Lava is all you need: R package reduces bias for correlations among censored variables

Jerlyn Malasig*, Monica Cordova-Medina*, LaShawn Tith*, Fitsum A. Ayele, and Kimberly A. Barchard

Department of Psychology, University of Nevada, Las Vegas

Author Note

*These authors contributed equally to this poster.

Poster to be presented at the Office of Undergraduate Research Symposium at University of Nevada, Las Vegas.

Contact Information: Kimberly A. Barchard, Department of Psychology, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, P.O. Box 455030, Las Vegas, NV, 89154-5030, USA, kim.barchard@unlv.edu

ABSTRACT

Data censoring occurs when researchers have only partial information about the value of a variable. For example, Carhart-Harris et al. (2018) studied treatment-resistant depression among participants taking psilocybin (magic mushrooms). If participants took extra psilocybin outside of the study context, then the dosage is known to be at least as much as a certain value, but it might have been higher. Left censoring occurs when the left-hand side of a distribution is obscured by censoring; right censoring when the right-hand side is obscured. The R package *lava* can estimate the correlation that would have been obtained between the uncensored variables when provided with the data from the censored variables. We conducted a Monte Carlo study to evaluate the extent to which *lava* estimates are biased for data sets of 500 cases with various correlations (-.95, -.50, -.05, .25, .50, and .95) and various degrees of left censoring (10% on both variables, 50% on both, 20% on one and 80% on the other, and 95% on both). When there was low to moderate censoring, *lava* estimates were unbiased. However, when there was 95% censoring on both variables, *lava* estimates were biased. When the correlation was -.05 or -.50, bias was large and negative (-.24 or -.35, respectively). For other correlations, bias was typically moderate (e.g., -.02 to .06). If researchers are interested in negative correlations between variables that may be left censored, we recommend they minimize censoring to avoid biased estimates.

INTRODUCTION

Data censoring is highly pervasive in research within various fields of study. Censoring is a condition in which the value of a measurement is partially observed. It occurs when there are measurement gaps below or above a certain threshold or between two data points. Distorted analysis may occur if censored data is not addressed. To account for these circumstances, certain methods are used in order to reduce bias in results. We will go into further detail on censored data by reviewing a study regarding psilocybin mushrooms, commonly known as magic mushrooms. The issues that occur within various types of research because of censored data, including psilocybin studies, are outlined here. The method that we will be assessing is the R package *lava* because it can be used when there is censoring on more than one variable. In the present study, we will evaluate its performance on estimating the correlation between uncensored variables given data from censored variables.

Several types of censoring are identified in research. First, right censoring is a common occurrence in survival and longitudinal studies where the time-to-event is being recorded but does not occur within the specified duration of the study for certain participants (Gijbels, 2010). In this case, the right-hand tail of the distribution consisting of the upper values is obscured. Second, left censoring occurs when the event of interest occurs prior to the commencement of the study, or because the limit of detection which the measurement device

lacks the sensitivity to capture all data points (Gijbels, 2010). In this case, the left-hand tail of the distribution is obscured.

A variety of studies are being conducted to test the effect of psilocybin on treatment-resistant depression. One such study by Carhart-Harris et al. (2017) was able to find decreased depressive symptoms observed through levels of serotonin which is a neurotransmitter commonly associated with stabilizing mood and reducing depression. Within this study, levels of serotonin were observed through blood oxygen level-dependent functional magnetic resonance imaging (fMRI) in 16 out of the 19 patients (Carhart-Harris et al., 2017). The fMRI did not record serotonin levels in the remaining three participants which could be due to lower levels in which the fMRI could not detect. Undetected serotonin would lead to left censoring in the data. Recording of dosage in psilocybin studies is also a concern in accuracy of measurements and could lead to right-censoring. In another study by Roseman et al. (2018), the effects of psilocybin on 20 volunteers with treatment resistant depression was observed. Volunteers were administered 10 milligrams of psilocybin at the beginning of the study and then another 25 milligrams a week later. They received final assessment at week five. The researchers can be sure that the volunteers consumed at least 35 milligrams of psilocybin within the study. It is uncertain if the volunteers took more than the amount given by the researchers during the 5-week study timeframe. If additional psilocybin was taken by any of the participants, outside of what was given by the researcher, right censoring would have occurred.

Many censored data analysis procedures exist for studies with one censored variable, but hardly any exist for studies with more than one censored variable. In this study, we account for more than one censored variable and estimate the correlation of those two variables. In terms of which type of censoring occurs, *lava* is set up to use the same formulas for left and right censoring where for left censoring the inverse of the observed data is taken account of (Barchard & Russel, 2020). This is due to left and right censoring catering to the opposite direction of distributions. We examined a method to estimate correlation, ρ_{XY}, from censored data as correlation allows the assessment for strength of relationship between two variables. Given two uncensored variables X and Y, and two censored variables x and y, where x covers part of X, and y covers a part of Y, *lava* allows us to estimate the correlation between X and Y, with observations from x and y. To test out *lava*'s ability to estimate ρ_{XY}, we conducted a Monte Carlo study to observe the calculated bias from the mean difference between the estimates of ρ_{XY} and the true values of ρ_{XY}.

