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

Dustin Fife[1]

[1] Rowan University

Author Note

Correspondence concerning this article should be addressed to Dustin Fife , 201 Mullica Hill Road Glassboro, NJ 08028 . E-mail: fife.dustin@gmail.com

Abstract

It is commonly advised to center predictors in multiple regression, especially in the presence of interactions or polynomial terms (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) will reduce multicollinearity. However, in this paper, I demonstrate that in some missing data situations, centering predictors biases parameter estimates and decreases precision. I also develop a Pearson-Lawyley-based (Aitken, 1935; Lawley, 1944) missing data correction (called r_{pl}) that does not require uncentered predictors, then evaluate the performance of this correction via Monte Carlo Simulation.

Keywords: missing data, selection, range restriction, interactions

Word count: X

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

It is commonly believed (and stated) that when performing a multiple regression, researchers ought to center predictor variables (which consists of subtracting the mean of the predictor variable[s] from every score; J. Cohen et al., 2013). Doing so often improves the interpretation of parameter estimates, especially when zero has no meaningful interpretation (e.g., IQ). It has also been suggested that centering predictors reduces multicollinearity, thereby increasing precision of parameter estimates (J. Cohen et al., 2013). Although this last advantage has been hotly debated (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Kromrey & Foster-Johnson, 1998), none (that I know of) have recommended against centering predictors (except when the metric of the predictor variable has a meaningful zero point). However, there is one situation where centering predictors will not only *decrease* precision, but it will also bias parameter estimates.

In this paper, I show how centering predictors can lead to substantial bias when data are missing (such as when subjects drop out of a study or some sort of selection procedure is operating). In order to do so, I will first briefly review the literature on centering predictors and show the mathematical advantages of doing so. After which, I will review the missing data literature and show how centering predictors may change a "Missing At Random" situation into on that is "Missing *Not* at Random," which I will then highlight via a simulated example. Finally, I will introduce a correction that allows researchers to center predictors without bias, and assess its performance via Monte Carlo Simulation.

Regression and Centering

In a multiple regression context, interactions often exist between two or more predictor variables. Suppose, for example, an academic institution is interested in assessing the impact of socioeconomic status (SES) on $FYGPA$, and wishes to correct for missing data (in this case, let us assume the university selected based on SAT scores). Further suppose that these two predictor variables interact, such that for those with high SAT scores, SES is more predictive of $FYGPA$ scores than for those with lower SAT scores. Mathematically,

$$FYGPA = b_0 + b_1SES + b_2SAT + b_3SES \cdot SAT$$

where b_3 will be some positive value (indicating that as SAT gets higher, SES will become more predictive of FYGPA). In this situation, the researcher might be inclined to center both SES and SAT scores. Doing so supposedly has two advantages. First, the coefficients for the transformed variables will have a more sensible interpretation (J. Cohen et al., 2013). The original zero points for the predictors are meaningless, which means that the intercept parameter is of little interest (in this case, the predicted $FYGPA$ for someone who has an SAT/SES of zero). Centering these predictors now yields a meaningful interpretation (the predicted $FYGPA$ for someone who has an average SAT and average SES score).

This interpretive advantage is heightened when interactions are present in the model. Recall that when an interaction is present, the relationship between, say 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, for example, is the *average* slope of $FYGPA$ on SES across all values of SAT.

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

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

(Note that the above equation only applies when each predictor is completely symmetrical). When both predictors are centered, the means become zero and the covariance between the two vanishes. This is what we call “nonessential multicollinearity,” or the collinearity that is attributable to the means of the predictors.

When the predictors are *not* symmetrical, some relationship between the two will remain. What remains is what is called "essential" multicollinearity.

Some (e.g., J. Cohen et al., 2013) argue that removing essential multicollinearity will increase the precision of parameter estimates. The rationale is simple, as multicollinearity tends to inflate standard errors. However, others (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Kromrey & Foster-Johnson, 1998) have demonstrated mathematically that precision is unaffected by centering.

Regardless of how it does (or does not) affect precision, centering is often considered wise practice, if at least for its interpretative advantages. However, when data are missing (such as due to selection or attrition), centering may inflate bias and standard errors.

