Evaluating the Quality of Correlation Estimates Under Data Point Censoring

Fitsum A. Ayele and Kimberly A. Barchard

Reference: Ayele, F. A., & Barchard, K. A. (2020, November 16-20). *Evaluating the Quality of Correlation Estimates Under Data Point Censoring* [Poster presentation]. University of Nevada, Las Vegas Office of Undergraduate Research Virtual Undergraduate Research Symposium, Las Vegas, NV.

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 are a common but routinely unrecognized phenomenon in psychological research. When a data point is censored, it means we only have partial information on its true value, knowing the value is no larger than, or at least as large as, a given limit of detection. Censoring can distort results and invalidate conclusions. Fortunately, there are a variety of methods that correct the effect of censoring when estimating univariate statistics like medians and means. However, few methods exist that allow researchers to estimate a correlation when both variables have been censored. Therefore, the purpose of this paper is to evaluate a method for estimating correlations with censored data. The method in question is implemented in the *lava* package in R. Using *lava*, it is possible to estimate the theoretical correlation between two uncensored variables (ρ_{XY}), based upon the available data from the censored variables. In the present study, we aimed to assess the accuracy of *lava* estimates of ρ_{XY}. We used bias as an indicator of the accuracy of the estimates. Bias was calculated as the mean difference between the estimate ($\hat{\rho}_{XY}$) and the true value (ρ_{XY}). Usually, the estimates were unbiased. Any noteworthy bias was limited to negative values of ρ_{XY} and high degrees of censoring. If researchers use this method with this type of data, their estimates will be closer to -1 than they should be. Therefore, if researchers are investigating large negative correlations, they should attempt to minimize censoring.

Introduction

Oftentimes data in many settings are censored. When a data point is censored, it means that we only have partial information on its true value, knowing the value at least as large as (or no larger than) a given limit of detection (Fox, 2016). Censoring can also occur when researchers are recording the time until some event, and the event in question may or may not have occurred before/after a specified point in time. Censoring can skew results and invalidate conclusions (Barchard & Russell, 2020). Fortunately, there are a wide variety of tools available to researchers to correct the effect of censoring when estimating single variable statistics like medians and means. However, few methods exist that deal with correlations when both variables have been censored, and little is known about the effectiveness of these methods. Correlations are used to examine a plethora of relationships in psychological research and they are an important statistic in many datasets. Therefore, the aim of this study is to evaluate a method for estimating the correlation between two variables when they are both censored.

Censoring occurs in many circumstances. Here, we will outline just a couple scenarios. As mentioned above, if a data value is censored, then its true value is only partially observed. Censoring can occur due to difficulties in detecting the phenomenon. Usually, when using a measuring instrument or scale to measure the concentration of a thing of interest, a certain amount of the substance needs to be present in the sample before it can be detected (Wang et al., 2014). This is called the limit of detection. To illustrate, a certain level of pollutant in water must be present before a measuring instrument can detect it. When no pollutant is detected, the concentration might be any value below this lower limit. This is a case of left censorship, where the true value is less than or equal to the recorded value (Fox, 2016). In this scenario, researchers only know that the true value of the data point is less than or equal to the observed value, but they do not know the actual value. The data here are referred to as *left censored* because the left-hand tail of the frequency distribution (which shows low values of the variable) has been obscured. There are also cases where data values are right censored. Right censoring usually occurs when a longitudinal study is tracking the time until some event occurs, and the event is not recorded for some participants. When a data point is right censored, the event of interest does not have enough time to occur (Gijbels, 2010). An example comes from medical research. In many randomized control medical studies, scientists often wish to assess the efficacy of a drug or treatment. Unfortunately, some participants stop reporting back before the study finishes, or some participants do not experience the effect of treatment before the study ends. In these cases, researchers only know that the time to event is at least as long as the length of the study, or until people stopped reporting, but it might be longer. This is a simple example of right censorship. The term *right censored* is applied because the right-hand tail of the frequency distribution (which shows high values of the variable) has been obscured.

Unfortunately, some of the methods researchers use to deal with censored data are far from ideal. One approach is to use ad-hoc substitutions, or to arbitrarily substitute the censored value with another value (Fox, 2016). A researcher may set the censored point equal to the limit of detection, or set the value equal to zero. Another approach is to delete the censored data altogether. This effectively ignores the problem by deleting data points that have issues. Both of these approaches are very problematic because they often lead to incorrect research conclusions and do not give one a more complete picture of the dataset.

