Abstract

It is commonly advised to center predictors in multiple regression, especially in the presence of interactions (J. Cohen, Cohen, West, & Aiken, 2013). This will enhance the interpretation of regression parameters, and (arguably; Dalal & Zickar, 2012; Echambadi & Hess, 2007; Kromrey & Foster-Johnson, 1998) reduce multicollinearity. However, I demonstrate that with missing data, centering predictors may bias parameter estimates. I develop a Pearson-Lawyley-based (Aitken, 1935; Lawley, 1944) correction (called r_{pl}) that is insensitive to centering, then evaluate its performance via Monte Carlo Simulation.

Keywords: missing data, selection, range restriction, interactions

Word count: 1994

When Interactions Bias Corrections: A Missing Data Correction for Centered Predictors

Centering predictors often improves the interpretation of coefficients (J. Cohen et al., 2013). Some have suggested it increases precision by reducing multicollinearity (J. Cohen et al., 2013). Although this last advantage is controversial (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Kromrey & Foster-Johnson, 1998), none (that I know of) recommended against centering (except when the predictors have meaningful zero; J. Cohen et al., 2013). However, in one situation centering predictors will exacerbate bias.

In this paper, I show how centering can lead to substantial bias when data are missing. First, I review the advantages of centering. Next, I show how centering predictors may change a "Missing At Random" into a "Missing *Not* at Random" situation. Finally, I introduce a correction that allows researchers to center predictors without bias, and assess its performance via Monte Carlo Simulation.

Regression and Centering

Suppose a university wishes to assess the impact of socioeconomic status (SES) on first year GPA ($FYGPA$), and wishes to correct for missing data (let us assume the university selected based on SAT scores). Further suppose these two predictors interact (e.g., for those with high SAT scores, SES is more predictive of $FYGPA$). Mathematically,

$$FYGPA = b_0 + b_1 SES + b_2 SAT + b_3 SES \cdot SAT$$

The researcher might be inclined to center both SES and SAT scores. Doing so supposedly has two advantages. First, the coefficients for the centered variables are more interpretable (J. Cohen et al., 2013). Recall that with interactions, the relationship between SES and $FYGPA$ is non-linear; the slope between SES and $FYGPA$ changes depending on the value of SAT. When centered, the b_1 parameter is the *average* slope of $FYGPA$ on SES across all values of SAT.

The second purported advantage of centering is that is removes "nonessential" collinearity (Aiken & West, 1991; J. Cohen et al., 2013). The covariance between the interaction variable ($SES \cdot SAT$) and either predictor is a function of the means of SAT and SES (Aiken & West, 1991, p. 180, Equation A.13):

$$cov(SES, SES \cdot SAT) = s^2_{SES}\overline{SAT} + cov(SAT, SES)\overline{SES} \tag{1}$$

(The above equation only applies when each predictor is symmetrical). When both predictors are centered, the means are zero and $cov(SES, SES \cdot SAT)$ vanishes. This is what is called "nonessential multicollinearity," or collinearity attributable to the means of the predictors.

Some (e.g., J. Cohen et al., 2013) argue removing essential multicollinearity increases the precision of parameter estimates since multicollinearity tends to inflate standard errors. However, others (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Kromrey & Foster-Johnson, 1998) demonstrate precision is unaffected by centering.

Regardless of whether centering affect precision, it is considered wise practice, at least for its interpretative advantages. However, under missing data, centering may inflate bias.

Interactions and Missing Data

In concurrent validity designs, when interactions exist, data are "Missing Not at Random" (MNAR; Little & Rubin, 2014; Rubin, 1976), which means the probability of missingness is correlated with both observable and unobservable data. To understand why, consider our previous example. Suppose students were selected based on SAT scores and the researcher wishes to assess the correlation between socioeconomic status (SES) and $FYGPA$. However, they want to know the unattenuated correlation, but unfortunately only have incumbent data for SES. Assuming SES itself is not a cause of attrition (or selection), missingness was actually cased by two variables:

(1) SAT scores. Since these were recorded before selection, these data are MAR (meaning missingness is caused by observable data).

(2) $SAT \cdot SES$ scores. This product term is correlated with the probability of missingness such that it is independent of SAT and SES alone (since an interaction is present). Some of these product scores are missing (because they were not selected into the university), rending them unobservable. Consequently, these data are MNAR.

Notice that the data are MNAR, regardless of whether SES itself is a cause of missingness (again, because the product variable is missing for certain applicants). Had SES been measured before selection on SAT occurred (i.e., in a predictive validity design), the data would be MAR.

In most situations, the fact that the data are MNAR is not problematic. One need not actually model the cause of missingness to render a situation MAR. Rather, one simply needs a correlate of the cause of missingness (Collins, Schafer, & Kam, 2001). With uncentered variables, the correlation between the predictors and their product is high and can be control for by using applicant SAT scores. When we center the variables, however, that correlation vanishes.

