Published in final edited form as:

Biom J. 2022 June ; 64(5): 858–862. doi:10.1002/bimj.202100250.

Correcting conditional mean imputation for censored covariates and improving usability

Sarah C. Lotspeich, **Kyle F. Grosser**, **Tanya P. Garcia**

Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

Missing data are often overcome using imputation, which leverages the entire dataset to replace missing values with informed placeholders. This method can be modified for censored data by also incorporating partial information from censored values. One such modification proposed by Atem et al. (2017, 2019a, 2019b) is conditional mean imputation where censored covariates are replaced by their conditional means given other fully observed information. These methods are robust to additional parametric assumptions on the censored covariate and utilize all available data, which is appealing. However, in implementing these methods, we discovered that these three articles provide nonequivalent formulas and, in fact, none is the correct formula for the conditional mean. Herein, we derive the correct form of the conditional mean and discuss the bias incurred when using the incorrect formulas. Furthermore, we note that even the correct formula can perform poorly for log hazard ratios far from **0**. We also provide user-friendly R software, the imputeCensoRd package, to enable future researchers to tackle censored covariates correctly.

Keywords

random censoring; reproducibility; r package; survival analysis; trapezoidal rule

1 | CHALLENGES WITH CURRENT APPROACHES

A popular way to overcome missing data is imputation, where missing values are replaced by informed placeholders that leverage the entire dataset. Imputation methods can similarly be applied to censored data, for example, by replacing a censored value with its conditional mean given other fully observed information. This method is known as conditional mean

Correspondence Tanya P. Garcia, Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA. tpgarcia@email.unc.edu.
Sarah C. Lotspeich and Kyle F. Grosser contributed equally to this work.

CONFLICT OF INTEREST
The authors have declared no conflict of interest.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.
This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

SUPPORTING INFORMATION
Additional supporting information may be found in the online version of the article at the publisher's website.

imputation. It is considered a single imputation approach but is a perfectly valid alternative to multiple imputation (Little, 1992). Recently, Atem et al. (2017, 2019a, 2019b) developed a conditional mean imputation estimator for censored covariates that is robust to additional parametric assumptions and utilizes all available data. However, in implementing the estimator of Atem and colleagues, we found that these articles actually state the conditional mean incorrectly and provide nonequivalent formulas. To remedy this, we derive the correct form of the conditional mean and provide implementation with a user-friendly R package called imputeCensoRd.

2 | WHAT IS THE CORRECT FORM OF THE CONDITIONAL MEAN?

To replicate the setup of Atem and colleagues, we consider a linear regression model, $\mathrm{E}(Y|X, \mathbf{Z}) = \alpha + \beta X + \boldsymbol{\gamma}^T\mathbf{Z}$, relating the outcome Y to censored and fully observed covariates X and $\mathbf{Z}$, respectively. Data are available on a sample of n subjects, but in place of X is the observed covariate value $W = \min(X, C)$, where C is a right-censored value and we define the event indicator $\delta = \mathrm{I}(X \le C)$. We assume X is independent of C given $\mathbf{Z}$ so that X is missing at random from censored subjects (Little & Rubin, 2002). We focus on right censoring to mimic Atem and colleagues, but modifications for left censoring follow closely. Conditional mean imputation involves computing $\mathrm{E}(X|X > C, \mathbf{Z})$ for each censored subject, which we prove below is

$$\mathrm{E}(X \mid X > C_i, \mathbf{Z}_i) = C_i + \frac{1}{S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)}} \int_{C_i}^{\infty} S_0(x)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} \mathrm{d}x, \tag{1}$$

where $S_X(t|\mathbf{Z}) = \mathrm{P}(X \ge t|\mathbf{Z})$ is the conditional survival function of $X|\mathbf{Z}$, $S_0(t) = S_X(t|\mathbf{Z} = \mathbf{0})$ is the baseline survival function for X, and $\boldsymbol{\lambda}$ are the log hazard ratios from the Cox model for $X|\mathbf{Z}$. An alternative to the Cox model here would be a fully parametric approach to modeling $S_X(x|\mathbf{Z}_i)$, similar to Royston (2007).

