Running head:

When Interactions Bias Selected Corrections

Dustin A. Fife

Department of Psychology

Rowan University

Christopher Lacke

Department of Mathematics

Rowan University

Jorge L. Mendoza

Department of Psychology

University of Oklahoma

Dustin Fife

Department of Psychology

University of Oklahoma

Norman, OK

73071

e-mail: dfife@ou.edu

Abstract

When Interactions Bias Selected Corrections

Introduction

In the early 1900s, Pearson developed a set of procedures aimed at correcting correlation coefficients that had been attenuated through selection (Pearson). Since their inception, these corrections (and their extensions; e.g., rubin) have been widely used in psychology (references), medicine (reference), education (references), and even xxxx (references). Yet these missing data strategies require an assumption that is rarely recognized, let alone tested, namely that the relationship between the outcome variable and the variable of interest is linear.

Clearly, in certain situations, this assumption is untenable (e.g., the "U-shaped" relationship between stress and performance). Lord and Novick (1968) noted that test scores have a tendency to violate both the assumption of linearity and homoscedasticity at the extremes of a distribution. Empirical data exist that suggest these violations are not uncommon. In particular, several researchers have found concave relationships between test scores and criteria (Arneson et al., 2011; Cullen et al., 2004; Lee & Foley, 1986). We are inclined to believe most researchers would recognize this as a serious limitation before proceeding with corrections. However, violating the assumption of linearity may be more common than most researchers assume. In particular, if there exists an interaction between the outcome variable and the predictor variables, the linearity assumption will be violated and parameter estimates will be biased. Given that interaction effects are more common than not (reference), we are inclined to believe this may be a serious problem.

The purpose of this paper is three-fold: first, we demonstrate how interactions bias parameter estimates and investigate the degree of bias. Second, we develop a new set of correction procedures that explicitly model interactions and eliminate bias. Finally, we turn to the literature to determine the extent to which interactions may have biased previous corrections.

For the remainder of the paper, we will proceed as follows. We first review the correction procedures (e.g., Case II, Case III, Multiple Imputation). Next, we demonstrate through Monte Carlo simulations how interactions bias parameter estimates under incidentally selected range restriction. Subsequently, we and review what others have demonstrated about the sensitivity of these estimates to non-linearity. Next, we introduce our corrections and investigate their performance through Monte Carlo simulation. Finally, we conclude with recommendations about how researchers might identify and overcome non-linearity problems in applied missing data scenarios.

Case II, Case III, and Missing Data

Thorndike (reference) originally classified Pearon's correction procedures into Case II and Case III. (Case I also exists, but it's rare...or soemthing like that). The Case II correction is used when the researcher is interested in estimating the correlation coefficient between the outcome of interest (Y) and a variable that is a cause of missingness. For example, suppose an organization uses cognitive ability scores to select employees into their organization. The organization later wishes to compute the corrected (i.e., unselected) correlation coefficient between cognitive ability and job performance (Y). In this case, the variable of interest (cogntive ability) has been *directly* used for selection (or, in missing data nomenclature, cognitive ability is a cause of missingness). In this situation, one simply needs to know the unrestricted variance of cognitive ability to obtain an unbiased estimate of the population correlation coefficient:

$$r_{XY} = \frac{r'_{XY}(\frac{s_X}{s'_X})}{\sqrt{1 - r'^2_{XY} + r'^2_{XY}\frac{s^2_X}{s'^2_X}}} \tag{1}$$

Now let us suppose the researcher is interested in investigating the relationship between a third variable (say conscientiousness) and the outcome (job performance). The Case II correction will not work because conscientiousness has been incidentally selected (i.e., missingness in conscientiousness occurred because of cognitive ability). Instead, the Case III correction is used

$$r_{XY} = \frac{r'_{XY} - r'_{ZX}r'_{ZY} + r'_{ZX}r'_{ZY}(\frac{s_Z}{s'_Z})}{\sqrt{[1 - r'^2_{ZX} + r'^2_{ZX}\frac{s^2_Z}{s'^2_Z}][1 - r'^2_{ZY} + r'^2_{ZY}\frac{s^2_z}{s'^2_Z}]}} \tag{2}$$

Note that this assumes that missingness is caused by cognitive ability alone. Or, put a different way, once we control for cognitive ability, there is no correlation between the probability of missingness and conscientiousness.

