The Journal of Open Source Software

ccostr: An R package for estimating mean costs with censored data

Lars Børty[1], **Rasmus Brøndum**[1, 2], **and Martin Bøgsted**[1, 2, 3]

1 Department of Hematology, Aalborg University Hospital, Aalborg, Denmark **2** Clinical Cancer Research Center, Aalborg University Hospital, Denmark **3** Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

DOI: 10.21105/joss.01593

Software

- Review
- Repository
- Archive

Submitted: 09 July 2019
Published: 05 September 2019

Summary

Censoring is a frequent obstacle when working with time to event data, as e.g. not all patients in a medical study can be observed until death. For estimating the distribution of time to event the Kaplan-Meier estimator is useful, but when estimating mean costs it is not, since costs, as opposed to time, typically don't accumulate at a constant rate. Often costs accumulate at a higher rate at the beginning (e.g. at diagnosis) and end (e.g. death) of the study.

Several methods for estimating mean costs when working with censored data have been developed. Of note is the work by Lin, Feuer, Etzioni, & Wax (1997), who proposed three different estimators. The first, $LinT$, partitions the time period into small intervals and then estimates the mean costs by weighting the mean total cost of fully observed individuals in the interval with the probability of dying in the interval. The two others, $LinA$ and $LinB$, weight the mean total cost within each interval with the probability of being alive at respectively the start or end of the interval.

Later Bang & Tsiatis (2000) proposed another method based on inverse probability weighting, where complete (fully observed) cases are weighted with the probability of being censored at their event time. Two estimators were presented: the simple weighted estimator, BT, using total costs for fully observed cases, and the partitioned estimator, BTp, utilizing cost history. Hongwei Zhao & Tian (2001) proposed an extension of the BT estimator, ZT, which includes cost history from both censored and fully observed cases. The ZT estimator was later simplified by Pfeifer & Bang (2005).

In Hongwei Zhao, Bang, Wang, & Pfeifer (2007) they demonstrated the similarity of the different estimators when using the distinct censoring times for defining intervals. They concluded that the following equalities hold for the estimates of mean cost: $\hat{\mu}_{BT} = \hat{\mu}_{LinT}$ and $\hat{\mu}_{LinA} = \hat{\mu}_{LinB} = \hat{\mu}_{BTp} = \hat{\mu}_{ZT}$. The estimators can be split into two classes: those that use and those that do not use cost history. As cost history contributes additional information these estimators are in general more efficient, and should be chosen if cost history is available.

Previous implementations of these estimators into statistical software have been done in Stata, first by Kim & Thompson (2011) who implemented the method from Lin et al. (1997), and later by Chen, Rolfes, & Zhao (2015) who implemented the BT and ZT estimators, and in SAS by Honwei Zhao & Wang (2010). To our knowledge none of the methods have previously been implemented in an R package.

Estimators

The R package ccostr includes four different estimators of the mean cost. The average sample, AS, estimator simply averages the total cost per individual, disregarding censoring, giving a

downwards biased estimate since costs after censoring are not accounted for. The complete case, CC, estimator averages the cost of only fully observed cases, biasing the estimate towards the average cost for individuals with shorter survival, typically downwards biased. These two naive estimators are included as reference for the estimators accounting for censoring. For dealing with censored data we implement the BT and ZT estimators for handling situations with or without cost histories.

Assume we observe $\{(T_i, \Delta_i, [M_i(u), 0 \leq u \leq T_i]), i = 1, ..., n\}$, where n is the number of individuals, T_i is the observation time, $M_i(u)$ the cost until time u, and Δ_i is event indicator for individual i, with $\Delta_i = 1$ or $\Delta_i = 0$ for fully observed and censored cases, respectively. Then the estimates are given by:

Naive "Available Sample estimator" and "Complete Case estimator":

$$\hat{\mu}_{AS} = \frac{\sum_{i=1}^{n} M_i}{n} \qquad \hat{\mu}_{CC} = \frac{\sum_{i=1}^{n} \Delta_i M_i}{\sum_{i=1}^{n} \Delta_i}.$$

where $M_i = M_i(T_i)$ denotes the total cost.

Bang and Tsiatis's estimator (also known as Weighted Complete Case estimator):

$$\hat{\mu}_{BT} = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i M_i}{\hat{K}(T_i)}$$

Where $\hat{K}(T_i)$ is the Kaplan-Meier estimator of the probability of censoring at time T_i, i.e. the time of event for individual i.

Zhao and Tian's estimator (also known as Weighted Available Sample estimator):

$$\hat{\mu}_{ZT} = \frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \frac{M_i}{\hat{K}(T_i)} + (1 - \Delta_i) \frac{\{M_i - \overline{M(C_i)}\}}{\hat{K}(C_i)} \right],$$

where $\overline{M(C_i)}$ is the average of cost until time C_i among individuals with event time later than C_i, and $\hat{K}(C_i)$ is the Kaplan-Meier estimator of the censoring probability at the time T_i.

