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

Reference: Cordova-Medina, M. Malasig, J., McDowell, V., Ayele, F.A., Barchard, K.A. (2021, May). *They’re Lava Hot! Two Methods that Reduce the Effect of Censoring on Correlations.* [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

Data point censoring occurs when the value of a variable is only partially known. For example, censoring can occur in studies using rating scales. Right censoring occurs when the highest value on the rating scales does not distinguish people at the right end of the dimension; left censoring occurs when the lowest value does not distinguish people at the left end. Censoring leads to under-estimates of correlations. The R package *lava* estimates the correlation between uncensored variables when given data from censored variables. Lava has two models: The correlation model is easier to use, but the regression model ensures the estimate is within the allowable range for a correlation [-1, 1]. We conducted a Monte Carlo study to evaluate these methods in terms of bias: the mean difference between the true correlation and the lava estimate. We examined a wide range of censoring patterns (e.g., 10% censoring on x and 70% on y; 30% censoring on both), a variety of correlations (±.95, ±.75, ±.50, ±.25), and two sample sizes (200, 5000). With low to moderate censoring, *lava* estimates were unbiased for all combinations of correlation and sample size. However, with high censoring and strong negative correlations, *lava* estimates were still biased. Increasing sample size lowered the bias. Since the two models had similar results, we recommend the regression model, which is guaranteed to provide a reasonable estimate. If researchers are interested in strong negative correlations, we encourage them to use a large sample and minimize censoring to avoid biased estimates.

Keywords: data censoring, R package *lava*, correlation model, regression model, technology

Introduction

Censored data refer to data in which the value of an observation or measurement is only partially known (Gijbels, 2010). Because censored data biases results, many researchers tend to either delete the censored data or substitute the data with determined values such as the mean of observed data. Both methods are also likely to bias the results because the missing or substituted values might lead to biased interpretation (Fox, 2016).

Right censored data can occur in surveys if the rating scales do not distinguish people at the right end of the dimension (Barchard & Russell, 2020). The scales fail to account for variability among the high scorers. In longitudinal and survival studies, right censoring can occur when the event of interest is not observed (e.g., drop out, lost to follow up, end of study, etc.). In both studies, the participant’s scores would be shown to be artificially lower than reality and not accurately reflect their true scores. It is also possible that the event of interest occurred within the time limit of the study, but researchers were unable to document the data.

Left censored data can occur in surveys if the rating scales do not distinguish people at the left end of the dimension (Barchard and Russel, 2020). The scale is unable to account for variability among the low scorers just as right censoring does not distinguish people at the right end of the dimension. Left censoring could also occur because of a measuring device’s Limit of Detection (LOD). The LOD is the point at which the measurement device is not sensitive enough to detect data points (Gijbels, 2010). For example, if a bathroom scale is used to weigh a piece of paper, the scale will likely measure the weight at zero. The measurement device (e.g. scale) is not sensitive enough to measure the paper's weight due to its LOD.

To analyze cases of left censoring and right censoring, we researched studies that examine the relationship between screen time and happiness levels. Censored data can occur in these types of studies, especially with the use of surveys to obtain data. 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.

Additionally, right censoring occurs in studies if participants left a study before it ended and the event of interest was not observed. Perry and Schleifer (2018) conducted a six-year study to examine the relationship between pornography use and the probability of divorce. Data were collected from three different groups who participated in one of the following time periods: 2006-2010, 2008-2012, or 2010-2014. About 36% of the participants in each group left before the study ended. Among the participants who left early and did not get a divorce, right censoring occurred.

To examine left censoring, we found a study by Utz and Beukeboom (2011) that 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, the 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). Self-reported rating scales will lead to left censoring if the lower end of the dimension includes people who vary substantially in their technology use. One of the questions asks, "How often do you visit the profiles of close friends?" Because the researcher did not define "almost never" for participants, their answers on this data point could differ immensely. Thus, left censoring occurred if participants who select 1 vary greatly in their social media use.