Decades after Pearson developed the correction procedures that are now maistream in organizational psychology, Rubin (reference) developed a more general framework under which to view selection. Rubin demonstrated that if one measures the cause of missingness and include it in a missing data model, bias can be eliminated (provided the appropriate missing data strategies are used; reference). His framework identifies three forms of missing data: Missing Completely at Random (MCAR), which means that the probability of missingness is uncorrelated with either the observable or the unobservable data; Missing at Random (MAR), which means that the probability of missingness is correlated with the observable data, but *not* the unobservable data; and Not Missing At Random (NMAR), which means that the probability of missingness is correlated with both unobservable and observable data. Under either the MAR or MCAR conditions, unbiased parameter estimates can be obtained using either Maximum Likelihood methods (e.g., Full Information Maximum Likelihood or the Expectation Maximization algorithm) or Multiple Imputation (MI).

Though not originally conceptualized as an extension of the selection literature, various authors (references) have noted that selection is a special case of missing data.

Like Rubin's framework, Case II and Case III both require population estimates from the variable that causes selection. By supplying these estimates (i.e., population variances), the cause of missingness becomes "observable" (or "ignorable" in Rubin's terminology, xxx, page xx).

The advantage of Rubin's framework, however, is that it allows greater flexibility than the standard Case II and Case III corrections. For example, if selection is not caused by one variable, but by a battery of variables (or a function of a battery of variables), Rubin's framework allows seamless estimation of the parameters of interest. (Although the Pearson-Lawley multivariate correction can also handle many of these more complicated situations).

For this paper, we focus on situations where three or more variables are used (e.g., a selection variable, an incidentally selected variable, and an outcome variable), and seek to identify how Case III and other missing data strategies (e.g., Maximum Likelihood and Multiple Imputation) are affected by nonlinearity caused by interactions.

Example and Demonstration

Throughout this paper, we will make use of a simple example. Suppose a university selects students based on SAT scores (Z). Later, they wish to estimate the correlation between high school GPA (X) and freshman GPA (Y). Further suppose HS GPA and SAT scores interact with one another in producing freshman GPA. Perhaps, for example, HS GPA is less predictive of first year performance for those higher in SAT scores than it is for those who are lower. In this situation, estimates between HS GPA (X) and first year GPA will be biased if corrected with Case III.

To demonstrate this fact, we did the following:

1. Generate 100 pairs of scores (X and Z) with a correlation of 0.3. X scores were generated to have a mean of 3.0 and a standard deviation of .4, while Z scores had a mean

of 500 and a standard deviation of 100. These were designed to simulate HS GPA and SAT scores, respectively.

2. Create an interaction variable by multiplying X and Z together.

3. Generate Y scores. Y was generated in such a way it had a mean of 3.0 and standard deivation of 0.4. It also had standardized β weights of 0.3 with each of the variables (Z, X, and ZX).

4. Select the top 50% of scores based on Z. All scores on the Z variable were sorted, then the bottom 50% of scores on X, ZX, and Y were set to missing.

5. Create a random sample. For comparison, we will randomly select 50% of the scores from the original dataset and compute the correlation coefficient.

6. Estimate the correlation between X and Y using the Case III and EM algorithm corrections. These corrections were compared to the random sample estimate.

7. Repeat 10,000 times. In order to simulate the sampling distribution and assess variability, we performed the Monte Carlo 10,000 times.

Figure 1 shows the distribution of both correction procedures (and the random sample), across the 10,000 iterations. Notice that both the EM and Case III overestimate the population correlation. On average, both Case III and the EM overcorrect by 0.188, relative to a random sample.

