Correlating Positively and Negatively Keyed Items: The Problem and The Solution

Kimberly A. Barchard, University of Nevada, Las Vegas
James A. Russell, Boston College

ABSTRACT

Correlations of positively and negatively keyed items are biased if they target high or low functioning, so that normal range test takers get the lowest or highest possible scores. R package *lava* corrects for such censoring. Estimates are accurate and precise unless variables have 70% censoring and correlate -.9; then estimates are biased -.06. Regression model confidence intervals were superior.

INTRODUCTION

Censoring occurs when items capture variation among people with moderate and high levels on a trait, but give identical scores to everyone with low levels. For example, items about suicidal thoughts distinguish those who sometimes think of suicide from those who are actively planning it, but do not distinguish sadness levels among those who are not suicidal.

When positively and negatively keyed items are intended to capture high and low functioning, respectively, they may all suffer from censoring. Correlations between such censored items are biased. See Figure 1. If two items are intended to be opposites, but both are censored, they will not correlate -1. For example, if items measure opposite halves of a normally distributed trait, they correlate -.467 (Russell & Carroll, 1999).

R package *lava* (Holst, 2020; Holst & Budtz-Jørgensen, 2013) corrects for the effect of censoring on correlations. The purpose of this poster was to evaluate the accuracy of point and interval estimates of the correlation between uncensored variables X and Y, based upon the data from censored variables x and y.

METHOD

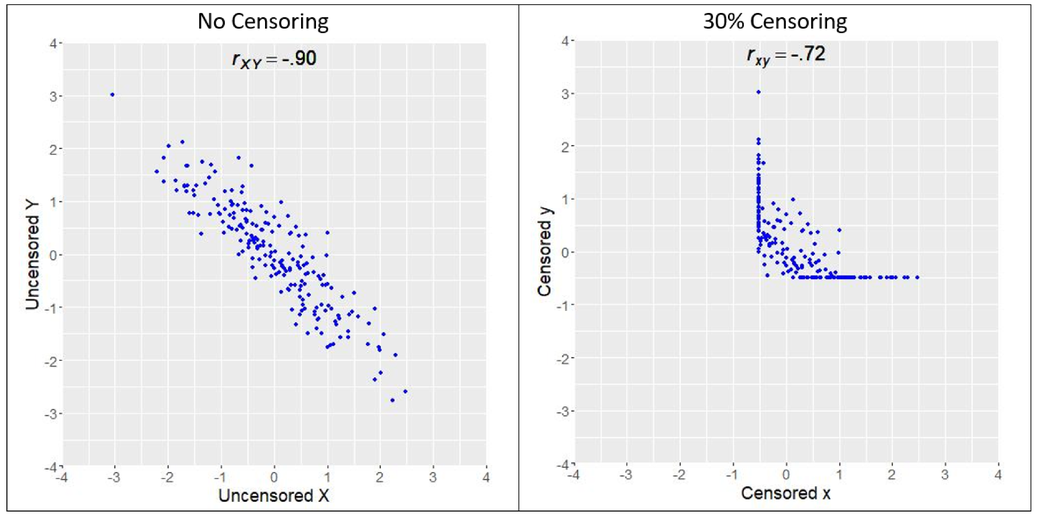
We simulated samples of 200 cases with bivariate normal distributions on X and Y, which correlated -.3, -.7, -.9, or -1. Then we imposed low, moderate, or high levels of censoring on both variables (30%, 50%, or 70%) to create x and y. For example, for 30% censoring, we set all values below the 30th percentile to be equal to the 30th percentile. We provided *lava* with data from censored x and y and asked it to estimate the correlation among uncensored X and Y.

We used two point estimates: one based upon the correlation model and one based upon the regression model. The regression model estimates were constrained to fall within the allowable values for a correlation [-1, 1].

We used three interval estimates: the correlation model Ward confidence interval, the regression model Ward confidence interval, and the regression model profile confidence interval.

Figure 1

Effect of Censoring on a Correlation



Note. 200 cases were generated from a population with a bivariate standard normal distribution (left panel). Then 30% censoring was applied to the x and y variables (right panel).

Use R package *lava* to estimate correlations between positively and negatively keyed items measuring high and low levels of a trait.

RESULTS

Point Estimates

The point estimates performed well (see Table 1). For both the correlation model and the regression model, point estimates always converged upon a solution, were unbiased, and had small root mean square errors, unless both variables had 70% censoring and the original correlation was -.9. In that case, *lava* overestimated the magnitude of the correlation (bias = -.06). Even for this extreme data, these point estimates were far more accurate than the uncorrected correlations (r_{xy} = -.22, bias = .68).

Interval Estimates

The confidence intervals performed well (see Table 2), converging upon a solution, capturing ρ_{XY} at least 95% of the time, and having narrow widths, except under two circumstances.

When ρ_{XY} = -1, the confidence intervals sometimes failed to converge, had widths of 2 (thus providing no information about ρ_{XY}), or were substantially biased. However, perfect correlations are rare in real datasets. Thus, perhaps these poor results can be safely ignored.

When there was 70% censoring on both variables and ρ_{XY} was -.9, the regression profile confidence intervals captured ρ_{XY} the most often (89% of the time), but were five times as wide as the Ward confidence intervals (.27 compared to .04). The correlation Ward confidence intervals were sometimes excessively wide (e.g., 1.06). Therefore, the regression Ward confidence intervals were best for this kind of data.

Nonetheless, the regression Ward confidence intervals for ρ_{XY} = -.90 and -1 were slightly biased and captured ρ_{XY} rarely, making it hard to test if two variables (such as positively and negatively keyed items intended to measure the same construct) are exact opposites. Therefore, researchers who are interested in opposites should aim to minimize censoring where possible.