If censoring is not properly dealt with, then a distortion in results and biased interpretations are likely to occur (Pesonen, et. al 2015). Although many censored data analysis tools exist to correct this distortion, these tools are normally specialized to deal with censoring on only one variable and deal with univariate statistics such as central tendencies and variabilities (Helsel, 2012). Few methods exist to deal with censoring on more than one variable and in estimating other statistics, such as a correlation. Correlations describe the relationship of two variables and are widely used in psychometrics as a basis in many statistical analyses. Fortunately, the R package *lava* can estimate correlations when two variables are censored (Holst et al., 2015).

In this study, we assessed how well *lava* estimated the correlation between uncensored variables, X and Y, based upon data from censored variables, x and y. The correlation between uncensored variables, X and Y, will be referred to by ρ_{XY}. We specified the use of two models provided by *lava*: correlation and regression. The correlation model will allow for a direct calculation of an estimate of ρ_{XY} based on censored variables, x and y, whereas the regression model will both give an estimate and ensure it is within the permissible range of -1 and 1. We assessed how the two models fared in creating point estimates of ρ_{XY} by looking at the bias generated. Bias is measured as the mean difference between ρ_{XY} and *lava*'s estimates of ρ_{XY}. We also incorporated a variety of parameters (ρ_{XY}, censoring patterns, and sample sizes) to see if bias varied. Running this study ensured if *lava* is suitable for estimating correlations when censored data occur on more than one variable.

Method

In order to compare the performance of *lava's* correlation model and regression model, we varied the patterns of censoring on x and y, ρ_{XY}, and sample size. In total, we ran 112 cells of data; each cell represented a unique combination of censoring pattern, $\rho_{XY,}$ and sample size.

We included a wide variety of censoring patterns, wide range of correlations, and two sample sizes to evaluate which conditions produced more accurate estimations and which conditions produced less accurate estimations. Patterns of censoring on x and y included 10% censoring on both variables, 30% censoring on both variables, 50% censoring on both variables, and 70% censoring on both variables; this represents low, moderate, and heavy patterns of censoring. We also included mixed censoring patterns because in research, degrees of censoring tend to be different on each variable. Thus, we included 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 were used to determine how the strength and direction of correlation could influence *lava*'s performance. Sample sizes included 200 and 5000. 200 was used to represent a moderate sample size. 5000 was used to represent a large sample size; we also wanted to replicate the sample size that was used in the technology studies we researched.

For each cell, we ran 1000 trials in which we generated a random set of data for which X and Y had a bivariate normal distribution. We then censored x and y to the required degrees. We provided the x and y data to *lava* and asked it to estimate the correlation between X and Y, using both a correlation model and a regression model. First, we determined the proportion of trials where each model failed to converge on an estimate of $\rho_{XY.}$ Next, we calculated bias to assess the quality of the point estimates of $\rho_{XY.}$ Bias was calculated as the mean difference between actual values of ρ_{XY} and *lava* estimates of $\rho_{XY.}$ In order to analyze the results across our 112 cells, we used between-within analysis of variance (ANOVA) and graphed an interaction that was significant. The between-cell factors were $\rho_{XY,}$ censoring pattern, and sample size. The within-cell factor was the estimation method used (correlation or regression). The dependent variable was bias.

Results

The results from our Monte Carlo study have been formatted into two tables. See Table 1 and Table 2. The tables show the bias for *lava* estimates of ρ_{XY} for each estimation method (correlation and regression) for each unique combination of censoring pattern, $\rho_{XY,}$ and sample size. The *lava* estimates of ρ_{XY} given by the two models (correlation and regression) were similar for each cell. We also found that each model converged in 100% of the trials.