The reason for this overcorrection can be illustrated in Figure 2. This image shows a scatterplot of 2,000 datapoints between X and Y, using the same parameters as the Monte Carlo. With Case III (and the EM), the procedures will use the available data to project the estimate into the unselected population. In the figure, the available data is represented by the color orange, and the best-fitting line for the available data is the solid orange line. The missing data are represented by the blue-colored dots (and the corresponding regression line is shown in blue). The solid black curve is a lowess line through both the available and unavailable data. Notice that at about the halfway point, the lowess curve bends upward, indicating that the relationship between X and Y is steeper for those with high *X* scores (and, in addition, it is more predictive for those for whom we have available data). Any attempt to project the estimates from those with higher *X* scores into the range of those who have lower *X* scores will tend to overestimate the correlation coefficient.

Figure 2 demonstrates why the assumption of linearity is so important to standard correction procedures; if the curves alter trajectories outside the available data, there is no way to estimate corrections using standard corrections. Others have noted corrections are generally not robust to violations of the linearity assumption. For example, some (e.g., Greener & Osburn, 1979, 1980; Gross, 1982; Gross & Fleischman, 1983) have investigated how correction procedures perform when either (or both) of these assumptions are violated. Greemer and Osburn (1980) noted that corrected estimates generally perform poorly, and in some cases lead to overcorrection (depending on the form of the distribution, Gross & Fleischman, 1983). In addition, Gross and Fleishman (1987) concluded that unless X and Y are strongly correlated and the sample size is large, it may be best to leave estimates uncorrected.

Potential Solutions to the Non-Linearity Problem

Given the problems noted previously, Culpepper (2016) developed a correction for nonlinear relationships that was adapted from the econometrics literature (Harvey, 1976). His procedure is designed to correct for direct range restriction (i.e., Case II), and assumes a quadratic relationship between *X* and *Y*. Culpepper's correction models heteroskedasticity using a set of coefficients that map the predictor(s) onto the residual variance. Monte Carlo simulations demonstrate that his procedure is able to yield unbiased estimates of unattenuated correlation coefficients.

Although Culpepper's procedure works well under the situations it was designed to work (direct range restriction, quadratic relationship), it is not sufficient to correct for interaction effects for three reasons:

1. If the relationship of interest is the correlation between X and Y, and if selection has occurred on Z, a *direct* range restriction correction formula is not appropriate since this is an *indirect* situation.

2. An interaction between (for example between X and Z) is unlikely to manifest itself as a quadratic relationship between X and Y. Because Culpepper's correction assumes a quadratic relationship, it will likely not work.

3. Culpepper's correction estimates the correlation using the model's estimate of R^2. If the expected relationship is negative, the researcher must change the sign of the estimate. Although this is not difficult to do, there may be situations where the correction changes signs of the uncorrected estimate (e.g., Ree, Carretta, & Albert, 1994). With Culpepper's correction, there would be no way to determine what the true sign is of the uncorrected estimates.

Given these limitation, and given the high frequency of interaction effects, we developed a new set of procedures aimed at correction non-linearity in the presence of interaction terms. In the next section, we develop the framework for the new set of procedures before we proceed to test the performance of the procedures via a Monte Carlo Simulation.

New Corrections

The corrections that we introduce borrow from the mathematical framework of the Pearson-Lawley (PL) correction procedure (reference). Recall that the PL procedure requires two inputs:

1. The unrestricted variance/covariance matrix of the variables responsible for missingness.

2. The restricted variance/covariance matrix of both the variables responsible for missingness as well as the outcome variable(s).

In the case of our example, the cause of missingness is SAT scores. However, if we were to only use population estimates of SAT scores in the PL procedure, the results would be biased. The reason for this is because the interaction term (ZX, or the SAT by HS GPA interaction) is missing not at random (MNAR; Rubin). Recall that data are MNAR if the probability of missingness is correlated with unobservable data. Although selection is technically only caused by Z, the net effect is that missingness is correlated with both Z *as well as* XZ (because Z is correlated with XZ, which is correlated with missingness). Because these ZX scores are missing, the data are MNAR.

