Censored Data: Evaluating Two Methods of Estimating Correlations within R package *Lava*

Cassandra K. Hoffman, Eden K. Thiess, Fitsum A. Ayele, and Kimberly A. Barchard
University of Nevada, Las Vegas

Reference: Hoffman, C. K., Thiess, E. K., Ayele, F. A., & Barchard, K. A. (2021, June 14-15). *Censored data: Evaluating two methods of estimating correlations within R package lava*. [Poster presentation]. American Association of Behavioral and Social Sciences Annual Conference, Las Vegas, NV, United States.
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

Censored data, which occur when exact values of variables are unknown, can distort statistical results and invalidate conclusions. Holst et al.'s (2015) R package *lava* can be used to estimate bivariate correlations for censored data using a correlation model or a regression model. To compare these two models, the present study simulated normally distributed data with various correlations and sample sizes, and then censored them to the required degrees. For both models, *lava* always converged on an estimate and bias was minimal unless both variables had 80% censoring and the correlation was large and negative. Moreover, bias decreased when sample size increased from 500 to 1000. Although both models produced unbiased correlation estimates when samples were large and censoring was low to moderate, the regression model ensured estimates were valid (between -1 and 1). Future research may wish to investigate *lava's* performance for non-normal distributions and smaller sample sizes.

INTRODUCTION

Within the realm of data analysis, one obstacle that arises across many fields of research is determining how to correct for the effects of censored data. Censored data occur when researchers are only able to obtain partial information about the value of a variable. They may know the value is less than or equal to (or greater than or equal to) some limit of detection, but they are unable to determine the variable's true value. This can happen if a measuring device is unable to detect the value of a variable outside a specific range of concentration levels. Alternatively, in the context of longitudinal/survival studies, censoring may occur when the event of interest does not have enough time to take place. This may happen if, for example, some participants stop reporting back, and/or the event is not recorded for some participants before the study concludes.

Two common types of censoring that occur are left and right censoring. Left censoring occurs when the value of a variable is greater than zero but less than or equal to the lower limit of detection (the lowest amount of a substance or concentration that can be distinguished from the absence of any substance) (Pesonen et al., 2015). For example, consider a rehabilitation clinic that is issuing drug tests to its residents. There is a specific concentration of a given drug that must be present in a sample for it to produce a positive test result. However, if the drug is present at a concentration level that falls at or below the limit of detection, a negative test result will be given. Similarly, right censoring occurs when the value of a variable is greater than or equal to an upper limit of detection. Moreover, it arises when the event of interest has not occurred by a specified point in time and is a common occurrence in longitudinal studies (Leung et al., 1997). For example, if researchers conducted a study to determine the survival rate of patients with dementia, right censoring would occur if some participants survived past the conclusion of the study or were lost to follow up. In these instances, the researchers would only know that the survival times of these participants were at least as great as the timepoint at which the study ended (or participants were lost to follow up), but they would not know the exact values.

Censored data often occur in the field of psychopharmacology. When researchers study the effects of medications, there are often differences that fall outside the limit of detection or after the study has concluded. For example, one study examined the relationship between antidepressant use and suicide attempts among adolescents (Valuck et al., 2004). To do this, researchers examined insurance claims and prescription refills for adolescents (aged 12-18). Inclusion criteria involved a diagnosis of Major Depressive Disorder between 1998 and 2003, as well as follow up information for at least six months. After analyzing the data, no significant link was found between antidepressant consumption and suicide attempts (Valuck et al., 2004). This is an example of a study that is susceptible to right censored data. In the experiment, the researchers accounted for these suicide attempts by looking at the insurance claims. It is possible that attempts may have occurred that were either treated at home, were not severe enough to make an insurance claim, or took place after the study concluded.

It is important to note that if both variables are censored in the same direction (both left-censored or both right-censored), bias is greatest for negative correlations. Only when one variable is right-censored and the other variable left-censored, do we see greater bias for positive correlations. If researchers do not correct for the effect of censored data, they can potentially distort statistical results and invalidate conclusions (Barchard & Russell, 2020). Different techniques have been proposed to correct for censored data including imputation, complete-data analysis, and likelihood-based approaches (Leung et al., 1997). These methods suggest replacing incomplete data with substituted values, removing incomplete data from analysis, and estimating values from available data, respectively. A vast majority of these methods allow censoring on either the predictor or the criterion variable, but not both. There are few existing methods that are capable of estimating correlations when both variables have been censored. As correlational studies are paramount to providing insight into the relationships between variables, researchers need an accurate way of assessing their data when censored values are present for both variables. One such method with this capability was introduced by Holst et al. (2015) and has been implemented in the R package *lava*.