After running the between-within analysis of variance (ANOVA) for all 112 cells, we found two significant interaction effects. The first interaction was between $\rho_{XY,}$ censoring pattern, and sample size, $F(42, 112) = 7.45$, $p < .05$. This is illustrated in Figure 1 (n = 200) and Figure 2 (n = 5000). *Lava* estimates of ρ_{XY} were unbiased when the sample size was 200 and censoring consisted of one of the following patterns: 10% on x and y, 30% on x and y, 10% on x and 70% on y, 30% on x and 70% on y, or 50% on x and y. The *lava* estimates of ρ_{XY} for a sample size of 200 were moderately biased when the censoring pattern was 50% on x and 70% on y or 70% on x and y, and the correlation was -.95 or -.75. Our figures show that increasing sample size from 200 to 5000 reduced that bias. When the sample size was 5000, all estimated values of ρ_{XY} were unbiased except when the censoring pattern was 70% on x and y, and correlation was -.95. In this instance, there was moderate bias. The second interaction was between $\rho_{XY,}$ censoring pattern, and model, $F(42, 112) = 15.25$, $p < .05$. Although there was a significant interaction, the difference in bias between the correlation and regression models was trivial (<.01).

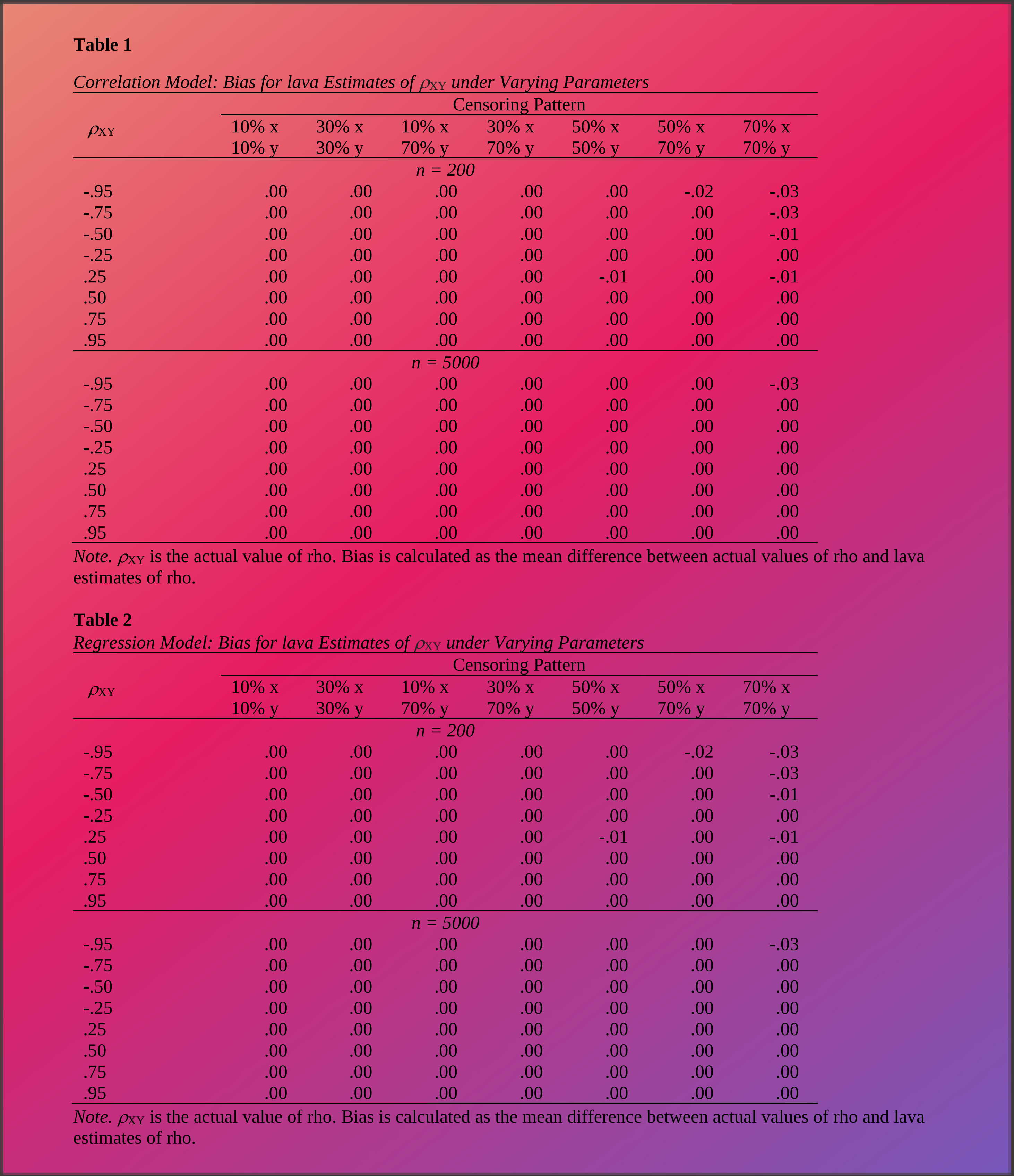


Table 1

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

ρ_{XY}	Censoring Pattern						
	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

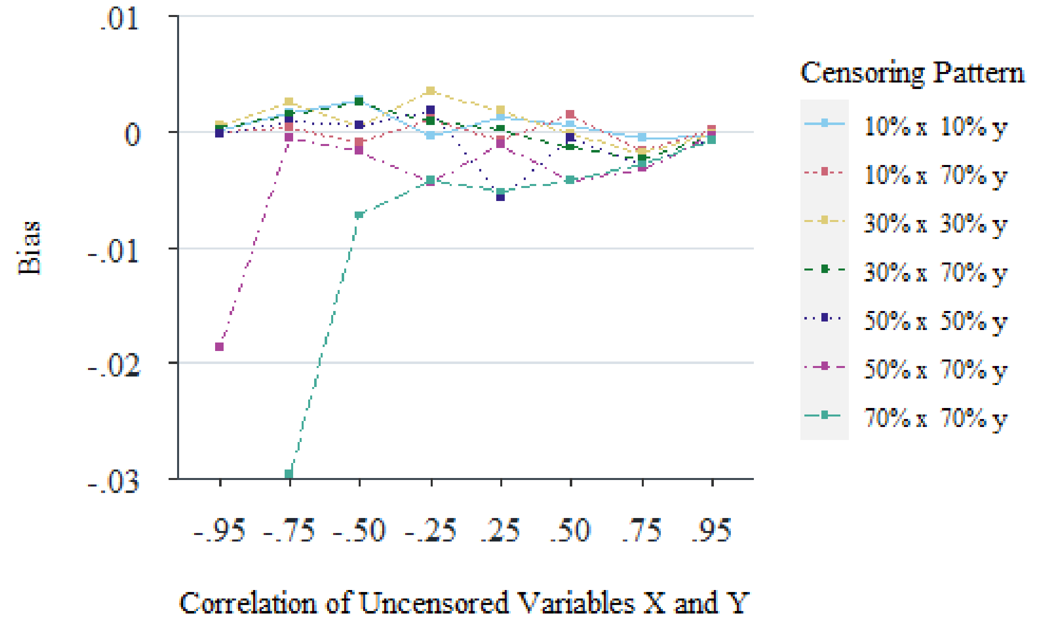
Regression Model: Bias for lava Estimates of ρ_{XY} under Varying Parameters

ρ_{XY}	Censoring Pattern						
	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

