They’re *Lava* Hot!

Two Methods that Reduce the Effect of Censoring on Correlations

Monica Cordova-Medina, Jerlyn Malasig, Victoria McDowell,
Fitsum A. Ayele, and Kimberly A. Barchard
Department of Psychology, University of Nevada, Las Vegas

UNLV

INTRODUCTION

DATA CENSORING

- Data point censoring occurs when the value of a variable is only partially known.
- Distorted results may occur if censored data are not addressed.
- Right and left censoring can occur in studies that use surveys to obtain data (Barchard & Russell, 2020).
- Right censoring occurs when the highest value on the rating scales does not distinguish people at the right end of the dimension (Barchard & Russell, 2020).
- Left censoring occurs when the lowest value does not distinguish people at the left end of the dimension (Barchard & Russell, 2020).

SURVEY STUDIES ON SOCIAL MEDIA

- **Right Censoring**
  - Chae (2018) studied the relationship between social media and happiness amongst Korean women between the ages of 20 and 39 years of age. To obtain social media frequency, participants were asked how often, on an average weekday, they use social media platforms. Participants replied using a 7-point scale (1 = never to 7 = more than 10 times a day). Right censoring occurred if participants who select 7 vary greatly in their social media use.
- **Left Censoring**
  - Utz and Beukeboom (2011) researched the relationship between the amount of time a person spent on social media and the happiness level they felt about their romantic relationship. To measure frequency on social network sites, participants in the study were asked questions that reported their time engaging in social media profile maintenance and grooming using a 7-point scale (1= almost never to 7 = daily)) left censoring occurred if participants who select 1 vary greatly in their social media use.

RESEARCH QUESTION

- The R package *lava* can estimate the correlation between uncensored variables, X and Y, based upon the data from censored variables, x and y, using either a correlation model or a regression model. Does the correlation model or regression model produce more accurate estimates?
- How does the correlation method and regression method perform under varying conditions (censoring patterns, ρ_{XY}, and sample sizes)?

METHOD

- We ran 112 cells of data; each cell represented a unique combination of censoring pattern, ρ_{XY}, and sample size.
- Patterns of censoring on x and y included: 10% censoring on both, 30% censoring on both, 50% censoring on both, 70% censoring on both, 10% censoring on x and 70% censoring on y, 30% censoring on x and 70% censoring on y, and 50% censoring on x and 70% censoring on y.
- ρ_{XY} values of -.95, .95, -.75, .75, -.5, .5, -.25, and .25.
- Sample sizes included 200 and 5000.
- We provided the x and y data to *lava* and asked it to estimate the correlation between X and Y (ρ_{XY})using both the correlation method and regression method.
- To assess *lava*’s performance:
  - We determined the proportion of trials where each model failed to converge on an estimate of ρ_{XY}.
  - We calculated bias – mean difference between actual values of ρ_{XY} and *lava* estimates of ρ_{XY}.
- We analyzed our data using between-within analysis of variance (ANOVA)

RESULTS

- Each model converged on an estimate 100% of the trials.
- See Table 1 and Table 2.
  - When n=200, lava estimates of ρ_{XY} were biased when censoring pattern was 50% on x and 70% on y or 70% on x and y, and the correlation was -.95 or -.75.
  - When n=5000, lava estimates of ρ_{XY} were biased when censoring pattern was 70% on x and y, and correlation was -.95.
- Figure 1 and Figure 2 show a significant interaction between censoring pattern, ρ_{XY} , and sample size, $F(42, 112) = 7.45$, $p < .05$.
- There was a second interaction between censoring pattern, ρ_{XY} , and model, $F(42, 112) = 15.25$, $p < .05$. Difference in bias between the correlation and regression models was trivial (<.01).

DISCUSSION

- Correlation model and regression model performed similarly.
- When sample size was 200, lava estimates of ρ_{XY} were unbiased when the censoring was 50% or below for both variables.
- When the sample size increased to 5000, lava estimates of ρ_{XY} were closer to the true correlation of uncensored variables and unbiased when the censoring pattern was below 70%.
- At high censoring patterns (70% on both x and y), estimates were moderately biased when the correlation was strong and negative.