Missing Data

To understand how centering may exacerbate bias, let us review the missing data nomenclature. Rubin (1976) developed a framework under which to view missing data. He considered three missing data situations: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not At Random (MNAR; Little & Rubin, 2014; Rubin, 1976).

MCAR

When data are MCAR, the probability of missingness is unrelated to either the observable or the unobservable data. Put differently, those values missing can be considered a random sample of all the available data (Graham, 2012). In this situation, nearly all methods of handling missing data (e.g., listwise deletion, mean imputation, maximum likelihood) will yield unbiased parameter estimates (Enders, 2010; Graham, 2012).

As an example, suppose some of the students have missing SES scores because of a computer outage that selectively wiped some students' data. Because the probability of a computer outage is unlikely to be related to $FYGPA$, this is a MCAR situation.

MAR

When data are MAR, the probability of missingness is correlated with the observable data, but *not* the unobserved data. For example, suppose the university selected students into the university based on their SAT scores. Naturally, those not selected will be missing $FYGPA$ scores. However, because we have measured and recorded the cause of missingness (SAT scores), it is possible to obtain unbiased estimates of model parameters, provided that the appropriate method of analysis is used (e.g., maximum likelihood methods or multiple imputation).

MNAR

Finally, when data are MNAR, the probability of missingness is correlated with *both* the observable data and unobservable data. For example, suppose at the end of the first academic year, not only are $FYGPA$ scores missing for those who were not selected into the university, but some are missing because they dropped out of the university due to lack of motivation. In this instance, motivation is the cause of missingness, but because the researcher did not measure motivation, the cause of missingness is unobservable.

When data are MNAR, it is difficult to obtain unbiased estimates without making quite restrictive assumptions (Enders, 2010; Heckman, 1979).

Interactions and Missing Data

In concurrent validity designs, when interactions exist in a regression model, the data are technically MNAR. To understand why, consider our previous example. Again, suppose students were selected based on SAT scores. Now let us further suppose the researcher is interested in assessing the correlation between socioeconomic status (SES) and $FYGPA$ on the current cohort of applicants. However, they want to know the correlation in the unselected population, but unfortunately only have applicant 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 technically observable and missingness due to this is MAR.

(2) $SAT \cdot SES$ scores. This product term is correlated with the probability of missingness in such a way that 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. Because we have an independent correlate of missingness (the product) that cannot be observed (because scores of students not selected into the university are missing), the 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 altogether problematic. Recall that one need not actually model the cause of missingness to render a situation MAR. Rather, one simply needs to model a correlate of the cause of missingness (Collins, Schafer, & Kam, 2001). With uncentered variables, the correlation between each of the predictors and their product is quite high and thus, even though the product term is a cause of missingness, we can actually control for it using the applicant SAT scores. When we center the variables, however, that correlation vanishes and the MNAR-ness of the data is exacerbated.

Naturally, a resourceful researcher might decide to include the interaction term as a predictor in a regression model, assuming that by including the cause of missingness they will render the data MNAR. Alas, this is not so because the product scores ($SES \cdot SAT$) for those who were not selected into the university are still missing. Consequently, without modification to the algorithms, there is no way to obtain an unbiased estimate using current missing data techniques.

Demonstration

To illustrate this problem (centering predictors exacerbates bias when interactions are present), I performed a simulation by doing the following[1]:

(1) Generate 100 fictitious SES and SAT scores. These scores were generated from a random normal distribution. The means were set to 10 and 3, respectively, their standard deviations to 3 and .2, and their covariance to 90 (which is equivalent to a 0.3 correlation).
(2) Standardize the predictor variables. On half of the iterations (see step 7), the predictor variables were centered around their mean.
(3) Create a product variable ($SAT \cdot SES$) by multiplying SES and SAT scores.
(4) Generate 100 $FYGPA$ scores, using the following equation:

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

The values for $b_0 - b_3$ were chosen such that $FYGPA$ had an expected value of 3.0 and a standard deviation of 0.4. The standardized slopes ($b_1 - b_3$) were set to 0.3. (Note that the value of unstandardized values of $b_0 - b_2$ changed depending on whether the current iteration was standardized.)
(5) Simulate selection on SAT. To do this, I set SES, $SES \cdot SAT$, and Y to missing for those individuals who had SAT scores below the mean (approximately 0 or 500, depending on whether the current iteration was standardized). For comparison, I also created a separate dataset which was simply a random sample of half the scores.
(6) Compute the corrected and uncorrected correlation between SES and $FYGPA$. To correct, I used both the expectation maximization (EM) algorithm (via the em.norm function in the norm package in R; Schafer, Novo, & Fox, 2010), as well as the Case III correction (via the caseIII function in the selection package in R; Fife, 2016). Note

[1]Complete access to the code that generated the data is freely available from the author:

that the standard Case III correction ignores the fact that there is an interaction present. For comparison, I also computed the simple correlation in the random sample.

(7) Repeat 10,000 times. To estimate bias and assess standard errors, these steps were repeated 10,000 times.

The results of this simulation are presented in Figure 1. The shaded boxes represent the distribution of estimates from the uncentered conditions while the open boxes represent those from the centered ones. Notice that, even when the variables are not centered, Case III and the EM are biased, though not by much (an average of 0.01 for both estimates). Again, the reason they are only slightly biased is because SAT is highly correlated with $SAT \cdot SES$. When centered, however, bias is much worse (an average of 0.18 for both). In addition, standard errors are slightly larger when centered (0.12 for Case III/EM in the centered condition and 0.11 in the uncentered condition).

These results demonstrate a clear advantage to *not* centering variables when an interaction is present (at least when missing data are involved).

Potential Solutions

The obvious solution to the bias problem is to simply not center variables. However, this may not be ideal if a researcher is keen on the interpretive benefits of centering. Consequently, I offer a correction.

Recall that the Pearson-Lawley correction (Aitken, 1935; Lawley, 1944) provides a way to correct estimates for missing data that occurs on one or more variables. It is a multivariate extension of the traditional Case III correction and requires two inputs:

(1) The unrestricted (unbiased) variance/covariance matrix of the variables responsible for missingness. In this case, that would be SES and $SES \cdot SAT$.

(2) The restricted (biased) variance/covariance matrix of all the variables in the model. In this case, that would be SES, SAT, $SES \cdot SAT$, and $FYGPA$.

In matrix form, we need the following population variance/covariance matrix

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

And the following restricted variance/covariance matrix

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

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

To compute the covariance between SAT and $SAT \cdot SES$, we can use Equation 1. Unfortunately, this requires knowing SES^2, or the unrestricted (unbiased) variance of SES. Because these data were collected on incumbents, we don't have access to that information. However, this parameter can be acquired using the PL correction, assuming we have access to the incumbent data for SAT. (If that information is unavailable, we direct the reader to Fife, Hunter, & Mendoza, 2016, who offer corrections for situations where incumbent data is unavailable). One simply inputs the 1×1 matrix of SAT^2 (i.e., the variance) as the unrestricted variance/covariance matrix, then subsequently, inputs the restricted estimates for the SES/SAT variances, as well as their covariance.

After performing the PL correction, we now have most[2] of the inputs necessary for Equation 1. While we are at it, we might as well compute the population covariance between SES and $SES \cdot SAT$ (Aiken & West, 1991, p. 180, Equation A.13):

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

[2]The mean of the incidentally restricted variable (SES in this case) may not be known. However, one can estimate this using the following equation: $\overline{SES} = b_0 + b_1 \times \overline{SAT}$, where b_0 and b_1 are the regression coefficients from the model predicting SES from SAT.

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

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

At this point, we have a corrected variance/covariance matrix of the predictors:

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

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

This variance/covariance matrix can then be inputted into the PL equation (as before) to obtain a doubly corrected variance/covariance matrix between the predictors and the outcome. For simplicity, we will call this estimate r_{pl}, for Pearson-Lawley. The standard correction (using Case III and ignoring the interaction term), we will 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 variance/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 variance/covariance matrix.