Correction Procedure

Recall the Pearson-Lawley equation (Aitken, 1935; Lawley, 1944) corrects estimates when missingness occurs on one or more variables. As a multivariate extension of Case III, it requires two inputs:

(1) The unrestricted covariance matrix of the variables responsible for missingness (in this case, SES and $SES \cdot SAT$):

$$\Sigma = \begin{bmatrix} \sigma^2_{SAT} & \sigma_{SAT,SES\cdot SAT} \\ \sigma_{SES\cdot SAT,SAT} & \sigma^2_{SES\cdot SAT} \end{bmatrix}$$

(2) The restricted covariance matrix of all the variables (in this case, SES, SAT, $SES \cdot SAT$, and $FYGPA$):

$$\widetilde{\Sigma} = \begin{bmatrix} \tilde{\sigma}^2_{SAT} & \tilde{\sigma}_{SAT,SES} & \tilde{\sigma}_{SAT,SES\cdot SAT} & \tilde{\sigma}_{SAT,FYGPA} \\ \tilde{\sigma}_{SES,SAT} & \tilde{\sigma}^2_{SES} & \tilde{\sigma}_{SES,SES\cdot SAT} & \tilde{\sigma}_{SES,FYGPA} \\ \tilde{\sigma}_{SES\cdot SAT,SAT} & \tilde{\sigma}_{SES\cdot SAT,SES} & \tilde{\sigma}^2_{SES\cdot SAT} & \tilde{\sigma}_{SES\cdot SAT,FYGPA} \\ \tilde{\sigma}_{SAT,FYGPA} & \tilde{\sigma}_{SES,FYGPA} & \tilde{\sigma}_{SES\cdot SAT,FYGPA} & \tilde{\sigma}^2_{FYGPA} \end{bmatrix}$$

(Note: anything with a tilde represents the restricted estimate).

To compute $\sigma_{SAT,SAT\cdot SES}$, we can use Equation 1.This requires knowing the unrestricted variance of SES, which may be unavailable. However, this parameter can be acquired using the PL correction, using incumbent data for SAT. One simply inputs the 1×1 matrix of SAT^2 (i.e., the variance) as the unrestricted covariance matrix, then inputs the restricted estimates for the covariance matrix of SES/SAT.

After performing the PL correction, we now have most[1] of the inputs necessary for Equation 1. We can also compute the population covariance between SES and $SES \cdot SAT$ (Aiken & West, 1991, p. 180, Equation A.13):

$$cov(SES, SES \cdot SAT) = s^2_{SAT}\overline{SES} + cov(SAT, SES)\overline{SAT} \tag{2}$$

We also need the variance of the interaction term (Aiken & West, 1991, p. 179, Equation A.8):

$$\sigma^2_{SAT\cdot SES} = \sigma^2_{SAT}\overline{SES}^2 + \sigma^2_{SES}\overline{SAT}^2 + 2\sigma_{SES,SAT}\overline{SES}\cdot\overline{SAT} + \sigma^2_{SES}\sigma^2_{SAT} + \sigma^2_{SES,SAT} \tag{3}$$

(Though not necessary, we could also use Equation 2 to estimate $cov[SAT, SES \cdot SAT]$). At this point, we have a corrected variance/covariance matrix of the predictors:

[1] The mean of SES may not be known, but can be estimated: $\overline{SES} = b_0 + b_1 \times \overline{SAT}$, where b_0 and b_1 are the regression coefficients from the model predicting SES from SAT.

$$\Sigma' = \begin{bmatrix} \sigma^2_{SAT} & \sigma\prime_{SAT,SES} & \sigma\prime_{SAT,SES \cdot SAT} \\ \sigma\prime_{SES,SAT} & \sigma\prime^2_{SES} & \sigma\prime_{SES,SES \cdot SAT} \\ \sigma\prime_{SES \cdot SAT,SAT} & \sigma\prime_{SES \cdot SAT,SES} & \sigma\prime^2_{SES \cdot SAT} \end{bmatrix}$$

(Note: anything with a prime ($\prime$) indicates a corrected estimate).

This variance/covariance matrix can be inputted into the PL correction (as before) to obtain a corrected covariance matrix between all variables. I call this estimate r_{pl}, for Pearson-Lawley. The Case III correction, I call r_{c3}.

To review, the PL-based correction (r_{pl}) for centered predictors is performed as follows:

1. Use the PL to estimate the covariance matrix between SES and SAT.
2. Use Equations 1-3 to complete the third rows/columns in Σ'.
3. Use the corrected Σ' to obtain the final corrected covariance matrix.