IMPLICATIONS

- We are confident in *lava’s* ability to produce accurate estimates of ρ_{XY} when degrees of censoring are low to moderate on at least one of the variables.
- If researchers are interested in strong negative correlations, we encourage them to use a large sample and minimize censoring to avoid biased estimates.
- Since the two models had similar results, we recommend the regression model, which is guaranteed to provide a reasonable estimate.

LIMITATIONS AND FUTURE RECOMMENDATIONS

- We do not know if *lava* would produce unbiased estimates if sample size is smaller. We recommend running more trials to observe *lava’s* performance when a small sample size is used.
- Our data was normally distributed. Our results may not generalize to cases where data has different distributions. We recommend researchers assess *lava’s* performance when data has a nonnormal distribution(bimodal, trimodal, etc.).

Table 1

Correlation Model: Bias for lava Estimates of ρ_{XY} under Varying Parameters

	Censoring Pattern						
ρ_{XY}	10% x 10% y	30% x 30% y	10% x 70% y	30% x 70% y	50% x 50% y	50% x 70% y	70% x 70% y
			n = 200				
-.95	.00	.00	.00	.00	.00	-.02	-.03
-.75	.00	.00	.00	.00	.00	.00	-.03
-.50	.00	.00	.00	.00	.00	.00	-.01
-.25	.00	.00	.00	.00	.00	.00	.00
.25	.00	.00	.00	.00	-.01	.00	-.01
.50	.00	.00	.00	.00	.00	.00	.00
.75	.00	.00	.00	.00	.00	.00	.00
.95	.00	.00	.00	.00	.00	.00	.00
			n = 5000				
-.95	.00	.00	.00	.00	.00	.00	-.03
-.75	.00	.00	.00	.00	.00	.00	.00
-.50	.00	.00	.00	.00	.00	.00	.00
-.25	.00	.00	.00	.00	.00	.00	.00
.25	.00	.00	.00	.00	.00	.00	.00
.50	.00	.00	.00	.00	.00	.00	.00
.75	.00	.00	.00	.00	.00	.00	.00
.95	.00	.00	.00	.00	.00	.00	.00

Note. ρ_{XY} is the actual value of rho. Bias is calculated as the mean difference between actual values of rho and lava estimates of rho.

Table 2

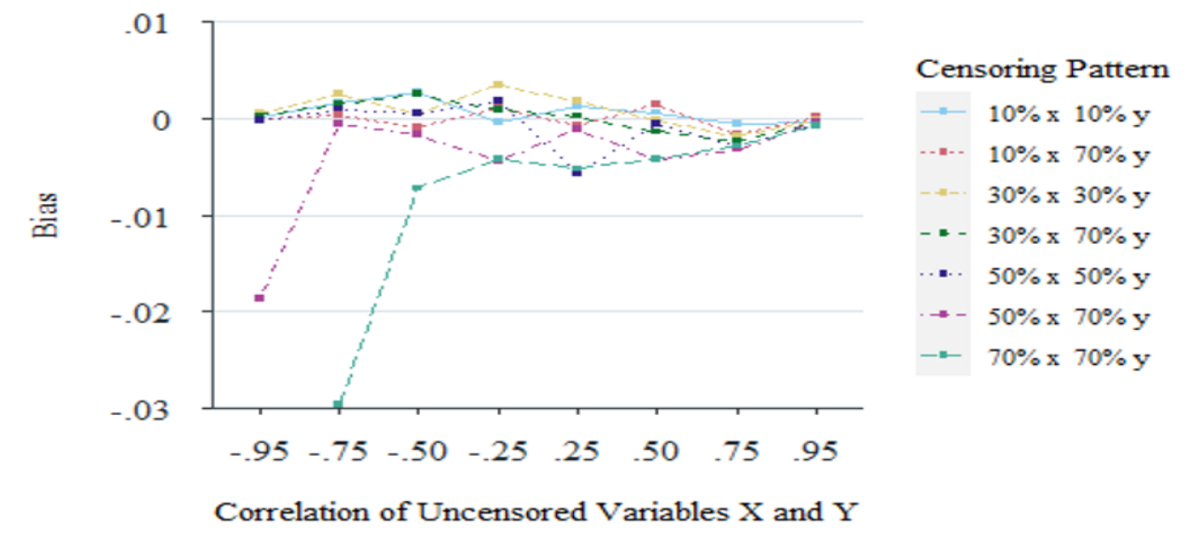
Regression Model: Bias for lava Estimates of ρ_{XY} under Varying Parameters