Holst's R package *lava* contains both a correlation model and a regression model that can be used to estimate correlations from censored data. Given two uncensored variables X and Y, and two censored variables x and y, where x covers part of X, and y covers part of Y, it is possible to use either model to estimate the correlation between X and Y using the available data from x and y. This research seeks to compare the correlation model with the regression model in terms of bias (the mean difference between *lava* estimates ($\hat{\rho}_{XY}$) and true values (ρ_{XY})) and determine which model produces the least biased estimates. We conducted a Monte Carlo simulation study to examine the effect of left censoring (on both x and y) and assess the accuracy and precision of *lava's* estimates. Additionally, the proportion of trials where *lava* failed to converge on an estimate were observed to determine which model is more reliable for producing estimates.

METHOD

To assess the conditions that lead to good and poor performance for the estimates, we simulated a large range of population correlations, large sample sizes, and a wide variety of censoring patterns. Our study had a total of 84 cells. We selected ρ_{XY} values of .25, .50, .75, -.25, -.50, and -.75 to assess *lava's* accuracy in estimating correlations for different ρ_{XY} values. We chose these values to ensure that the results are generalizable and comprehensive, covering an array of different correlation values, so that they can be applicable to many different datasets. We also chose large sample sizes of 500 and 1000 to provide R package *lava* with enough information to create the best possible estimates for ρ_{XY} that it could. Additionally, we used seven patterns of censoring to see the effects of this censoring in a variety of situations. Our censoring patterns include 20% and 20%, 40% and 40%, 60% and 60%, 80% and 80%, 20% and 40%, 20% and 60%, and 20% and 80%.

In our study, we ran 1000 trials for each cell and then replicated each cell for a total of 2000 trials for each combination of factors. Within each of these trials, we generated a random dataset for which the predictor and criterion variables had a multivariate normal distribution with the specified ρ_{XY} value and sample size, and then we censored that data to the required degree. We provided this data to *lava* and asked it to estimate the true population correlation for each trial (for both the correlation model and the regression model) and calculate the proportion of trials for which it failed to converge on an estimate. To determine the quality of these estimates, we calculated bias (the difference between $\hat{\rho}_{XY}$ and ρ_{XY}) for each trial and then the mean bias (the mean difference between $\hat{\rho}_{XY}$ and ρ_{XY}) for each cell. Lastly, to summarize the results across the 84 cells and assess how different combinations of factors affect *lava's* estimates, we used between-within analysis of variance (ANOVA) and graphed significant interactions. The between-cell factors used for comparison were ρ_{XY}, sample size, and censoring pattern. The within-cell factor used for comparison was the estimation method (correlation or regression). The outcome variable was bias.

RESULTS

First, we assessed the proportion of trials where *lava* failed to converge on an estimate. We found that *lava* was able to converge on an estimate 100% of the time, regardless of the parameters set and the model used (correlation or regression).

Second, we assessed bias for all conditions. See Table 1. Overall, bias was only found when the population correlation was large and negative, or small and positive, and both variables were censored at 80%. For the sample size of 500, we found bias when the correlation was small and positive (.25), or moderate to large and negative (-.05 and -.75), and both variables were censored at 80%. For the sample size of 1000, we found bias when the correlation was moderate to large and negative (-.05 and -.75), and both variables were censored at 80%.