One way to approximate Equation (1) is to use the trapezoidal rule and incorporate the indicator $\mathrm{I}(W_{(j)} \ge C_i)$ for right censoring. This leads to $\widehat{\mathrm{E}}(X_i \mid X_i > C_i, \mathbf{Z}_i) =$

$$C_i + \frac{1}{2} \left[\frac{\sum_{j=1}^{n-1} \mathrm{I}\left(W_{(j)} \ge C_i\right)\left\{S_0\left(W_{(j+1)}\right)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} + S_0\left(W_{(j)}\right)^{\exp\left(\lambda^T \mathbf{Z}_i\right)}\right\}\left(W_{(j+1)} - W_{(j)}\right)}{S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)}} \right] \tag{2}$$

where $W_{(1)} < W_{(2)} < \cdots < W_{(n)}$ are the ordered, observed values of $W = \min(X, C)$. To best capture this integral, $S_0(W_{(n)})$ should be approximately 0 since $\lim_{t\to\infty} S_0(t) = 0$. Also, given that the subintervals in the trapezoidal rule are defined on the observed values, larger sample sizes should lead to better approximations.

To derive Equation (1), we begin simply with the definition of expectation, which gives us that $\mathrm{E}(X \mid X > C_i, \mathbf{Z}_i) = \int_{-\infty}^{\infty} x\mathrm{P}(x \mid x > C_i, \mathbf{Z}_i)\mathrm{d}x$, where only X is considered random given observed data $(C_i, \mathbf{Z}_i)$ for a given subject. It follows from Bayes' theorem that $\mathrm{P}(X|X\ C_i, \mathbf{Z}_i) = \mathrm{P}(X, X > C_i|\mathbf{Z}_i)/\mathrm{P}(X > C_i|\mathbf{Z}_i)$, where we recognize $\mathrm{P}(X > C_i|\mathbf{Z}_i)$ as the conditional survival function of $X|\mathbf{Z} = \mathbf{Z}_i$ at time C_i and thus $\mathrm{P}(X|X\ C_i, \mathbf{Z}_i) = \mathrm{P}(X, X > C_i|\mathbf{Z}_i)/S_X(C_i|\mathbf{Z}_i)$. This gives us

$$\mathrm{E}(X \mid X > C_i, \mathbf{Z}_i) = \frac{1}{S_X(C_i \mid \mathbf{Z}_i)} \int_{-\infty}^{\infty} x\mathrm{P}(x, x > C_i \mid \mathbf{Z}_i)\mathrm{d}x\,. \tag{3}$$

Under the assumption that X is independent of C given $\mathbf{Z}$,

$$\mathrm{E}(X \mid X > C_i, \mathbf{Z}_i) = \frac{1}{S_X(C_i \mid \mathbf{Z}_i)} \int_{C_i}^{\infty} x\mathrm{P}(x \mid \mathbf{Z}_i)\mathrm{d}x\,. \tag{4}$$

Now, using integration by parts, $\int_{C_i}^{\infty} x\mathrm{P}(x \mid \mathbf{Z}_i)\mathrm{d}x$ equals

$$\begin{aligned} &\left\{\lim_{x\to\infty} - xS_X(x \mid \mathbf{Z}_i)\right\} - \{-C_iS_X(C_i \mid \mathbf{Z}_i)\} + \int_{C_i}^{\infty} S_X(x \mid \mathbf{Z}_i)\mathrm{d}x \\ &= C_iS_X(C_i \mid \mathbf{Z}_i) + \int_{C_i}^{\infty} S_X(x \mid \mathbf{Z}_i)\mathrm{d}x\,. \end{aligned} \tag{5}$$

Finally, plugging Equation (5) into Equation (4) yields $\mathrm{E}(X \mid X > C_i, \mathbf{Z}_i) = C_i + \frac{1}{S_X(C_i \mid \mathbf{Z}_i)}\int_{C_i}^{\infty} S_X(x \mid \mathbf{Z}_i)\mathrm{d}x$, and since X is modeled with a Cox model, $S_X(t \mid \mathbf{Z}_i) = S_0(t)^{\exp\left(\lambda^T\mathbf{Z}_i\right)}$. With this, we arrive at Equation (1).

Remark.

Even the correct formula, Equation (2), can be volatile for log hazard ratios farther from the null, $\boldsymbol{\lambda} = \mathbf{0}$, sometimes leading to biased inference. (See Web Appendix B for more information.)