	Censoring Pattern						
ρ_{XY}	10% x 10% y	30% x 30% y	10% x 70% y	30% x 70% y	50% x 50% y	50% x 70% y	70% x 70% y
			n = 200				
-.95	.00	.00	.00	.00	.00	-.02	-.03
-.75	.00	.00	.00	.00	.00	.00	-.03
-.50	.00	.00	.00	.00	.00	.00	-.01
-.25	.00	.00	.00	.00	.00	.00	.00
.25	.00	.00	.00	.00	-.01	.00	-.01
.50	.00	.00	.00	.00	.00	.00	.00
.75	.00	.00	.00	.00	.00	.00	.00
.95	.00	.00	.00	.00	.00	.00	.00
			n = 5000				
-.95	.00	.00	.00	.00	.00	.00	-.03
-.75	.00	.00	.00	.00	.00	.00	.00
-.50	.00	.00	.00	.00	.00	.00	.00
-.25	.00	.00	.00	.00	.00	.00	.00
.25	.00	.00	.00	.00	.00	.00	.00
.50	.00	.00	.00	.00	.00	.00	.00
.75	.00	.00	.00	.00	.00	.00	.00
.95	.00	.00	.00	.00	.00	.00	.00

Note. ρ_{XY} is the actual value of rho. Bias is calculated as the mean difference between actual values of rho and lava estimates of rho.

Figure 1

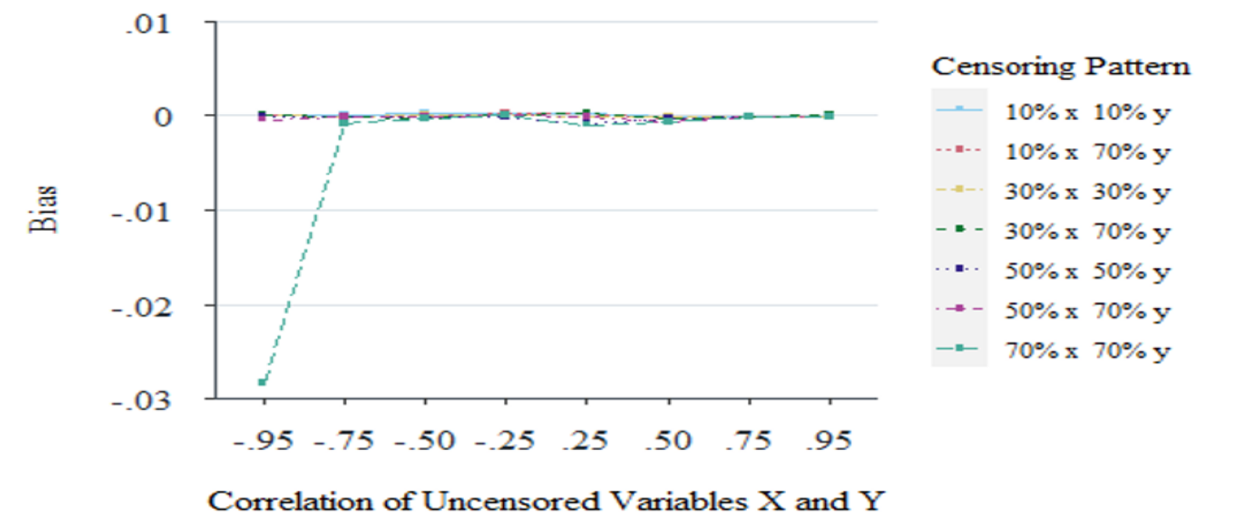
Interaction Between Correlation of Uncensored Variables X and Y, Censoring Pattern, and Sample Size (n = 200)



Note. Bias was calculated as the mean difference between actual values of rho and lava estimates of rho. Results have been averaged between the correlation and regression models.

Figure 2

Interaction Between Correlation of Uncensored Variables X and Y, Censoring Pattern, and Sample Size (n = 5000)



Note. Bias was calculated as the mean difference between actual values of rho and lava estimates of rho. Results have been averaged between the correlation and regression models.