Third, we conducted between-within ANOVA and graphed significant interactions. Two three-way interaction effects were found. The first significant interaction was between ρ_{XY}, censoring pattern, and sample size, $F(30, 84) = 11.32$, $p < .001$. Figure 1 shows the effect of ρ_{XY} and censoring pattern on bias for the sample size set to 500. Figure 2 shows the effect of ρ_{XY} and censoring pattern on bias for the sample size set to 1000. These graphs show that bias is greatest when ρ_{XY} is strong and negative and censoring is high. Additionally, they show that increasing the sample size reduces that bias. The second significant interaction was between ρ_{XY}, censoring pattern, and model, $F(30, 84) = 10.73$, $p < .001$. Figure 3 shows the effect of ρ_{XY} and censoring pattern on bias for the correlation model. Figure 4 shows the effect of ρ_{XY} and censoring pattern on bias for the regression model. These figures again show that bias is largest when ρ_{XY} is strong and negative and censoring is high. Notably, while ANOVA calculations found a significant interaction between ρ_{XY}, censoring pattern, and model, the differences between models were trivial. The biggest difference found between models occurred when the population correlation value was -.75 and both variables were censored at 80% (the difference between models for this condition was < .0004). However, one advantage of the regression model is that it always produces estimates within the allowable range for a correlation [-1,1].

DISCUSSION

In our research we sought to assess the ability of R package *lava* to provide correlation estimates from censored data. We investigated the accuracy of point estimates of ρ_{XY} for a wide variety of initial values of ρ_{XY}, sample sizes, and censoring patterns, across the correlation and regression models. Overall, we found that the differences in estimation methods (correlation and regression) were trivial.

Our results showed that bias was minimal unless the correlation was large and negative and both variables had 80% censoring. Although both the correlation model and the regression model seem to be reasonably effective and accurate methods of providing estimates for ρ_{XY}, the regression model is preferred as it ensures estimates fall within the allowable range for correlations [-1, 1]. Based on our findings, researchers can confidently use the regression model to estimate ρ_{XY} for most initial values of ρ_{XY}, sample sizes, and patterns of censoring. Furthermore, when the initial value of ρ_{XY} is large and negative and censoring is high, our results show that both models produce substantially biased estimates. Therefore, if researchers are interested in large, negative correlations, they should make every attempt to minimize censoring. To minimize left censoring, they could use more fine-grained measuring devices with lower limits of detection. In the context of survival studies, in order to limit the occurrence of right censored points, researchers may wish to extend their studies to capture more of the target event in their data. In conclusion, we turn to the initial question posed at the start of the study: Are point estimates of ρ_{XY} more biased for the correlation model or the regression model? Considering that both models produced extremely similar results, we conclude that either the correlation or regression model may be used to estimate ρ_{XY}. However, the regression model is superior because it always produces estimates between -1 and 1.

Our study did not have any instances where *lava* failed to converge, further emphasizing the effectiveness of using R package *lava* in studies. The results of this study show that using R package *lava* with normally distributed data will provide researchers with accurate estimates for ρ_{XY}. While the correlation and regression models have been effective in providing minimally biased estimates when provided with left-censored data, there is one implication to be considered. As mentioned, and demonstrated by our results, bias is greatest for negative correlations when both variables are censored in the same direction. Although our study used left-censored data for both variables, we expect research findings to be similar to ours when using right-censored data for both variables. Future research could examine this supposition.

Our study also contains a few limitations that should be considered for future research. The first limitation is that all of the data provided in the study had a normal distribution. Therefore, if research data are provided which do not reflect this normal distribution, it is possible that the results may be more or less biased than the results our study showed. Similarly, if the data set contains numerous outliers, it is possible that the estimates provided may be more or less biased. A final limitation to consider is the sample sizes used. In our study, we only ran moderate and large sample sizes. We did not test to see the effects that the study would have on a much smaller sample size. The significant three-way interaction between sample size, censoring pattern, and ρ_{XY} suggests that increasing sample size for strong negative correlations with highly censored values reduces bias. Therefore, a much smaller sample size than the one used in this study could result in substantially biased results.

Future research could examine some of the limitations of our study. Trials could be done to assess ***lava's*** accuracy when estimating population correlations using data that do not have a normal distribution (i.e., skewed, polychotomous, bimodal, or trimodal data) and/or contain outliers. Additionally, trials could be done to assess ***lava's*** performance when working with smaller sample sizes. While the performance of *lava* under these conditions is not yet known, *lava* has been shown to provide accurate estimates under several conditions and appears to be a useful method of reducing bias when data are censored.

REFERENCES

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

Holst, K. K., Budtz-Jørgensen, E., & Knudsen, G. M. (2015). A latent variable model with mixed binary and continuous response variables. *arXiv preprint arXiv:1507.01182*.