3 | IMPACTS OF IMPUTING WITH INCORRECT FORMULAS

3.1 | Misplacing the hazard ratio

The formula provided in Atem et al. (2017) is $\widehat{\mathrm{E}}(X \mid X > C_i, \mathbf{Z}_i)$

$$= C_i + \frac{1}{2}\left[\frac{\sum_{j=1}^{n-1}\boxed{\mathrm{I}(W_{(j)} > C_i)}\left\{S_0\left(W_{(j+1)}\right) + S_0\left(W_{(j)}\right)\right\}\boxed{\exp(\lambda^{\mathrm{T}}\mathbf{Z}_{\mathrm{i}})}\left(W_{(j+1)} - W_{(j)}\right)}{S_0(C_i)\exp\left(\lambda^T\mathbf{Z}_i\right)}\right], \quad (6)$$

where the boxes highlight the differences from the correct formula, Equation (2). The impact of using an indicator function with an exclusive, $\mathrm{I}(W_{(j)} > C_i)$, rather than inclusive, $\mathrm{I}(W_{(j)} \geq C_i)$, inequality is discussed in Section 3.2. For now, we examine the term $\left\{S_0\left(W_{(j+1)}\right) + S_0\left(W_{(j)}\right)\right\}^{\boxed{\exp\left(\lambda^T\mathbf{Z}_i\right)}}$, which incorrectly assumes

$$\left\{S_0\left(W_{(j+1)}\right) + S_0\left(W_{(j)}\right)\right\}^{\exp\left(\lambda^T\mathbf{Z}_i\right)} = \left\{S_0\left(W_{(j+1)}\right)^{\exp\left(\lambda^T\mathbf{Z}_i\right)} + S_0\left(W_{(j)}\right)^{\exp\left(\lambda^T\mathbf{Z}_i\right)}\right\} . \quad (7)$$

This equality is only guaranteed to hold when $\boldsymbol{\lambda}^T\mathbf{Z}_i = 0$; otherwise, Equation (6) is incorrect. By properties of the survival function, recall that $0 \leq S_0(t) \leq 1$ for all t. With this in mind, it can be shown that

$$\left\{S_0\left(W_{(j+1)}\right)^{\exp\left(\lambda^T\mathbf{Z}_i\right)} + S_0\left(W_{(j)}\right)^{\exp\left(\lambda^T\mathbf{Z}_i\right)}\right\} < \left\{S_0\left(W_{(j+1)}\right) + S_0\left(W_{(j)}\right)\right\}^{\exp\left(\lambda^T\mathbf{Z}_i\right)} \quad (8)$$

if $\boldsymbol{\lambda}^T\mathbf{Z}_i < 0$ and the opposite relationship (i.e., >) holds if $\boldsymbol{\lambda}^T\mathbf{Z}_i > 0$. As such, Equation (6) will overestimate or underestimate the correct conditional means when $\boldsymbol{\lambda}^T\mathbf{Z}_i < 0$ or > 0, respectively.

Next, the formula in Atem et al. (2019b) computes $\widehat{\mathrm{E}}(X \mid X > C_l, \mathbf{Z}_l)$ as

$$C_i + \frac{1}{2}\left[\frac{\sum_{j=1}^{\boxed{n}}\boxed{\mathrm{I}\left(W_{(j)} > C_i\right)}\left\{S_0\left(W_{(j+1)}\right) + S_0\left(W_{(j)}\right)\right\}\boxed{\exp\left(\lambda^T\mathbf{Z}_i\right)}\left(W_{(j+1)} - W_{(j)}\right)}{S_0(C_i)\boxed{\exp\left(\lambda^T\mathbf{Z}_i\right)}}\right] . \quad (9)$$

Since both the numerator and denominator are being multiplied by $\exp(\boldsymbol{\lambda}^T\mathbf{Z}_i)$, this term cancels out. Effectively, this gives $C_i + \left\{\int_{C_i}^{\infty} S_0(x)\mathrm{d}x\right\}/S_0(C_i)$, which ignores available covariate information from $\mathbf{Z}_i$.