METHOD

In order to evaluate the accuracy of R package *lava* in estimating the correlation between uncensored variables X and Y based on data from censored variables x and y, we varied the patterns of censoring for x and y along with ρ_{XY} while keeping a fixed sample size of 500. In total, we ran 30 cells where each cell represents a unique combination of ρ_{XY}, censoring pattern, and sample size.

Patterns of censoring on x and y included: 10% censoring on both, 50% censoring on both, and 95% censoring on both representing small, moderate, and heavy patterns of censoring, respectively. We also included a mixed censoring pattern of 20% censoring on x and 80% censoring on y. This variety were selected to observe whether certain patterns influenced the accuracy of estimates.

ρ_{XY} values of -.95, .95, -.5, .5, -.05, and .25 were used. ρ_{XY} of $\pm$.95 was included to observe how a strong correlation and its direction could influence *lava*'s performance. A moderate ($\pm$.50) and minimal (-.05 and .25) ρ_{XY} were used for comparison.

For each cell, we ran 1000 trials in which we generated a random set of data for which X and Y had a bivariate normal distribution. We then censored x and y to the required degrees. We provided the x and y data to *lava* and asked it to estimate the correlation between X and Y (ρ_{XY}). To assess *lava*'s performance, we calculated bias. Bias was calculated as the mean difference between actual values of ρ_{XY} and *lava* estimates of ρ_{XY}.

RESULTS

A table was created to represent the results from the Monte Carlo study. See Table 1. The table shows the bias for *lava* estimates of ρ_{XY} for each unique combination of censoring pattern, ρ_{XY}, and sample size. The *lava* estimates of ρ_{XY} for all cells with censoring patterns of 10% on x and y, 50% on x and y, and 20% on x and 80% on y were unbiased. The *lava* estimates of ρ_{XY} were biased when the censoring pattern was 95% on x and y; bias was largest when the initial correlation was negative.

Table 1

Bias for lava Estimates of ρ_{XY} for Different Values of ρ_{XY} Under Different Patterns of Censoring

	Patterns of Censoring			
Correlation	10% x 10% y	50% x 50% y	20% x 80% y	95% x 95% y
-.95	.00	.00	.00	.06
-.50	.00	.00	.00	-.35
-.05	.00	.00	.00	-.24
.25	.00	.00	.00	-.04
.50	.00	.00	.00	-.02
.95	.00	.00	.00	.00

Note. Sample size was fixed at 500.

DISCUSSION

The results from our Monte Carlo study showed that *lava* estimates of ρ_{XY} for all unique combinations of sample size, ρ_{XY}, and censoring patterns of up to 50% were unbiased. The small difference between ρ_{XY} and *lava* estimates of ρ_{XY} means *lava* produced estimates close to the true correlation of uncensored variables. Mixed censoring on X and Y (20% and 80%) produced similar unbiased results. When the censoring pattern was high (95% on both x and y) and the correlation was negative, however, there was substantial bias. It was found that if there was same direction censoring on both variables (either both left or both right), *lava* produces severe biased estimates for certain negative correlations (low and moderate). However, if we had high left censoring on one variable and high right censoring on the other, we would have found that *lava* produced bias estimates for positive correlations. Given our results, we are confident of R package *lava's* performance when degrees of censoring are low to moderate on at least one of the variables. We recommend R package *lava* to other researchers in their studies with low to moderate censoring and encourage them to minimize censoring to avoid biased estimates.

The R-package *lava* assumes a normal distribution for the values of X and Y. If data sets reflect a distribution that is not normal (i.e. skewed), *lava* estimates could differ and potentially be even more bias. Therefore, it is recommended that future researchers investigate how different distribution patterns affect censored data analysis. Additionally, within this study we looked into the effect *lava* holds on estimating for correlation, future studies should look into regression models that will allow for estimations of future values and the regression of current data points, even when censoring occurs.

References

Barchard, K. A., & Russel, J. A. (2020) *Modelling and correcting the effect of data point censoring on correlations*. Unpublished manuscript. Department of Psychology, University of Nevada, Las Vegas.

Carhart-Harris, R. L., Roseman, L., Bolstridge, M., Demetriou, L., Pannekoek, J. N., Wall, M. B., Tanner, M., Kaelen, M., McGonigle, J., Murphy, K., Leech, R., Curran, H. V., & Nutt, D. J. (2017). Psilocybin for treatment-resistant depression: FMRI-measured brain mechanisms. *Scientific Reports*, *7*(1), *13187–11*. https://doi.org/10.1038/s41598-017-13282-7

Gijbels, I. (2010). Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(2), 178–188. https://doi.org/10.1002/wics.80

Roseman, L., Nutt, D. J., & Carhart-Harris, R. L. (2018). Quality of acute psychedelic experience predicts therapeutic efficacy of psilocybin for treatment-resistant depression. *Frontiers in Pharmacology*, 8:974. https://doi.org/10.3389/fphar.2017.00974