Recall that the corrections from Aiken and West (1991) require that the data are symmetrical. What is unknown is how robust this correction is in the presence of skewness. It is also unknown how this correction fares in terms of standard errors. In the following section, I introduce the Monte Carlo Simulation I used to assess the performance of the correction under a variety of conditions.

Method

The Monte Carlo simulation was nearly identical to the simulation in the demonstration, with the exception of the parameters varied. The parameters varied are shown in Table 1.[3] In short, I did 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) 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.
(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.

There is one other detail worth mentioning. The population value of the correlation is less tractable when the data are skewed. Since skewness tends to attenuate correlation coefficients, I instead compared the average r_{pl} (i.e., averaged across the conditions listed in Table 1) and r_{c3} values to the averaged random sample values. In the results that follow, bias values are reported relative to the random sample. That is,

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

[3] Many of these parameters were not varied because they made little to no difference in preliminary simulations. These preliminary simulations randomly varied every parameter using a random uniform distribution. Subsequently, a Random Forest (RF; Breiman, 2001) model was used to determine which parameters were predictive of bias. The benefit of RF is that it natively detects interactions, which is clearly important in this situation. After performing the RF, only b_{sat}, r, and skew were predictive of bias. (Note that we only predicted bias for the r_{pl} estimate. Had we also predicted bias for the r_{c3} estimate, other variables may have also been predictive). We also varied $p_{missing}$ and n since they will affect standard errors. Full details of this preliminary simulation are available from the author.

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 2 shows how bias changed 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 mean, collapsed across the conditions labeled on the x-axis. Note that I have only labeled the various values of b only once since they repeat across the plot and I wanted to avoid visual clutter. I have 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.

The first thing to notice is that, in nearly every condition, r_{pl} outperforms r_{c3}; across nearly all conditions, 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 fairly heavy skew. It performs poorest when skewness is positive, and r and b are high, reaching values of approximately -0.08 (meaning the actual correlation is underestimated by 0.08). It is also worth noting that r_{c3} almost always overestimates, while r_{pl} may underestimate or overestimate, depending on the values of skewness, r, and b.

Figure 3 shows the empirical standard errors from the same simulation. Here, standard errors are plotted against n, proportion missing (p), and skewness (s). As before, the light-colored line (with solid circles) is the r_{pl} estimate and the dark line (with open circles) is the r_{c3} estimate.

Not surprisingly, standard errors increase as n decreases and as p increases. In addition, skewness also influences standard errors; as data become more negatively skewed, standard errors increase, at least when more than 50% of data are missing. Finally, r_{pl} has larger standard errors than r_{c3}, at least when a large proportion of data are missing (>50%).

This advantage is much smaller for lower values of p.

Discussion

Centering predictors in multiple regression is often recommended as a method of enhancing the interpretation of parameters and reducing multicollinearity. In this paper, I have shown that a major disadvantage of centering predictors is that it may increase bias and decrease precision when data are missing. The reason is because centering predictors strips "nonessential" correlation between the interaction variable and the outcome. The net result of this is that other variables in the model 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 have developed a correction, which allows researchers to center predictors even when data are missing. Unfortunately, this correction relies on the assumption of skewness. However, the Monte Carlo simulation demonstrated that this correction (r_{pl}) was generally robust to fairly extreme skewness, and usually outperformed the standard Case III correction (which assumes no interactions exist between the predictor variables). Never did average bias values exceed 0.08. The Case III correction (r_{c3}), on the other hand, performed quite poorly; sometimes it exceeded 0.2 in bias, though it did tend to have smaller standard errors than the PL correction (at least when the proportion missing was more than 50%).

Because of r_{pl}'s marginal sensitivity to skew, I recommend caution when researchers attempt to use the correction. Univariate distributions ought to be inspected for symmetry and, when not symmetric, transformations may be applied. I would not, however, recommend using the standard Case III correction. As this simulation shows, Case III tends to over-correct when interactions exist.