The final formula comes from Atem et al. (2019a), which states that $\widehat{\mathrm{E}}(X \mid X > C_i, \mathbf{Z}_i) =$

$$C_i + \frac{\sum_{j=1}^{\boxed{n}} \boxed{\mathrm{I}\left(W_{(j)} > C_i\right)} \left\{ S_0\left(W_{(j+1)}\right)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} + S_0\left(W_{(j)}\right)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} \right\} \left(W_{(j+1)} - W_{(j)}\right)}{\boxed{2^{\exp\left(\lambda^T \mathbf{Z}_i\right)}} S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)}}. \tag{10}$$

This formula will be incorrect for all $\boldsymbol{\lambda}^T\mathbf{Z}_i \neq 0$. Specifically, for all $\boldsymbol{\lambda}^T\mathbf{Z}_i < 0$, Equation (10) will overestimate the imputed value, since $(1/2)^a > 1/2$ for all $a < 0 < 1$. Similarly, when $\boldsymbol{\lambda}^T\mathbf{Z}_i > 0$ the imputed values are underestimated. The degree of over-/underestimation worsens as $\boldsymbol{\lambda}^T\mathbf{Z}_i$ deviates from 0. In addition, the upper bounds of the sums for Equations (9) and (10) are invalid because we cannot evaluate at $W_{(n+1)}$.

3.2 | Underestimating the integra

Note that Equation (2) is equivalent to

$$C_i + \frac{1}{2}\left[\frac{\sum_{j=1}^{n-1} \mathrm{I}\left(W_{(j)} > C_i\right) \left\{ S_0\left(W_{(j+1)}\right)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} + S_0\left(W_{(j)}\right)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} \right\} \left(W_{(j+1)} - W_{(j)}\right)}{S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)}}\right] \tag{11}$$

$$+ \frac{1}{2}\left[\frac{\left\{ S_0(m_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} + S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} \right\} (m_i - C_i)}{S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)}}\right], \tag{12}$$

where $m_i = \min\{W_{(j)} : W_{(j)} > C_i\}$, that is, the first observed value is greater than C_i. Thus, using the exclusive indicator $\mathrm{I}(W_{(j)} > C_i)$, as in Equations (6)–(10), leads to underestimation of the integral, $\int_{C_i}^{\infty} S_X(x \mid \mathbf{Z}_i)\mathrm{d}x$, and biased imputed values. Underestimation does not occur only if $S_0(m_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} = S_0(C_i)^{\exp\left(\lambda^T \mathbf{Z}_i\right)} = 0$.

4 | EMPOWERING FUTURE USABILITY

To help others correctly use conditional mean imputation, we implemented it in the imputeCensoRd R package available on GitHub (see the Supporting Information). We believe that this note and software will broaden adoption of the conditional mean imputation approach for covariate censoring. We note that Atem et al. (2017) contain, in addition to the single conditional mean imputation approach discussed herein, a valid multiple imputation based on the predictive distribution of the censored covariate. Important areas of future work include handling missing data in Y and/or $\boldsymbol{Z}$ (though we note that Atem et al., 2019a, accommodate multiple censored $\boldsymbol{X}$) and imputation for covariates that are censored by a limit of detection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research was supported by the National Institute of Environmental Health Sciences grant T32ES007018 and the National Institute of Neurological Disorders and Stroke (NINDS) grant K01NS099343.

Funding information

National Institute of Environmental Health Sciences, Grant/Award Number: T32ES007018; National Institute of Neurological Disorders and Stroke, Grant/Award Number: K01NS099343; Bundesministerium für Bildung und Forschung, Grant/Award Number: LivSysTransfer (031L0119); Deutsche Forschungsgemeinschaft, Grant/Award Number: RTG 2624 (427806116)

DATA AVAILABILITY STATEMENT

Data sharing not applicable – no new data generated.

REFERENCES

Atem FD, Matsouaka RA, & Zimmern VE (2019a). Cox regression model with randomly censored covariates. Biometrical Journal, 61, 1020–1032. [PubMed: 30908720]

Atem FD, Qian J, Maye JE, Johnson KA, & Betensky RA (2017). Linear regression with a randomly censored covariate: Application to an Alzheimer's study. Journal of the Royal Statistical Society. Series C, Applied Statistics, 66, 313–328. [PubMed: 28239197]

Atem FD, Sampene E, & Greene TJ (2019b). Improved conditional imputation for linear regression with a randomly censored predictor. Statistical Methods in Medical Research, 28, 432–444. [PubMed: 28830304]

Little RJA (1992). Regression with missing X s: A review. Journal of the American Statistical Association 87, 1227–1237.

Little RJA, & Rubin DB (2002). Statistical analysis with missing data. Wiley.

Royston P (2007). Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. Stata Journal, 7, 445–464.