): 382–92.

Wang, Mei-Cheng, Ron Brookmeyer, and Nicholas P Jewell. 1993. "Statistical Models for Prevalent Cohort Data." *Biometrics*, 1–11.