Application

We have implemented the functions above in an R package, ccostr. The package includes two main functions, the first is ccmean() which calculates the mean cost until a time limit, specified with the parameter "L=", and takes as input a data frame in the following format:

```
library(ccostr)
head(hcost)
```

```
##      id start  stop  cost   trt delta  surv
## 1     1     1     1  3694     0     0   575
## 2     1     1     9     1     0     0   575
## 3     1     1     9    12     0     0   575
## 4     1     1    34   106     0     0   575
## 5     1     1   237    68     0     0   575
## 6     1     1   237    86     0     0   575
```

The data shown above are simulated data from the Stata hcost package (Chen et al., 2015). Applying ccmean() on the data with a time limit of L = 1461, gives results identical to hcost in Chen et al. (2015). The option addInterpol adds a small value to the numerator and denominator of the fraction used for interpolation of cost at unobserved times, and is only used here to mimic the implementation in hcost, by default it is set to zero.

```
ccmean(hcost, L = 1461, addInterPol = 1)
```

```
## ccostr - Estimates of mean cost with censored data
##
##   Observations Induviduals Events Limits TotalTime MaxSurv
## N         9704         160     61   1461    122401    2082
##
##    Estimate Variance      SD     95UCI    95LCI
## AS 63725.42 19193502 4381.04  72312.26 55138.59
## CC 74779.13 37572385 6129.63  86793.21 62765.05
## BT 86175.16 51593885 7182.89 100253.62 72096.70
## ZT 80134.84 23726332 4870.97  89681.94 70587.74
##
## Mean survival time: 1165.04 With SE: 41.94
```

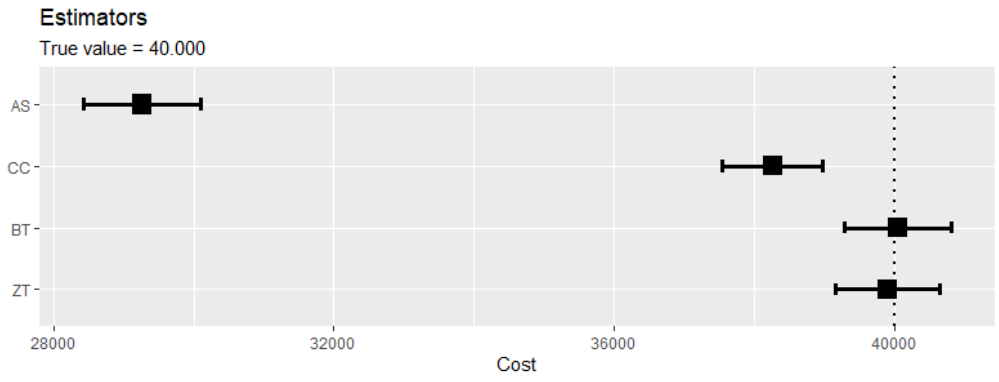
The second main function in ccostr is simCostData(). This function simulates data in the correct format according to the method in Lin et al. (1997), and may be used for testing purposes:

```
sim <- simCostData(n = 1000, dist = "unif", censor = "heavy", L = 10)
head(sim$censoredCostHistory)
```

```
##   id start     stop       cost     surv delta
## 1  1     0 1.000000 10189.0404 1.654524     0
## 2  1     1 1.654524   813.4767 1.654524     0
## 3  2     0 1.000000 11137.0038 6.158623     0
## 4  2     1 2.000000  1722.1321 6.158623     0
## 5  2     2 3.000000  1672.4579 6.158623     0
## 6  2     3 4.000000  1858.5673 6.158623     0
```

The true mean cost of the simulated dataset is 40,000 (Lin et al., 1997). Applying the ccmean function to the simulated data yields the result below. We here present the result graphically using the built-in plotting function for an object of the ccobject class.

```
library(ggplot2)
simMean <- ccmean(sim$censoredCostHistory)
plot(simMean) +
  geom_hline(yintercept = 40000, linetype = "dotted", size = 1) +
  labs(subtitle = "True value = 40.000")
```



References

Bang, H., & Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika*, *87*(2), 329–343. doi:10.1093/biomet/87.2.329

Chen, S., Rolfes, J., & Zhao, H. (2015). Estimation of mean health care costs and incremental cost-effectiveness ratios with possibly censored data. *Stata Journal*, *15*(3), 698–711. doi:10.1177/1536867X1501500305

Kim, L. G., & Thompson, S. G. (2011). Estimation of life-years gained and cost effectiveness based on cause-specific mortality. *Health Economics*, *20*(7), 842–852. doi:10.1002/hec.1648

Lin, D. Y., Feuer, E. J., Etzioni, R., & Wax, Y. (1997). Estimating Medical Costs from Incomplete Follow-Up Data. *Biometrics*, *53*(2), 419. doi:10.2307/2533947

Pfeifer, P. E., & Bang, H. (2005). Non-parametric estimation of mean customer lifetime value. *Journal of Interactive Marketing*, *19*(4), 48–66. doi:10.1002/dir.20049

Zhao, H., & Tian, L. (2001). On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics*, *57*(4), 1002–1008. doi:10.1111/j.0006-341X.2001.01002.x

Zhao, H., & Wang, H. (2010). Cost and Cost-Effectiveness Analysis with Censored Data. In D. E. Faries, R. Obenchain, J. M. Haro, & A. C. Leon (Eds.), *Analysis of observational health care data using SAS* (pp. 363–381). SAS Press.

Zhao, H., Bang, H., Wang, H., & Pfeifer, P. E. (2007). On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine*, *26*(24), 4520–4530. doi:10.1002/sim.2882