Recall that the corrections from Aiken and West (1991) require symmetrical data. What is unknown is how robust r_{pl} is to skewness. In the following section, I introduce the Monte Carlo I used to assess r_{pl} under a variety of conditions.

Method

To assess the performance of r_{pl}, I performed a simulation by doing the following:

(1) Generate n skewed SES and SAT scores, with means of 5 and 500, respectively, and variances of one. The skewness values varied as shown in Table 1.[2]

(2) Center the predictor variables.

(3) Create a product variable ($SAT \cdot SES$) by multiplying SES and SAT scores.

(4) Generate 100 $FYGPA$ scores, using the regression weights shown in Table 1.

[2] Many of these parameters were not varied because they made little difference in preliminary simulations. Only b_{sat}, r, and skew affected of bias. Full details of this preliminary simulation are available from the author.

(5) Simulate selection on SAT, by omitting SES, $SES \cdot SAT$, and Y values for those who fell below the p percentile of SAT.

(6) Compute the correlation between SES and $FYGPA$ using r_{pl} and r_{c3}.

(7) Repeat 10,000 times.

Both r_{pl} and r_{c3} were averaged across conditions and compared to the average estimates obtained from the random sample:

$$Bias = \hat{r} - r$$

where $\hat{r}$ is the estimate of interest (either r_{c3} or r_{pl}) and r is the mean estimate from the random sample.

Results

Figure 1 shows bias as a function of skewness (s), the correlation between SES and SAT (r), and the slope predicting $FYGPA$ from SAT (b_{ses}, though to save space in the plot, I have labeled it b). Each dot in the plot represents the cell mean, averaged within the conditions labeled on the x-axis. I labeled the various values of b only once since they repeat across the plot and I wanted to avoid visual clutter. I also added a horizontal line at zero to indicate where Bias = 0. The r_{pl} estimate is in gray with closed circles, while the r_{c3} estimate is in black with open circles.

In nearly every condition, r_{pl} outperforms r_{c3}; the r_{pl} (gray) estimates are very near the horizontal line. The only time r_{c3} performs as good or better than r_{pl} is when skewness is positive, and r and b are high. Otherwise, r_{pl} always outperforms the other estimate. In addition, r_{pl} is generally unbiased, even under heavy skew. It performs poorest when skewness is positive, and r and b are high, reaching approximately -0.08 (meaning the actual correlation is underestimated by 0.08). Also, r_{c3} almost always overestimates, while r_{pl} may underestimate or overestimate, depending on the values of skewness, r, and b.

Discussion

Centering predictors is often recommended to enhance parameter interpretation and reduce multicollinearity. I have shown a major disadvantage of centering predictors: they increase bias under missing data. Centering predictors strips “nonessential” correlation between the interaction the predictor variables. Subsequently, the predictors are unable to augment the missing data model and mitigate bias.

Fortunately, there need not be a trade-off between bias and the advantages of centering predictors. In this paper, I developed a correction that allows researchers to center predictors when data are missing. This correction assumes symmetrical predictors. However, the simulation demonstrated that r_{pl} was robust to fairly extreme skewness, and usually outperformed Case III (which assumes no interactions exist between the predictors). Never did average bias exceed 0.08. The Case III correction (r_{c3}), on the other hand, performed poorly, sometimes exceeding 0.2 in bias.

Because of r_{pl}’s marginal sensitivity to skew, I recommend caution when researchers use the correction. Univariate distributions ought to be inspected for symmetry and transformed when appropriate. I would not, however, recommend using r_{c3} when interactions exist.

Although r_{pl} minimizes bias when centering predictors, there is no reason not to use it when variables are *not* centered. When variables are left uncentered, *some* bias is expected (because the data are technically MNAR). Consequently, I recommend researchers inspect predictor/criterion relationships for potential interactions before applying Case III. If interactions are suspected, the r_{pl} correction will generally lead to unbiased estimates of the population correlation.

References

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: SAGE Publications, Inc.

Aitken, A. C. (1935). Note on Selection from a Multivariate Normal Population. *Proceedings of the Edinburgh Mathematical Society*, *4*(2), 106–110. doi:10.1017/S0013091500008063

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Routledge.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

Dalal, D. K., & Zickar, M. J. (2012). Some Common Myths About Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research Methods*, *15*(3), 339–362. doi:10.1177/1094428111430540

Echambadi, R., & Hess, J. D. (2007). Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models. *Marketing Science*, *26*(3), 438–445. doi:10.1287/mksc.1060.0263

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean Centering in Moderated Multiple Regression: Much Ado about Nothing. *Educational and Psychological Measurement*, *58*(1), 42–67. doi:10.1177/0013164498058001005

Lawley, D. N. (1944). A Note on Karl Pearson's Selection Formulae. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, *62*(01), 28–30. doi:https://doi.org/10.1017/S0080454100006385

Little, R. J. A., & Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi:10.1093/biomet/63.3.581