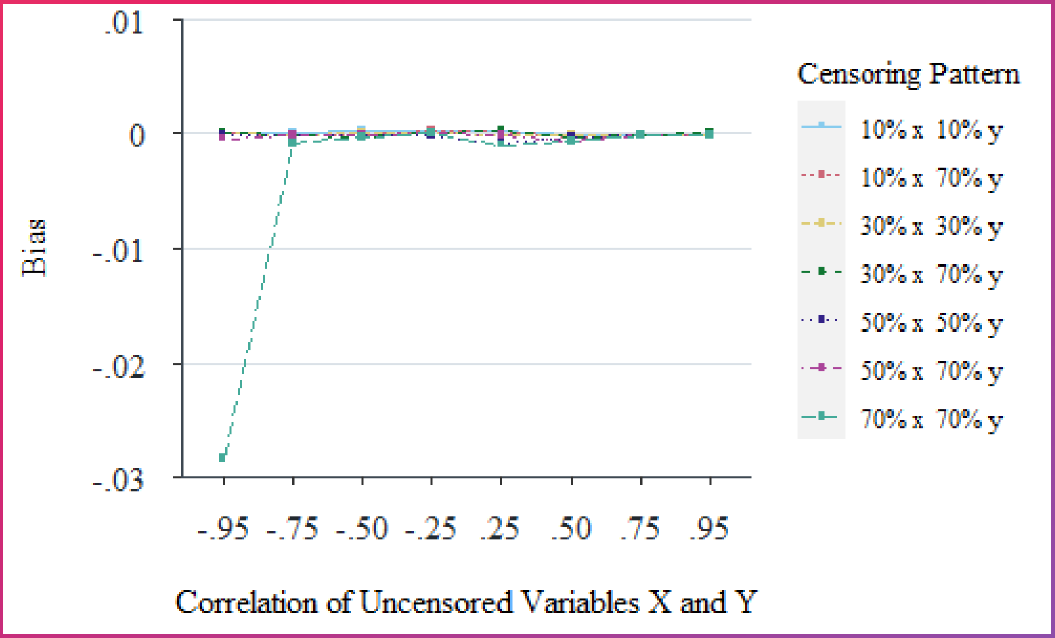
Interaction Between Correlation of Uncensored Variables X and Y, Censoring Pattern, and Sample Size (n = 200)



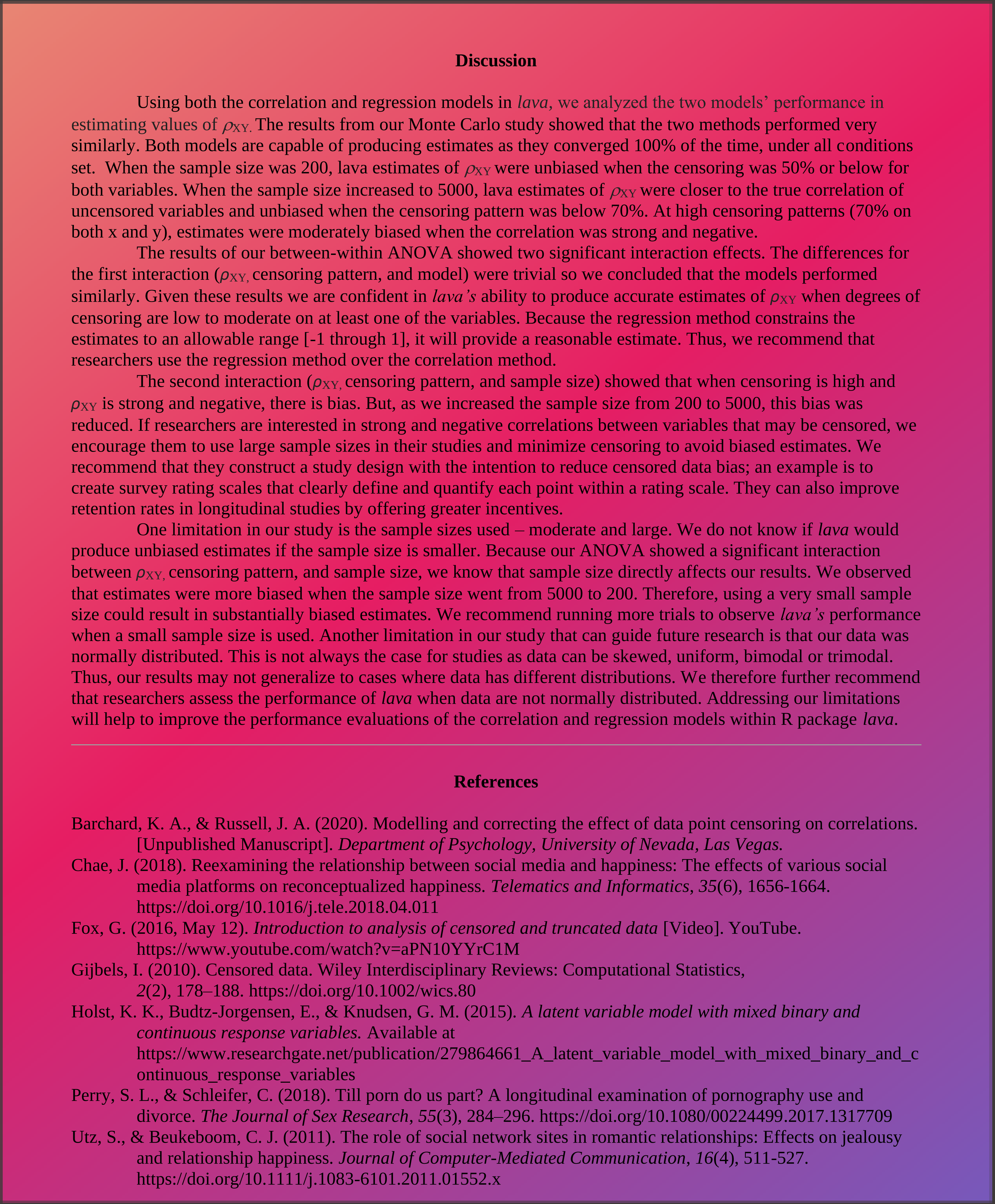
Note. Bias was calculated as the mean difference between actual values of rho and lava estimates of rho. Results have been averaged between the correlation and regression models.