DISCUSSION

To estimate the relationships between positively and negatively keyed items, researchers should use the R package *lava*, particularly if items were designed to measure high or low levels of a trait or to identify high or low functioning individuals. *Lava* provides useful point and interval estimates of the correlation between uncensored variables based upon the data from censored variables. When researchers are interested in positively and negatively keyed items with strong relationships, they should attempt to reduce censoring by using a wider range of response options or less extreme samples.

REFERENCES

Holst, K. K., & Budtz-Jørgensen, E. (2013). Linear latent variable models: The lava-package. *Computational Statistics*, *28*(4), 1385-1452. doi:1.1007/s00180-012-0344-y

Holst, K. K. (2020). *lava: Latent Variable Models (Version 1.6.8)* [Computer software]. https://CRAN.R-project.org/package=lava

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*(1), 3 – 30. https://doi.org/10.1037/0033-2909.125.1.3

Table 1

Point Estimates of ρ_{XY}

Sample Characteristics				Correlation Model Estimates of ρ_{XY}				Regression Model Estimates of ρ_{XY}			
Censoring on X	Censoring on Y	ρ_{XY}	Average of r_{xy} across trials	Proportion that converged	Average across trials	Bias	RMSE	Proportion that converged	Average across trials	Bias	RMSE
.3	.3	-.30	-.25	1.00	-.30	.00	.07	1.00	-.30	.00	.07
.3	.3	-.70	-.56	1.00	-.70	.00	.04	1.00	-.70	.00	.04
.3	.3	-.90	-.70	1.00	-.90	.00	.02	1.00	-.90	.00	.02
.3	.3	-1.00	-.77	1.00	-1.00	.00	.00	1.00	-1.00	.00	.00
.5	.5	-.30	-.20	1.00	-.30	.00	.08	1.00	-.30	.00	.08
.5	.5	-.70	-.40	1.00	-.70	.00	.05	1.00	-.70	.00	.05
.5	.5	-.90	-.45	1.00	-.90	.00	.02	1.00	-.90	.00	.02
.5	.5	-1.00	-.47	1.00	-1.00	.00	.00	1.00	-1.00	.00	.00
.7	.7	-.30	-.13	1.00	-.31	-.01	.11	1.00	-.31	-.01	.11
.7	.7	-.70	-.21	1.00	-.71	-.01	.10	1.00	-.71	-.01	.10
.7	.7	-.90	-.22	1.00	-.96	-.06	.08	1.00	-.96	-.06	.08
.7	.7	-1.00	-.22	1.00	-.98	.02	.02	1.00	-.98	.02	.02

Note. RMSE = Root mean square error. ρ_{XY} is the population correlation between the uncensored variables X and Y. r_{xy} is the sample correlation between the censored variables x and y. Green font indicates desirable result; red font, undesirable.

Table 2

Interval Estimates of ρ_{XY}

Sample Characteristics				Correlation Model Ward Confidence Intervals					Regression Model Ward Confidence Intervals					Regression Model Profile Confidence Intervals				
Censoring on X	Censoring on Y	ρ_{XY}	Average of r_{xy} across trials	Proportion that converged	Average midpoint across trials	Proportion that include ρ_{XY}	Average width across trials	Maximum width across trials	Proportion that converged	Average midpoint across trials	Proportion that include ρ_{XY}	Average width across trials	Maximum width across trials	Proportion that converged	Average midpoint across trials	Proportion that include ρ_{XY}	Average width across trials	Maximum width across trials
.3	.3	-.30	-.25	1.00	-.30	.94	.27	.36	1.00	-.29	.95	.27	.36	1.00	-.29	.95	.27	.30
.3	.3	-.70	-.56	1.00	-.70	.96	.16	.25	1.00	-.69	.96	.16	.25	1.00	-.69	.97	.16	.22
.3	.3	-.90	-.70	1.00	-.90	.98	.06	.10	1.00	-.89	.98	.06	.10	1.00	-.89	.98	.06	.10
.3	.3	-1.00	-.77	1.00	-1.00	1.00	.00	.00	1.00	-.70	1.00	.60	2.00	.99	-1.00	1.00	.00	.00
.5	.5	-.30	-.20	1.00	-.30	.94	.31	.38	1.00	-.29	.94	.31	.38	1.00	-.29	.94	.31	.35
.5	.5	-.70	-.40	1.00	-.70	.94	.19	.28	1.00	-.69	.95	.19	.28	1.00	-.69	.96	.20	.28
.5	.5	-.90	-.45	1.00	-.90	.95	.09	.14	1.00	-.89	.96	.09	.15	1.00	-.89	.98	.09	.15
.5	.5	-1.00	-.47	.93	-1.00	1.00	.00	.53	.01	-1.00	1.00	.00	.00	1.00	-1.00	1.00	.00	.00
.7	.7	-.30	-.13	1.00	-.31	.93	.40	.56	1.00	-.29	.94	.39	.55	1.00	-.29	.95	.40	.44
.7	.7	-.70	-.21	1.00	-.71	.90	.28	.47	1.00	-.68	.91	.28	.46	.95	-.67	.98	.31	.38
.7	.7	-.90	-.22	1.00	-.96	.15	.06	1.06	1.00	-.96	.10	.04	.37	.12	-.78	.89	.27	.30
.7	.7	-1.00	-.22	1.00	-.98	.05	.03	.69	1.00	-.98	.00	.02	.25	.00	NA	NA	NA	NA

Note. ρ_{XY} is the population correlation between the uncensored variables X and Y. r_{xy} is the sample correlation between the censored variables x and y. Green font indicates desirable result; red font, undesirable.