Leung, K. M., Elashoff, R. M., & Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health, 18*(1), 83–104. https://doi.org/10.1146/annurev.publhealth.18.1.83

Luo, X., Le, C. T., Chu, H., Epstein, L. H., Yu, J., Ahluwalia, J. S., & Thomas, J. L. (2013). Analysis of cigarette purchase task instrument data with a left-censored mixed effects model. *Experimental and Clinical Psychopharmacology, 21*(2), 124–132. https://doi-org.ezproxy.uvu.edu/10.1037/a0031610

Pesonen, M., Pesonen, H., & Nevalainen, J. (2015). Covariance matrix estimation for left-censored data. *Computational Statistics & Data Analysis*, *92*, 13-25. https://doi.org/10.1016/j.csda.2015.06.005

Valuck, R. J., Libby, A. M., Sills, M. R., Giese, A. A., & Allen, R. R. (2004). Antidepressant treatment and risk of suicide attempt by adolescents with major depressive disorder. A propensity-adjusted retrospective cohort study. *CNS Drugs, 18*(15), 1119-1132. https://doi.org/10.2165/00023210-200418150-00006

Table 1

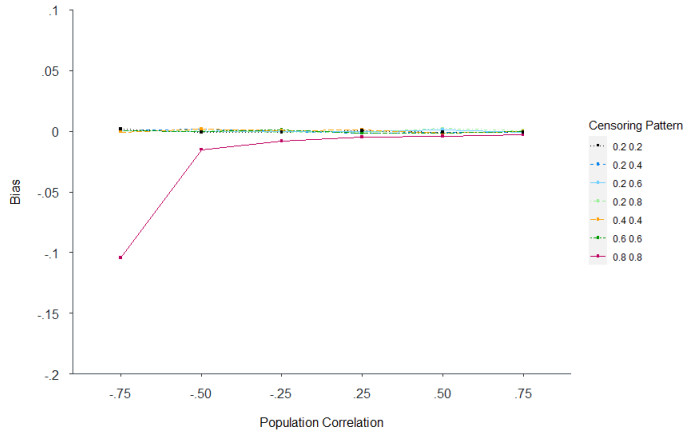
Bias for Lava's Correlation and Regression Models

Model	Censoring	ρ_{XY}					
		-.75	-.5	-.25	.25	.5	.75
		N = 500					
Correlation	20% x 20% y	.00	.00	.00	.00	.00	.00
	40% x 40% y	.00	.00	.00	.00	.00	.00
	60% x 60% y	.00	.00	.00	.00	.00	.00
	80% x 80% y	-.10	-.01	.00	-.01	.00	.00
	20% x 40% y	.00	.00	.00	.00	.00	.00
	20% x 60% y	.00	.00	.00	.00	.00	.00
	20% x 80% y	.00	.00	.00	.00	.00	.00
Regression	20% x 20% y	.00	.00	.00	.00	.00	.00
	40% x 40% y	.00	.00	.00	.00	.00	.00
	60% x 60% y	.00	.00	.00	.00	.00	.00
	80% x 80% y	-.10	-.01	.00	-.01	.00	.00
	20% x 40% y	.00	.00	.00	.00	.00	.00
	20% x 60% y	.00	.00	.00	.00	.00	.00
	20% x 80% y	.00	.00	.00	.00	.00	.00
		N = 1000					
Correlation	20% x 20% y	.00	.00	.00	.00	.00	.00
	40% x 40% y	.00	.00	.00	.00	.00	.00
	60% x 60% y	.00	.00	.00	.00	.00	.00
	80% x 80% y	-.07	-.01	.00	.00	.00	.00
	20% x 40% y	.00	.00	.00	.00	.00	.00
	20% x 60% y	.00	.00	.00	.00	.00	.00
	20% x 80% y	.00	.00	.00	.00	.00	.00
Regression	20% x 20% y	.00	.00	.00	.00	.00	.00
	40% x 40% y	.00	.00	.00	.00	.00	.00
	60% x 60% y	.00	.00	.00	.00	.00	.00
	80% x 80% y	-.07	-.01	.00	.00	.00	.00
	20% x 40% y	.00	.00	.00	.00	.00	.00
	20% x 60% y	.00	.00	.00	.00	.00	.00
	20% x 80% y	.00	.00	.00	.00	.00	.00

Note. This table shows the mean bias (the difference between *lava's* population correlation estimates and the true correlation values) for various combinations of model, censoring pattern, population correlation, and sample size.