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.



Discussion

Using both the correlation and regression models in *lava*, we analyzed the two models' performance in estimating values of ρ_{XY}. The results from our Monte Carlo study showed that the two methods performed very similarly. Both models are capable of producing estimates as they converged 100% of the time, under all conditions set. When the 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.

The results of our between-within ANOVA showed two significant interaction effects. The differences for the first interaction (ρ_{XY}, censoring pattern, and model) were trivial so we concluded that the models performed similarly. Given these results 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. Because the regression method constrains the estimates to an allowable range [-1 through 1], it will provide a reasonable estimate. Thus, we recommend that researchers use the regression method over the correlation method.

The second interaction (ρ_{XY}, censoring pattern, and sample size) showed that when censoring is high and ρ_{XY} is strong and negative, there is bias. But, as we increased the sample size from 200 to 5000, this bias was reduced. If researchers are interested in strong and negative correlations between variables that may be censored, we encourage them to use large sample sizes in their studies and minimize censoring to avoid biased estimates. We recommend that they construct a study design with the intention to reduce censored data bias; an example is to create survey rating scales that clearly define and quantify each point within a rating scale. They can also improve retention rates in longitudinal studies by offering greater incentives.

One limitation in our study is the sample sizes used – moderate and large. We do not know if *lava* would produce unbiased estimates if the sample size is smaller. Because our ANOVA showed a significant interaction between ρ_{XY}, censoring pattern, and sample size, we know that sample size directly affects our results. We observed that estimates were more biased when the sample size went from 5000 to 200. Therefore, using a very small sample size could result in substantially biased estimates. We recommend running more trials to observe *lava's* performance when a small sample size is used. Another limitation in our study that can guide future research is that our data was normally distributed. This is not always the case for studies as data can be skewed, uniform, bimodal or trimodal. Thus, our results may not generalize to cases where data has different distributions. We therefore further recommend that researchers assess the performance of *lava* when data are not normally distributed. Addressing our limitations will help to improve the performance evaluations of the correlation and regression models within R package *lava*.

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.*

Chae, J. (2018). Reexamining the relationship between social media and happiness: The effects of various social media platforms on reconceptualized happiness. *Telematics and Informatics*, *35*(6), 1656-1664. https://doi.org/10.1016/j.tele.2018.04.011

Fox, G. (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., Budtz-Jorgensen, 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

Perry, S. L., & Schleifer, C. (2018). Till porn do us part? A longitudinal examination of pornography use and divorce. *The Journal of Sex Research*, *55*(3), 284–296. https://doi.org/10.1080/00224499.2017.1317709

Utz, S., & Beukeboom, C. J. (2011). The role of social network sites in romantic relationships: Effects on jealousy and relationship happiness. *Journal of Computer-Mediated Communication*, *16*(4), 511-527. https://doi.org/10.1111/j.1083-6101.2011.01552.x