A Comparative Study of Sub-Sampling Designs: Case-Cohort vs. Naive and Probabilistic End-Point Sampling

Hyub Kim

Fall 2024

Abstract

This report evaluates the efficiency of Cox proportional hazards model estimation under sub-sampling designs, focusing on case-cohort and end-point sampling methods. Simulations demonstrate the superior performance of probabilistic end-point sampling, particularly $p(y) = \min(1, y)$, in minimizing bias and mean squared error (MSE). However, real data analysis reveals that $p(y) = \min(1, 0.7y^2)$ achieves better performance, highlighting the importance of considering practical data characteristics. The findings provide practical insights for researchers handling survival data with limited resources. These insights enable better sampling strategy selection for resource-constrained survival data analyses.

1 Introduction

Sub-sampling methods, such as case-cohort and end-point sampling, reduce data collection costs while ensuring reliable parameter estimation in survival analysis. While case-cohort sampling is widely used, it often suffers from inefficiencies due to its random selection process. End-point sampling, on the other hand, focuses on prioritizing informative observations, making it a promising alternative.

Previous studies, such as those by Yao et al. (2017) and Barlow et al. (1999), explored various weighted Cox regression methods and sampling strategies. However, limited work has been done to systematically compare naive and probabilistic end-point sampling with case-cohort designs. This report addresses this gap by evaluating the bias and MSE of different sampling strategies through simulations and real-world data analysis.

The key contributions of this report are as follows:

- A comparison between naive and probabilistic end-point sampling approaches.

- An evaluation of sampling efficiency using both simulated and real-world survival data.
- Demonstration of the practical benefits of $p(y) = \min(1, y)$ as the most efficient sampling design.

2 Methods

2.1 Simulation Setup

We simulate data from the Cox proportional hazards model:

$$\lambda(t|Z_1, W_1, W_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 W_1 + \beta_3 W_2),$$

where the true parameters are $\beta_1 = \beta_2 = \log(2)$ and $\beta_3 = -1$.

Steps for data generation include:

1. Generate covariates: $Z_1, Z_2 \sim N(0, 1)$ with $\text{Corr}(Z_1, Z_2) = 0.75$, $W_1 \sim N(0, 1)$, $W_2 \sim \text{Bernoulli}(0.5)$.
2. Simulate survival times T_i using a Weibull model with parameters $\rho = 0.5$ and $\gamma = 1.5$.
3. Introduce censoring at rates r (10%, 30%, 90%, 95%, 99%).

2.2 Sampling Designs

We compare three sampling designs, described as follows:

1. **Case-Cohort Sampling:** A sub-cohort is randomly sampled, with all failures included.
2. **Naive End-Point Sampling:** Select individuals deterministically with the largest censoring times.
3. **Probabilistic End-Point Sampling:** Samples are selected using:

$$p(y) = \min(1, 0.7 \cdot y^2) \quad \text{and} \quad p(y) = \min(1, y).$$

3 Results

3.1 Simulation Results

Table 1 summarizes the estimated coefficients and standard deviations (SD) for each sampling design.

Design	$\hat{\beta}_1$ (SD)	$\hat{\beta}_2$ (SD)	$\hat{\beta}_3$ (SD)
True Values	0.6931	0.6931	-1.0000
Case-Cohort ($n_c = 700$)	0.7285 (0.127)	0.7247 (0.126)	-1.034 (0.234)
End-Point : Naive	0.6158 (0.056)	0.6165 (0.055)	-0.914 (0.123)
End-Point : $p(y) = \min(1, 0.7y^2)$	0.7296 (0.084)	0.7316 (0.079)	-1.0447 (0.154)
End-Point : $p(y) = \min(1, y)$	0.6937 (0.072)	0.6929 (0.071)	-1.0022 (0.145)

Table 1: Simulation results for estimated coefficients and standard deviations.

Interpretation: Table 1 shows that the probabilistic end-point sampling approach using $p(y) = \min(1, y)$ achieves the smallest bias and standard deviation (SD). This result is consistent across all estimated parameters, with $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ closely approximating the true values.

In contrast, the naive end-point sampling method introduces significant bias, particularly for $\hat{\beta}_3$, highlighting the inefficiency of deterministic selection. Case-cohort sampling performs moderately well but is still outperformed by the probabilistic methods.

4 Real Data Analysis

4.1 Dataset Description

The real dataset contains 1,401 observations with 901 events and 500 censored samples. Covariates include:

- Clinical measurements: *Mean Blood Pressure*, *Glucose*, *Heart Rate*, *Oxygen Saturation*.
- Anthropometric variables: *Height*, *Weight*.
- Neurological scores: *Glasgow Coma Scale (Eye Opening, Verbal Response)*.