However, there are far more powerful methods of dealing with censored data that are much more acceptable. These analytic methods fall under the name of censored data analysis, and are widely available to researchers investigating censored data. A majority of these methods allow censoring on either the independent or outcome variable, but not both. In order to estimate the correlation between two censored variables, we need a method that allows censoring on both x and y. One of these methods was invented by Holst et al. (2015), and implemented in the *lava* package in R. Using *lava*, one can estimate the association between uncensored variables using the available data from censored variables. 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 *lava* to estimate the correlation between X and Y with observations from x and y (Holst et al., 2015). This method relies on maximum likelihood estimation (MLE) to estimate the covariance between uncensored variables given data from censored observations. Maximum likelihood covariance matrix estimators that allow the presence of left and right censored data have been found to produce relatively accurate estimates (Pesonen et al., 2015).

In order to test out *lava's* ability to estimate ρ_{XY}, we used the outcome variable bias. Bias was calculated as the mean difference between the estimates and the true values. Using bias, it is possible to assess the quality of the estimate in *lava*. Ideally, bias should be low (at or close to 0).

In order to evaluate if *lava's* estimates are accurate, we used a Monte Carlo study. We aimed to investigate if estimates of ρ_{XY} are biased when x and y are censored, and to assess how that bias varies for different values of ρ_{XY} and different patterns of censoring.

Method

This study used a total of 40 cells. Each cell was a unique combination of sample size, ρ_{XY}, and censoring pattern. We used a sample size of 500. We selected a large sample size in order to provide *lava* with ample information to create the best estimates of ρ_{XY} that it can. We used 10 values of ρ_{XY}: -.10, -.30, .30, -.50, .50, -.65, -.80, .80, -.95, and .95. A majority (six) of the values we chose were negative. We did this because when both variables are censored in the same direction, there is a greater effect on negative relationships. Hence, it would be more interesting and informative to assess the bias involved in correlation estimates in *lava* when the true relationship is negative. We also included four positive relationships to incorporate a wide variety of values of ρ_{XY} from which to assess the effect of censorship on the estimate. We used four different censoring patterns on x and y. Patterns that involved 20%, 40%, 75%, and 90% censoring on x and y, indicating low, moderate, heavy, and extremely heavy censorship, respectively. We selected these so that differences between the censoring patterns were large enough to observe a qualitative difference in the way the estimate behaved in terms of bias.

We ran 1000 trials for each cell. Within each trial, we produced a random set of data for which X and Y had a multivariate normal distribution with the selected correlation, and we censored x and y to the required percentages. We provided *lava* the x and y data. We then asked *lava* to estimate the correlation between X and Y (ρ_{XY}). The estimate $\hat{\rho}_{XY}$ used maximum likelihood estimation based upon the correlation model (Holst, 2020). In order to assess the quality of the estimates of ρ_{XY}, we calculated bias (the mean difference between $\hat{\rho}_{XY}$ and ρ_{XY}). Ideally, bias should be low (close to 0).

Results

We assembled the results of the Monte Carlo simulation into a table. See Table 1. Overall, we found that there was very little bias for positive values of ρ_{XY}. In addition, there was very little bias for 20% and 40% censoring across all values of ρ_{XY}. However, when it came to 75% and 90% censoring, and negative values of ρ_{XY}, there was substantial bias. An interesting finding is that for 90% censoring, if ρ_{XY} was not close to -1, the bias was negative and the estimates were closer to -1. However, if ρ_{XY} was really close to -1 (as in the value of -.95 used in this study), the bias was positive and the estimates were further from -1. This trend was not observed with 75% censoring, where the bias was always negative for negative values of ρ_{XY}, producing estimates closer to -1 than the true value of ρ_{XY}.

Table 1

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

	Estimates of Bias for Different Patterns of Censoring			
ρ_{XY}	20 %	40%	75%	90%
-0.95	.00	.00	-.02	.03
-0.80	.00	.00	-.06	-.12
-0.65	.00	.00	-.01	-.24
-0.50	.00	.00	-.01	-.25
-0.30	.00	.00	.00	-.11
-0.10	.00	.00	.00	-.02
0.95	.00	.00	.00	.00
0.80	.00	.00	.00	.00
0.50	.00	.00	.00	-.01
0.30	.00	.00	.00	-.01

Note. Percentages of censoring are the same on x and y.