Although r_{pl} is intended to minimize bias when centering predictors, there is no reason not to use it when variables are *not* centered. As shown in Figure 1, even if variables are left uncentered, some bias is expected. In this demonstration, average bias values only reached 0.01. However, there may be a different set of values (e.g., correlations between the predictors, sample sizes, means, variances) that might yield substantial bias. Consequently, I recommend applied researchers always 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

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

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

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fife, D. (2016). Selection: Estimating unattenuated correlations on range restricted correlations. Retrieved from https://CRAN.R-project.org/package=selection

Fife, D., Hunter, M. D., & Mendoza, J. L. (2016). Estimating Unattenuated Correlations With Limited Information About Selection Variables: Alternatives to Case IV. *Organizational Research Methods*, *19*(4), 593–615. doi:10.1177/1094428115625323

Graham, J. W. (2012). *Missing Data: Analysis and Design*. New York, NY: Springer.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161. doi:10.2307/1912352

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

Schafer, J. L., Novo, A. A., & Fox, J. (2010). *Norm: Analysis of multivariate normal datasets with missing values*. Retrieved from http://CRAN.R-project.org/package=norm

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

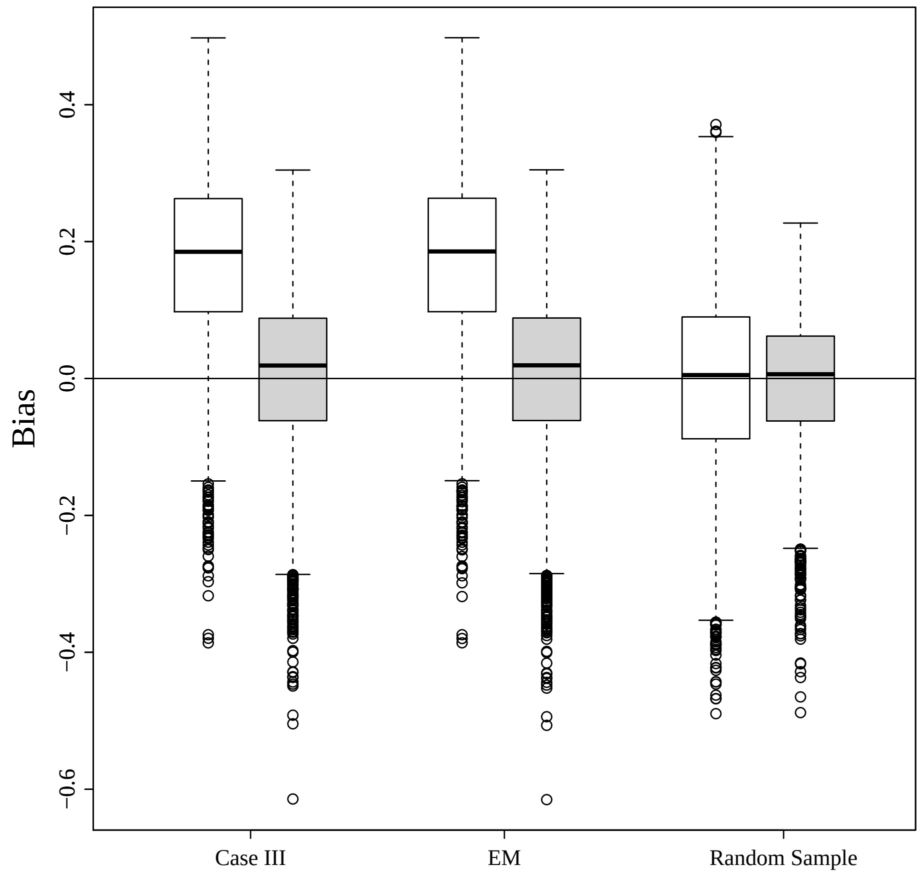

Figure 1. Boxplots showing the distribution of two correction procedures (Case III and the EM algorithm) relative to a random sample, across 10,000 iterations. The shaded boxes represent the distribution of estimates from the uncentered predictors, while the un-shaded boxes are from the centered predictors.

