Interval Censoring

Background and a review of Zhan and Sun (2010)

Brian Segal

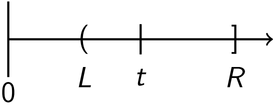
August 29, 2017

Outline

What is interval censoring?

Type II (General)

Event is known to occur between two time points.



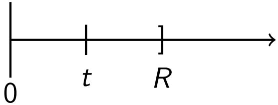
Notation

- t = event time (unobserved)
- L = left side of interval
- R = right side of interval

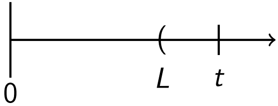
Type I (Current status)

Event is known to occur before or after a single time point:

- Left censoring ($L = 0$)



- Right censoring ($R = \infty$)

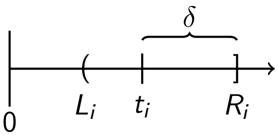


What's the problem with ignoring interval censoring?

Survival time is over-estimated

- Suppose time of event $t_i \in (L_i, R_i]$ is interval censored
- Assuming $t_i = R_i$ causes survival time to be over-estimated $(R_i \geq t_i)$

Example for patient i



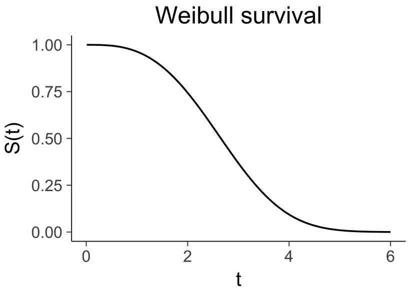
How much does this bias survival estimates?

It depends

Let $\delta = R_i - t_i$ be the common measurement error and suppose event times follow survival function S. Size of bias depends on:

- Size of measurement error δ
- Change in S between times t_i and R_i

Example: S is Weibull with shape and scale of 3



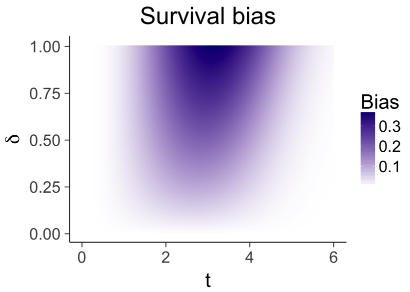
How much does this bias survival estimates?

It depends

Let $\delta = R_i - t_i$ be the common measurement error and suppose event times follow survival function S. Size of bias depends on:

- Size of measurement error δ
- Change in S between times t_i and R_i

Example: S is Weibull with shape and scale of 3 (see Appendix)



Bias is a problem when

- δ is large
- Slope of S is large

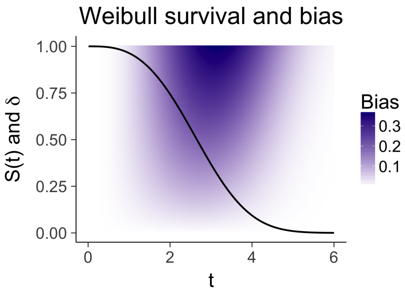
How much does this bias survival estimates?

It depends

Let $\delta = R_i - t_i$ be the common measurement error and suppose event times follow survival function S. Size of bias depends on:

- Size of measurement error δ
- Change in S between times t_i and R_i

Example: S is Weibull with shape and scale of 3 (see Appendix)



Bias is a problem when

- δ is large
- Slope of S is large

How do we avoid this bias?

Methods for interval-censored data

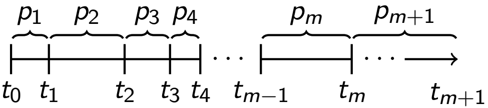
Use a likelihood proportional to

$$L = \prod_{i=1}^{n} \underbrace{[S(L_i) - S(R_i)]}_{\text{Pr(event between } L_i \text{ and } R_i\text{)}}$$

Nonparametric maximum likelihood estimator (NPMLE)

Turnbull estimator of $\hat{S}$ (1976): interval censoring counterpart to Kaplan-Meier

- Partitions timeline by all left and right censoring times, and estimates probability of each partition



Nonparametric maximum likelihood estimator (NPMLE)

Turnbull estimator of $\hat{S}$ (1976): interval censoring counterpart to Kaplan-Meier

- Pros
  - Consistent (with enough data, the estimate is correct)
  - Can incorporate right-censored data by setting $R_i = \infty$
- Cons
  - Statistical convergence is slower than Kaplan-Meier (need more data for a good estimate)
  - No closed form – requires iterative fitting algorithm

Nonparametric maximum likelihood estimator (NPMLE)

Side notes

- The Turnbull estimator (1976) is an EM algorithm, though the seminal EM paper was not published until 1977 (Dempster and Waird).