Discussion

At the outset of our research, we sought to assess the efficacy of *lava* estimates of ρ_{XY} from censored observations on x and y. We used bias as an indicator of the accuracy of the estimates. We found that any noteworthy bias was limited to negative values of ρ_{XY} and high degree of censorship (75%, 90%). The bias for positive relationships and low degrees of censorship was negligible. This is understandable, because when both variables are censored in the same direction, there is a much stronger effect of censorship on negative relationships than positive ones (Barchard & Russell, 2020); also, if x covers more of X, and y covers more of Y, it would stand to reason that estimates of ρ_{XY} will be more accurate than cases where censored observations cover less of the variables of interest.

Based upon the results from the present study, we conclude that *lava* is good at estimating the correlation between uncensored variables using data from censored variables. Any appreciable bias was restricted to a few of the cells used in our study. A majority of the cells had negligible bias. Therefore, with the exception of a few cases, *lava* provides reasonably accurate estimates of ρ_{XY}.

Lava produces accurate estimates of ρ_{XY} for positive to moderately negative values of ρ_{XY} and low degrees of censoring; but, for strong negative values of ρ_{XY} and high degrees of censoring, it is not as accurate. If researchers use this method with this type of data, their estimates will be biased, leading to estimates that are closer to -1 than the true value of the population correlation. Therefore, when researchers are interested in large negative correlations, they should attempt to have minimal censoring on their data points in order to avoid substantial bias. For example, if right censoring is caused by participants not experiencing the event before the study ends, researchers may wish to extend their study to avoid an abundance of right censored points. In addition, in order to reduce left censored values, researchers can use more sensitive measuring instruments with smaller lower levels of detection, leading to a minimization of left censored points in the dataset. Despite these issues, the *lava* package in R still fares well against other methods designed to handle censored data.

The efficacy of *lava* estimates has been compared to other statistical estimators in the past. In particular, lava estimates have been compared to the limited information estimator proposed by Muthen (1984). Results show that *lava* estimates are more accurate than other estimators (Holst et al., 2015). However, Holst et al. estimated the effect of two covariates on a single binary outcome variable. Up until now, no study has investigated the accuracy of point estimates of correlations in *lava*. Thus, this study adds to the growing literature on censored data and the *lava* package by providing an assessment of the accuracy of estimates of ρ_{XY}.

Our study addressed an important question, namely, are *lava* estimates of ρ_{XY} biased when data are censored? We obtained a useful insight that can inform future work on censored data analysis. The present study used the correlation model to assess estimates of ρ_{XY} in *lava*. The regression model has also been posited by Holst (2020) as a useful way to conduct censored data analysis in *lava*. To date, there is no work investigating the accuracy of point estimates of correlations in *lava* using the regression model. Therefore, future researchers may wish to conduct simulations assessing the accuracy of ρ_{XY} estimates using the regression model. It would be interesting to see a comparison of bias across such studies with our study.

Censored data and censorship are an ever-growing concern in psychological research. Therefore, there is an increasing need for methods that effectively handle such data. In the past (and even now), researchers have often considered these data intractable, and dealt with them in undesirable ways. However, there are more effective methods of handling these data that provide researchers with a more accurate picture of their datasets. In this study, we assessed a promising approach that uses the *lava* package in R. This work effectively demonstrates that *lava* is a useful method of conducting censored data analysis. We hope that future work on censored data can make use of the *lava* package in R.

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.

Fox, Gordon [Gordon Fox] (2016, May 12). *Introduction to analysis of censored and truncated data* [Video]. Youtube. https://www.youtube.com/watch?v=aPN10YYrC1M

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

Holst, K., K. (2020). *Estimating partial correlations with lava* [Unpublished manuscript]. Department of Biostatistics, University of Copenhagen.

Holst, K. K., Budtz-Jørgensen, E., & Knudsen, G. M. (2015). *A latent variable model with mixed binary and continuous response variables.* Available at https://www.researchgate.net/publication/279864661_A_latent_variable_model_with_mixed_binary_and_continuous_response_variables

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

Wang, Y., Gay, G. D., Botelho, J. C., Caudill, S. P., & Vesper, H. W. (2014). Total testosterone quantitative measurement in serum by LC-MS/MS. *Clinica Chimica Acta: International Journal of Clinical Chemistry, 436*, 263-267. https://doi.org/10.1016/j.cca.2014.06.009