Assessing the Quality of Correlation Estimates Based upon Censored Variables

Rosalba M. Gomez*, Joshua Z. Chang*, Zain N. Raja*, Fitsum A. Ayele, & Kimberly A. Barchard

* These authors have contributed equally to this project.

Reference: Gomez, R. M., Chang, J. Z., Raja, Z. N., Ayele, F. A., & Barchard, K. A. (in prep). *Assessing the Quality of Correlation Estimates Based upon Censored Variables* [Poster presentation]. American Association of Behavioral and Social Sciences Conference, 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

UNLV

Abstract

Data censoring occurs when some values are not fully known. For example, researchers might know participants are at least 55, but not know their exact ages. Censoring distorts correlations. The R package *lava* estimates correlations between uncensored variables based upon data from censored variables. This study examined whether those estimates are biased. We examined 10 values for the correlation and four censoring patterns. For each of these 40 cells, we conducted 1000 trials with 300 cases from a multivariate normal distribution. *Lava* estimates were biased when there was at least 70% censoring on both variables, but unbiased when average censoring was no more than 50%. If researchers are interested in moderate to large correlations, they should ensure there is no more than moderate censoring.

Introduction

Researchers are often faced with censored data, which can occur in psychological studies, clinical trials, or survival studies. Censored data occur when the value of a measurement or observation is only partially known. Censored data distort results and invalidate conclusions.

One common type of censored data is left-censored data. Left-censored data occur when values fall below the limit of detection and cannot be precisely observed. For example, in a study that looked at the relationship between depression and age, left-censoring may have occurred because some participants achieved the lowest possible score on a scale of depression (Falcaro et al., 2013). These data did not distinguish among people with the lowest possible score.

Several methods effectively correct the effect of censoring in the univariate context (e.g., Canales et al., 2018). However, few methods exist to correct for the effect of censoring in the bivariate or multivariate context. Holst et al. (2015) proposed a maximum likelihood estimator that can be used in bivariate contexts and multivariate contexts with a moderate number of observed items. This model has been implemented in the *lava* package in R (Holst et al., 2015).

The purpose of our study was to evaluate the accuracy of *lava* in the bivariate context. We used *lava* to estimate ρ_{XY} (the correlation between our two uncensored variables, X and Y) based upon the data from our censored variables, x and y. We used a Monte Carlo study to determine if the estimates of ρ_{XY} are biased for different values of ρ_{XY} and different patterns of censoring.

Method

Our study used a total of 40 cells where each cell is a specific combination of sample size, ρ_{XY}, and censoring pattern. We chose a fixed sample size of 300 because this size is relatively large enough and not too large to feasibly occur in real psychological studies. We chose -.95, -.80, -.40, -.25, 0, .10, .30, .50, .85, and .90 as our values of ρ_{XY}. These 10 values were chosen to represent a wide range of possible values. Additionally, we examined four patterns of censoring: 10% censoring on x and 90% on y, 50% censoring on both, 70% censoring on both, and 95% censoring on both. These patterns were chosen to represent low, moderate, and severe as well as cases in which only one variable was highly censored.

We generated 1000 trials for each cell in our study. Within each trial, we generated a random set of data for which X and Y had a multivariate normal distribution with the desired correlation. We censored x and y to the

required degrees, and we provided the x and y data to *lava*. We then used *lava* to estimate the correlation between X and Y. The quality of the estimate was determined by calculating bias, which is the mean difference between the estimate of ρ_{XY} and the true value of ρ_{XY}. A high-quality estimate should have a value of bias at or close to zero.

Results

The results indicate that *lava's* estimates of ρ_{XY} were unbiased unless there was severe censoring on both x and y (see Table 1). Estimates of ρ_{XY} were unbiased when one variable has a low degree of censoring and the other variable has a high degree of censoring and when both variables have 50% censoring. When x and y both had 70% censoring, small degrees of bias occurred for large negative correlations. When x and y both had 95% censoring, bias was severe, particularly for small to moderate negative correlations.

Table 1
Estimates of Bias for Various Patterns of Censoring and ρ_{XY}

	Censoring Pattern			
ρ_{XY}	90% x 10% y	50% both	70% both	95% both
-.95	.00	.00	-.03	.06
-.80	.00	.00	-.03	-.09
-.40	.00	.00	.00	-.43
-.25	.00	.00	.00	-.47
0	.00	.00	.00	-.34
.10	.00	.00	-.01	-.26
.30	.00	.00	.00	-.11
.50	.00	.00	-.01	-.04
.85	.00	.00	.00	-.01
.90	.00	.00	.00	-.01

Discussion

The R package *lava* can be used to estimate the correlation between uncensored variables based on the data from censored variables. In this study, we examined the accuracy of *lava's* estimates. Researchers can use *lava* to make unbiased correlation estimates when both of their variables are moderately censored (50%), or when only one of their variables is highly censored (90%). However, caution should be exercised when using *lava* to deal with severely censored data. In particular, the bias came close to or exceeded .10 when the true population correlation ranged from highly negative (-.80) to moderately positive (.30). Therefore, researchers should attempt to minimize censoring if their expected correlations fall within that range.

Lava assumes a bivariate normal distribution. One limitation of our study is that we did not examine its performance when the assumption of normality has been violated. Therefore, future research could look at the accuracy of *lava's* estimates for non-normal data, such as data that are skewed.

References

Canales, R. A., Wilson, A. M., Pearce-Walker, J. I., Verhougstraete, M. P., & Reynolds, K. A. (2018). Methods for handling left-censored data in quantitative microbial risk assessment. Applied and environmental microbiology, 84(20), e01203-18. doi.org/10.1128/AEM.01203-18

Falcaro, M., Pendleton, N., & Pickles, A. (2013). Analysing censored longitudinal data with non-ignorable missing values: Depression in older age. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(2), 415–430. doi-org.ezproxy.library.unlv.edu/10.1111/j.1467-985X.2011.01034.x

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