$$p_j^{\text{new}} = \overbrace{\frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\frac{\alpha_{ij} p_j^{\text{old}}}{\sum_{l=1}^{m+1} \alpha_{ij} p_l^{\text{old}}}\right)}_{\text{E step: } q_{ij}=\mathbb{E}[\Pr(t_{j-1}<T_i\leq t_j)]}}^{\text{M step: } p_j^{\text{new}}=\arg\max_{p_j} L(\boldsymbol{p}|\boldsymbol{q})}$$

- Faster algorithms exist

Hypothesis testing with NPMLEs

Comparing survival functions

- Very similar to right-censored data
- Log rank tests with modified calculations of
  - d_j: number of events at time t_j
  - n_j: number at risk at time t_j.
- Note: formulas for d_j and n_j very similar to updates for the Turnbull estimator

Regression

Common models

Similar to right-censored data, we can fit

- Semiparametric
  - Proportional hazards (Cox)
  - Proportional odds
  - Additive hazards
- Parametric
  - Accelerated failure time and generalizations
  - Piecewise exponential

Issues for Cox model

Computational

Baseline hazards do not cancel out of likelihood and must be estimated

Statistical

While baseline hazard converges at $n^{1/3}$ rate, regression coefficients still converge at $n^{1/2}$ rate (Huang and Wellner, 1997)

Key assumption: Non-informative interval censoring

Non-informative interval censoring

Except for the requirement that $L_i < t_i \leq R_i$, L_i and R_i contain no additional information about survival time.

Common violation

Sick patients are seen more often than healthy patients, so if $R_i - L_i$ is small, t_i is probably closer to L_i than R_i (expected survival time is shorter).

Implications

Estimates of baseline hazard might be wrong. How much does this affect estimates of regression coefficients?

Software

R packages

- CRAN survival view
- Anderson-Bergman (Preprint). icenReg: Regression Models for Interval Censored Data in R. Available here (also see the icenReg vignette).
- Gómez, G., Luz Calle, M., Oller, R., Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modeling*. 9: 259–297. Available here. (Does anyone have access?)

References

- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B*. 39: 1–38.
- Huang, J., Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In: Lin, D., Fleming, T., editors. Proceedings of the first Seattle symposium in biostatistics: survival analysis. New York: Springer-Verlag.
- Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society: Series B*. 38: 290–295.
- Zhang, Z., Sun, J. (2010). Interval Censoring. *Statistical Methods in Medical Research*. 19: 53–70.

Appendix

Overview

This appendix outlines the details mentioned in earlier slides. I deal with the simple case where the time shift δ is the same for all patients. While not likely to be the case in practice, I think it still provides some insights.

Notation and assumptions

Let the event times $T \sim F$, where $F(t) = \Pr(T \leq t)$. Let $S(t) = 1 - F(t)$ be the survival function and suppose that $\hat{S}_n$ is a consistent estimator of S for right censored data, such as the Kaplan-Meier estimator. That is, $\hat{S}_n(t) \to S(t)$ as $n \to \infty$. The subscript indexes $\hat{S}_n$ by the number of observations $i = 1, \ldots, n$.

Appendix

Bias from assuming $\delta = 0$

Ignoring interval censoring is equivalent to assuming $\delta = 0$. In this case, a patient's survival time is assumed to be R_i even though it is actually $t_i = R_i - \delta$. Consequently $\hat{S}_n(R_i)$ is an estimate of S not at time R_i, but at time $t_i = R_i - \delta$. That is,

$$\hat{S}_n(R_i) = \hat{S}_n(t_i + \delta) \to S(t_i).$$

Because S is monotone non-increasing and $\delta \geq 0$, we have $S(t_i) \geq S(t_i + \delta)$, which causes our estimate to be biased upward.

Appendix

Approximating the bias

This gives an asymptotic bias of

$$\begin{aligned}\text{bias}_n(t_i+\delta) &= \mathbb{E}[\hat{S}_n(t_i+\delta)] - S(t_i+\delta)\\ &\to S(t_i) - S(t_i+\delta). \end{aligned} \quad (1)$$

This shows that bias is a function of both the size of δ and the derivative of S (if S is nearly constant over $(t_i, t_i+\delta)$ then bias is near zero). To make this explicit (and assuming the density $f(t) = -\frac{d}{dt}S(t)dt$ exists at t_i) we can take a first order Taylor expansion of $S(t_i+\delta)$ about t_i to get that for sufficiently large n,

$$\begin{aligned}\text{bias}_n(t_i+\delta) &\approx S(t_i) - (S(t_i) - \delta f(t_i))\\ &= \delta f(t_i). \end{aligned} \quad (2)$$

I show (1) in earlier slides, though (2) is very similar.