Assessing the Quality of Correlation Estimates Based upon Censored Variables

Rosalba M. Gomez, Joshua Z. Chang, Zain N. Raja, Fitsum A. Ayele, and Kimberly A. Barchard
University of Nevada, Las Vegas

Introduction

Researchers are often faced with censored data. 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.

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. We chose -.95, -.80, -.40, -.25, 0, .10, .30, .50, .85, and .90 as our values of ρ_{XY}. 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. We generated 1000 trials for each cell. Within each trial, we generated a random set of data for which X and Y had a multivariate normal distribution.

Results

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 (Table 1). 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, for small to moderate negative correlations.

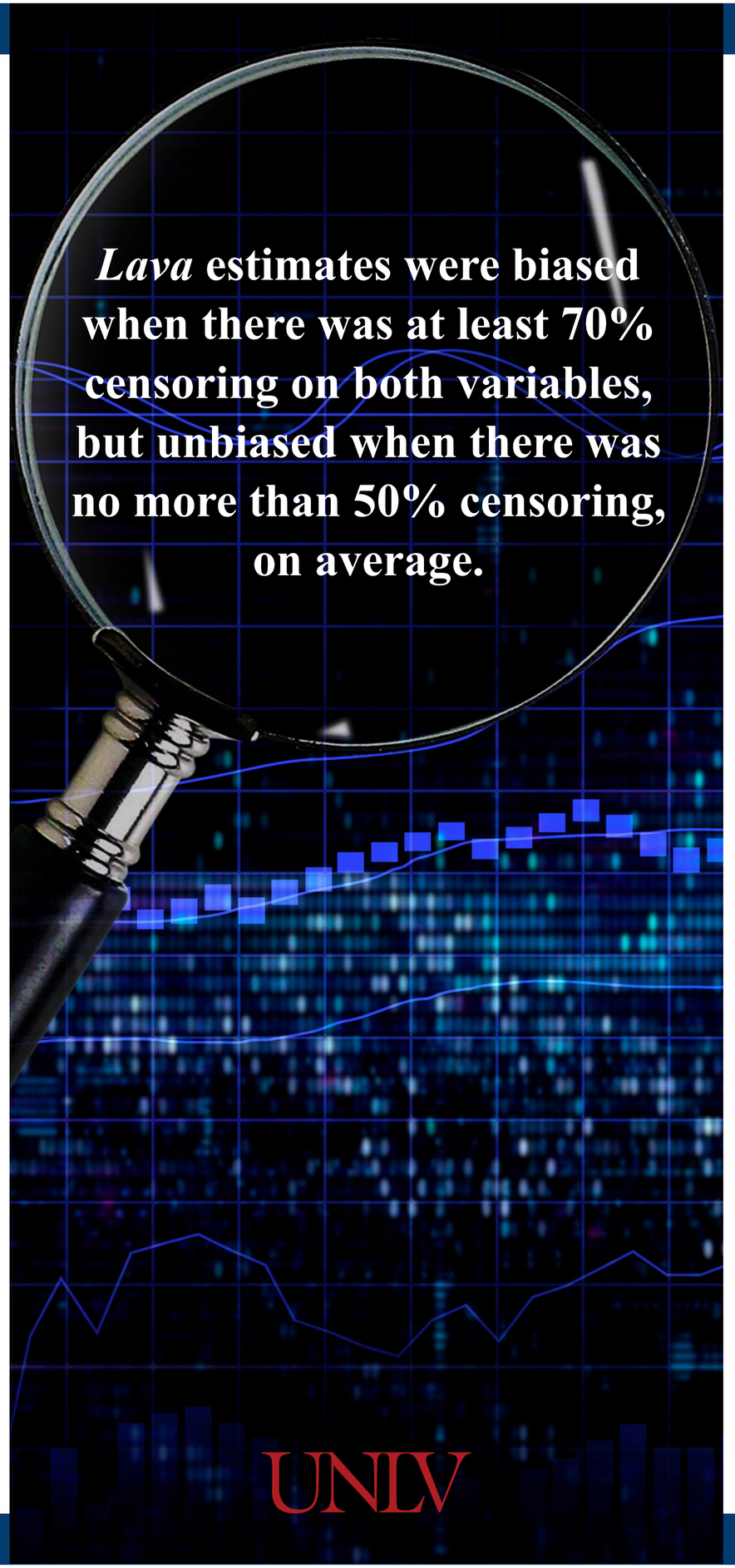

Table 1
Estimates of Bias for Various Patterns of Censoring and ρ_{XY}

ρ_{XY}	Censoring Pattern			
	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.

Contact Information
Rosalba Gomez gomezr24@unlv.nevada.edu
Joshua Chang changj20@unlv.nevada.edu
Zain Raja rajaz1@unlv.nevada.edu