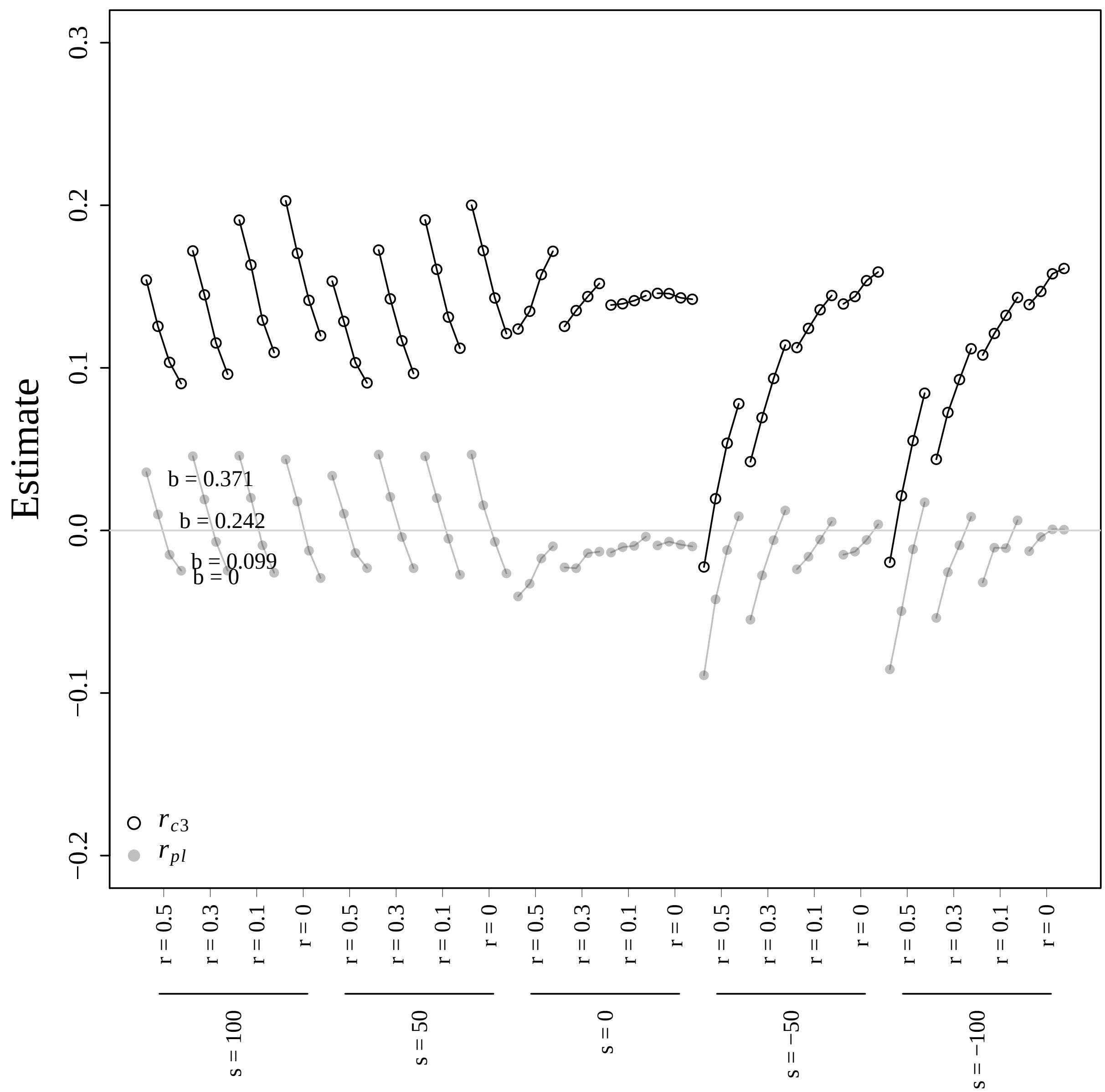

Figure 2. Average bias in estimating the correlation coefficient, under various conditions: correlation between the predictor variables (r), skewness (s), the slope between SAT and $FYGPA$ (b), and estimator (r_{pl} vs r_{c3}). Note that each line shows bias as a function of the values of b (b=0.371, 0.242, 0.099, 0). These values are repeated, though only the first are labeled.