Figure 1

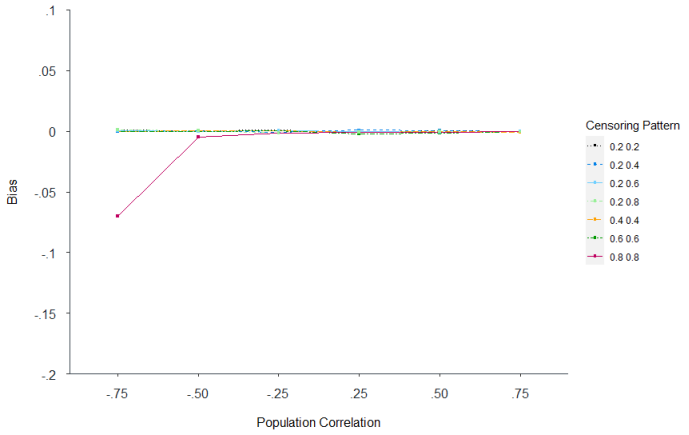
Effect of Population Correlation, Censoring Pattern, and Sample Size on Bias (Sample Size set to 500)



Note. To demonstrate the effect of population correlation, censoring pattern, and sample size on bias (the mean difference between *lava's* estimate and the true correlation value), bias was averaged across the correlation model and the regression model.

Figure 2

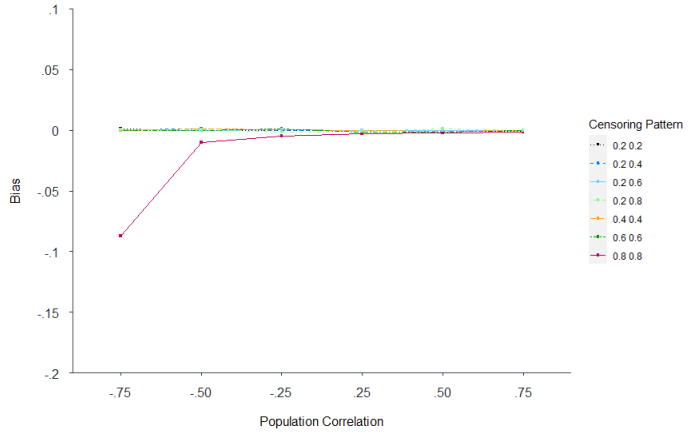
Effect of Population Correlation, Censoring Pattern, and Sample Size on Bias (Sample Size set to 1000)



Note. To demonstrate the effect of population correlation, censoring pattern, and sample size on bias (the mean difference between *lava's* estimate and the true correlation value), bias was averaged across the correlation model and the regression model.

Figure 3

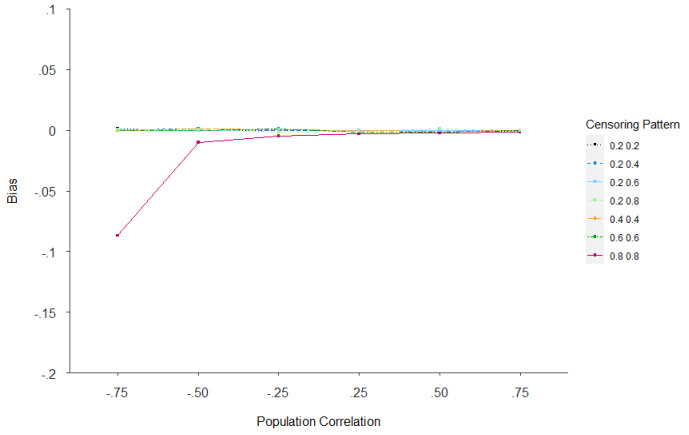
Effect of Population Correlation, Censoring Pattern, and Model on Bias (Correlation Model)



Note. To demonstrate the effect of population correlation, censoring pattern, and model on bias (the mean difference between *lava's* estimate and the true correlation value), bias was averaged across sample sizes (500 and 1000).

Figure 4

Effect of Population Correlation, Censoring Pattern, and Model on Bias (Regression Model)



Note. To demonstrate the effect of population correlation, censoring pattern, and model on bias (the mean difference between *lava's* estimate and the true correlation value), bias was averaged across sample sizes (500 and 1000).