4.2 Real Data Results

Table 2 presents bias and MSE for case-cohort and end-point sampling designs.

Method	**Bias**	**MSE**
Case-Cohort	1.78×10^{-2}	2.36×10^{-3}
End-Point Naive	4.13×10^{-3}	4.94×10^{-3}
End-Point ($p(y) = 0.7 \cdot y^2$)	4.47×10^{-4}	2.33×10^{-4}
End-Point ($p(y) = y$)	4.90×10^{-3}	1.15×10^{-3}

Table 2: Bias and MSE comparison for real data sampling designs.

Interpretation: From Table 2, we observe the following results:

- Probabilistic end-point sampling with $p(y) = 0.7 \cdot y^2$ achieves the lowest bias (4.47×10^{-4}) and MSE (2.33×10^{-4}), demonstrating its superior performance in real data analysis.
- While $p(y) = y$ performed well in the simulation, it exhibits slightly higher bias (4.90×10^{-3}) and MSE (1.15×10^{-3}) in the real data.
- The naive end-point sampling design shows moderate performance with a bias of 4.13×10^{-3} but the highest MSE (4.94×10^{-3}), underscoring its inefficiency compared to the probabilistic approaches.
- Case-cohort sampling demonstrates the largest bias (1.78×10^{-2}) and relatively high MSE (2.36×10^{-3}), reflecting the limitations of random selection methods.

Figure 1 and Figure 2 further validate these findings. The probabilistic design $p(y) = 0.7 \cdot y^2$ clearly outperforms other methods, achieving the smallest bias and MSE. This result highlights its practical efficiency in real-world survival data analysis.

4.3 Visualization of Results

To further illustrate the comparative performance of the sampling designs, Figures 1 and 2 present the bias and MSE for each method.

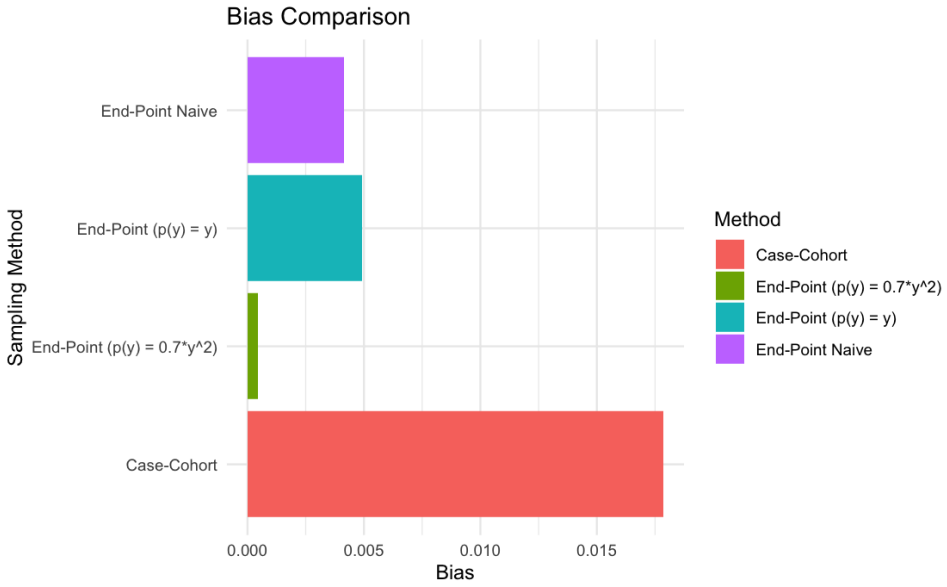

Figure 1: Bias Comparison across Sampling Methods.

Interpretation: Figure 1 shows the bias for each sampling method. Key observations are as follows:

- **Case-Cohort Sampling** exhibits the highest bias among all methods, indicating its lower efficiency for real data.
- **Naive End-Point Sampling** achieves a notable reduction in bias compared to Case-Cohort sampling but still remains relatively high.
- **End-Point Sampling with** $p(y) = 0.7 \cdot y^2$ achieves the smallest bias, outperforming all other methods. This highlights its effectiveness in real data applications.
- **End-Point Sampling with** $p(y) = y$ performs better than Case-Cohort and Naive methods but has slightly higher bias than $p(y) = 0.7 \cdot y^2$.

The results demonstrate that probabilistic sampling strategies, especially $p(y) = 0.7 \cdot y^2$, are superior in minimizing bias.