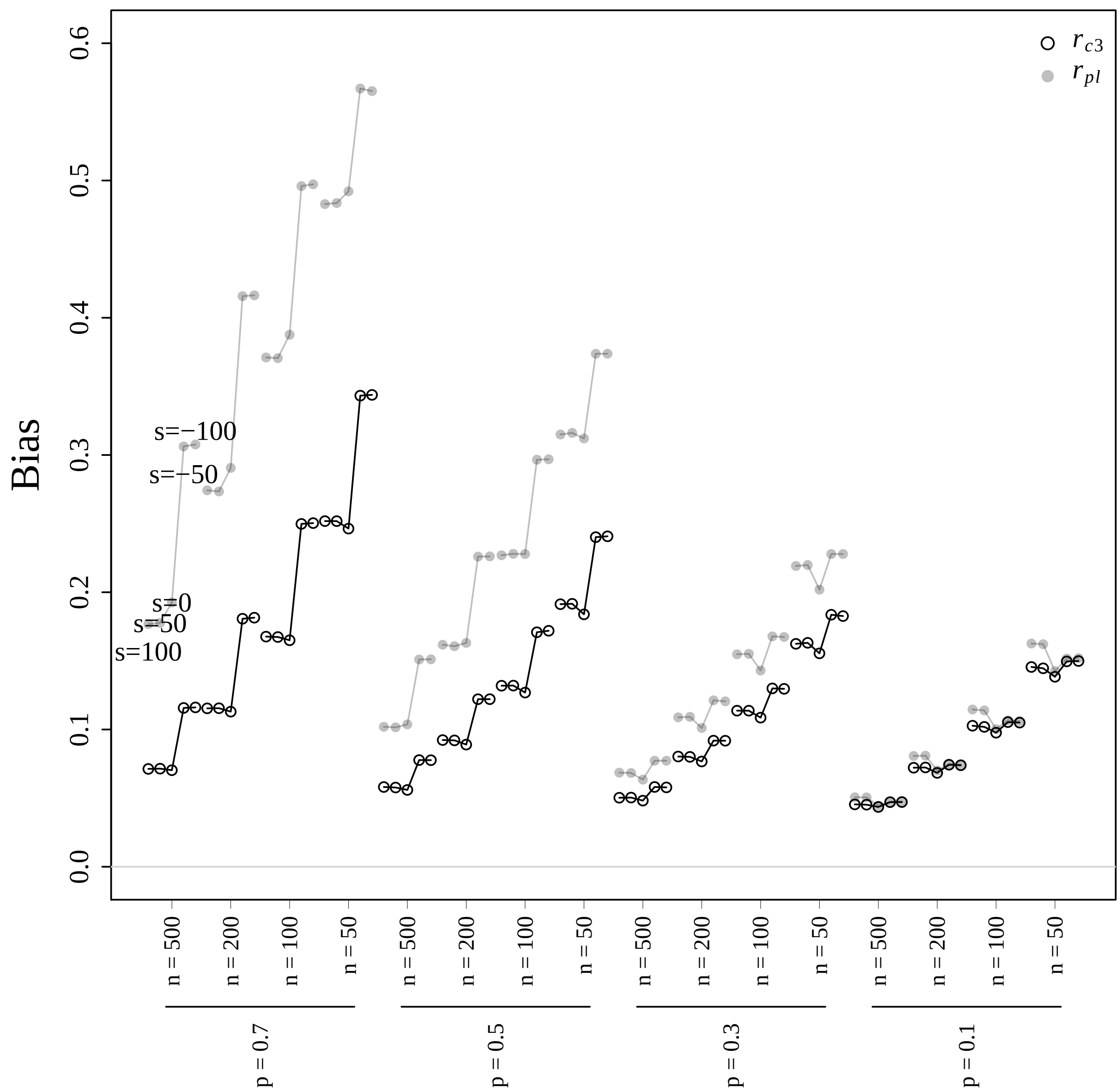

Figure 3. Standard errors in estimating the correlation coefficient, under various conditions: sample size (n), skewness (s), the proportion of missing data (p), and estimator (r_{pl} vs r_{c3}). Note that each line shows bias as a function of the values of s (s=100, 50, 0, -50, -100). These values are repeated, though only the first are labeled.