Table 1

Parameters Used for the Monte Carlo Simulation

Parameters	Values
b_{ses}	0.242
b_{sat}	0, 0.099, 0.242, 0.371
$b_{ses \cdot sat}$	0.242
r	0, 0.1, 0.3, 0.5
$s\bar{e}s$	5
$s\bar{a}t$	500
n	50, 100, 200, 500
$p_{missing}$	0.1, 0.3, 0.5, 0.7
skew	-100, -50, 0, 50, 100

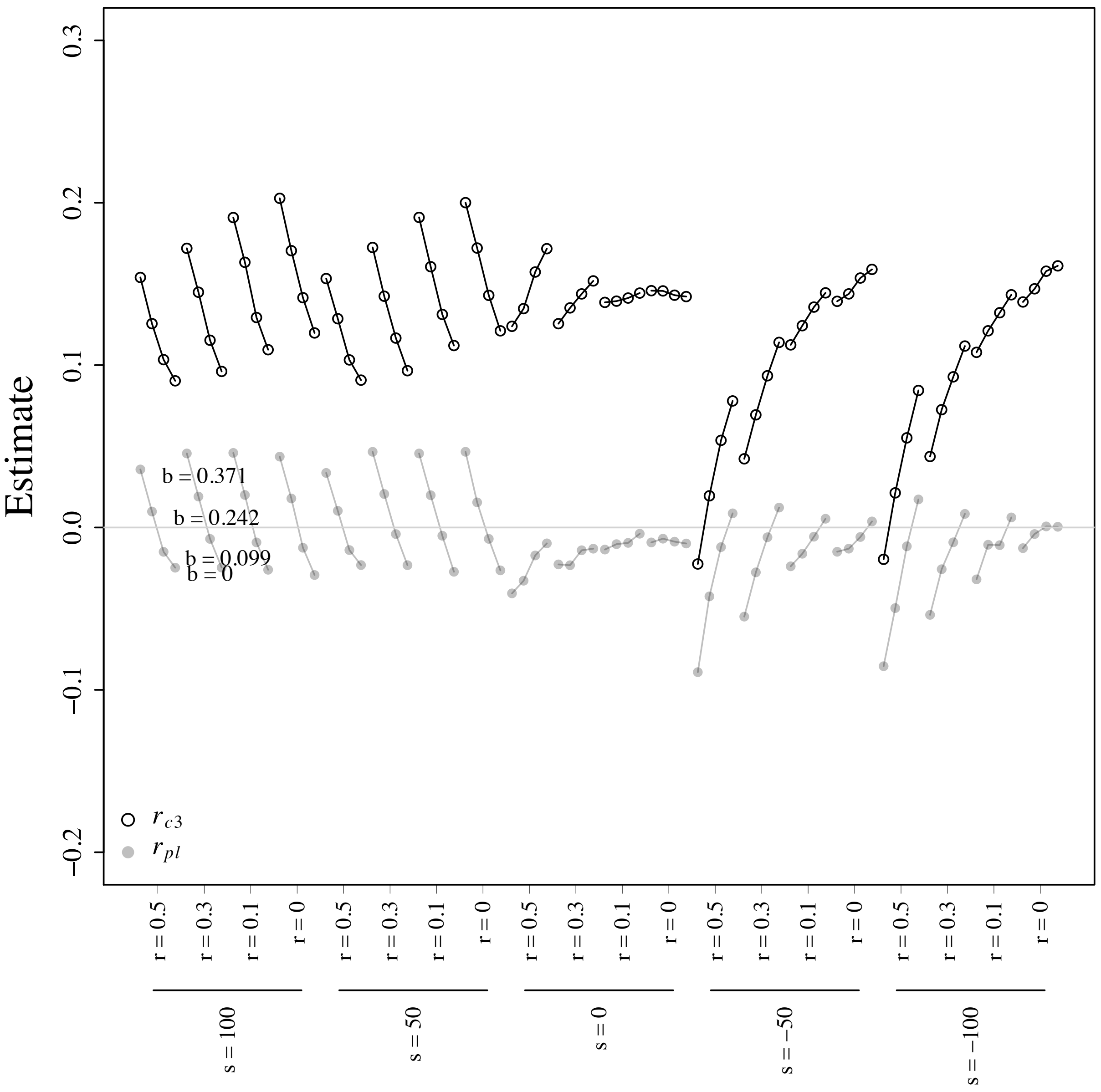

Figure 1. Average bias in estimating the correlation coefficient, under various conditions: correlation between the predictor variables (r), skewness (s), the slope between SAT and $FYGPA$ (b), and estimator (r_{pl} vs r_{c3}). Note that each line shows bias as a function of the values of b (b=0.371, 0.242, 0.099, 0). These values are repeated, though only the first are labeled.