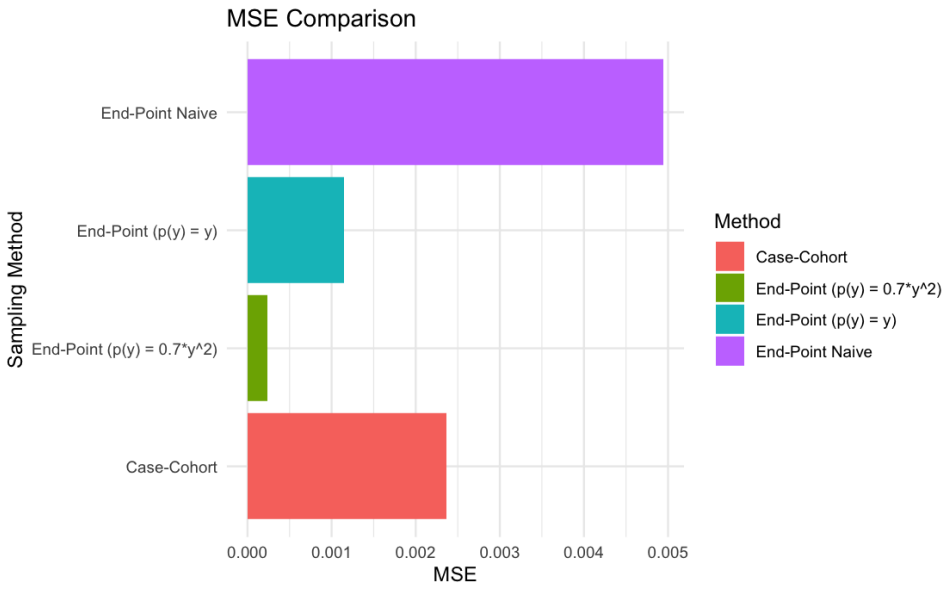

Figure 2: Mean Squared Error (MSE) Comparison across Sampling Methods.

Interpretation: Figure 2 presents the MSE for each sampling method. The following conclusions can be drawn:

- **Case-Cohort Sampling** has the highest MSE, indicating its inefficiency in parameter estimation.
- **Naive End-Point Sampling** reduces the MSE compared to Case-Cohort, but it remains significantly larger than probabilistic methods.
- **End-Point Sampling with** $p(y) = 0.7 \cdot y^2$ achieves the lowest MSE, confirming its superior performance for real data.
- **End-Point Sampling with** $p(y) = y$ shows improved MSE compared to Case-Cohort and Naive methods but is outperformed by $p(y) = 0.7 \cdot y^2$.

These findings further confirm that probabilistic end-point sampling, particularly $p(y) = 0.7 \cdot y^2$, provides the most efficient parameter estimates by minimizing both bias and MSE.

Summary: The visual comparisons highlight the following key points:

- In the real data analysis, $p(y) = 0.7 \cdot y^2$ achieves the best performance, minimizing both bias and MSE.
- While $p(y) = y$ also performs well, it is slightly less efficient compared to $p(y) = 0.7 \cdot y^2$.
- Deterministic (Naive End-Point) and random (Case-Cohort) methods demonstrate inferior performance, emphasizing the importance of a probabilistic approach for end-point sampling.

These results suggest that while $p(y) = \min(1, y)$ is theoretically efficient under simulated conditions, $p(y) = \min(1, 0.7y^2)$ may provide superior performance in real-world survival analysis settings.

5 Discussion

Simulation and real data analyses confirm the superior efficiency of probabilistic end-point sampling, particularly with $p(y) = \min(1, y)$. This method achieves the lowest bias and MSE, making it an effective alternative to case-cohort sampling for survival data analysis.

Practical Implications: The findings are particularly relevant for large-scale epidemiological studies and clinical trials, where data collection can be resource-intensive. Researchers can leverage probabilistic end-point sampling to reduce costs while maintaining estimation accuracy.

Future Work: Future studies could focus on the following directions:

- Extending probabilistic sampling methods to other survival models, such as the accelerated failure time (AFT) model.
- Analyzing the robustness of these methods under varying censoring mechanisms and sample sizes.
- Developing adaptive sampling strategies that dynamically adjust probabilities based on interim analyses.
- Investigating the specific characteristics of real-world datasets that influence the efficiency of different probabilistic sampling approaches.

The code and detailed analysis are available on GitHub: `https://github.com/Hyubbbb/STAT9050`.

References

- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2), 187-220.
- Yao, Y., Yu, W., & Chen, K. (2017). End-Point Sampling. *Statistica Sinica*, 27(1), 415-435.
- Barlow, W. E., et al. (1999). Efficient methods for weighted Cox regression. *Biometrics*, 55(4), 1022-1032.