Because of this, the unrestricted variance of Z is not sufficient. Rather, we need the following unselected variance/covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma^2_Z & \sigma_{Z,X} & \sigma_{Z,XZ} \\ \sigma_{X,Z} & \sigma^2_X & \sigma_{X,XZ} \\ \sigma_{XZ,Z} & \sigma_{XZ,X} & \sigma^2_{XZ,XZ} \end{bmatrix}$$

The first and second rows/columns of the matrix (i.e., the Z, X variance/covariance matrix) can be acquired using the PL correction, assuming we have access to the population variance of Z. (If that information is unavailable, we direct the reader to Fife...). One simply inputs the 1×1 matrix of Z (i.e., the variance) as the unrestricted variance/covariance matrix, then subsequently, inputs the restricted estimates for the Z/X variances, as well as their covariance.

The third row/column is more difficult to obtain. According to Aiken and West (p. 180, eq. A.15), the covariance between ZX and X is

$$\sigma_{XZ,X} = \sigma^2_X \times \bar{Z} + \sigma_{X,Z} \times \bar{X} \tag{3}$$

Likewise, the covariance between ZX and Z is

$$\sigma_{XZ,Z} = \sigma^2_Z \times \bar{X} + \sigma_{X,Z} \times \bar{Z} \tag{4}$$

Finally, to correct the variance ($\sigma^2_{XZ,XZ}$); See Aiken and West, p. 180, eq. A. 8),

$$\sigma^2_Z \bar{X}^2 + \sigma^2_X \bar{Z}^2 + 2\sigma_{X,Z}\bar{X}\bar{Z} + \sigma^2_X \sigma^2_Z + \sigma^2_{X,Z} \tag{5}$$

Once we have these estimates, we can again use the PL-correction to obtain our final corrected variance/covariance matrix.

To review, the PL-based correction for interaction terms is as follows:

1. Use the PL to estimate the variance/covariance matrix between X and Z
2. Use Equations 3-5 to complete the third rows/columns in Σ.
3. Use the corrected Σ to obtain the final corrected variance/covariance matrix.

As an example, suppose we have the dataset shown in Table 1. The observed (restricted) correlation between HS GPA and first year GPA is 0.53. If we use the PL correction on SAT/HSGPA, assuming selection has taken place on SAT, we obtain the following corrected variance/covariance matrix:

$$\Sigma = \begin{bmatrix} 10000 & 4.26 \\ 4.26 & 0.1 \end{bmatrix}$$

If we apply Equations 3-5, we get the following:

$$\sigma_{X,XZ} = 110.72$$

$$\sigma_{Z,XZ} = 35120.36$$

$$\sigma_{XZ,XZ} = 162427.03$$

Taken together, this yields the following matrix:

$$\Sigma = \begin{bmatrix} 10,000 & 4.26 & 35120.36 \\ 4.26 & 0.10 & 110.72 \\ 35120.36 & 110.72 & 162427.03 \end{bmatrix}$$

which can be inputted into the PL-correction, yielding the following corrected variance/covariance matrix:

$$\Sigma = \begin{bmatrix} 10,000 & 4.26 & 35120.36 & -18221.1936931697 \\ 4.26 & 0.10 & 110.72 & -56.9481670815452 \\ 35120.36 & 110.72 & 162427.03 & -83836.4103891712 \\ -18221.1936931697 & -56.9481670815452 & -83836.4103891712 & 43273.3338758884 \end{bmatrix}$$

which yields a correlation of -0.87.

Appendix

Suppose we have three variables: X, Z, and Y. Further suppose that missingness occurs due to Z and that a non-zero correlation exists between $Z \times X$ and Y (i.e., an XZ interaction is present).

In order to obtain an unbiased estimate of the population variance/covariance matrix for the selected and unselected variables (i.e., Z, X, ZX, and Y), we may supply a population covariance matrix between X, Z, and ZX. Let us assume the researcher has access to population variances of Z. Under that situation, the X/Z variance/covariance matrix can be obtained using the standard Case II correction. However, the third row/column covariance matrix is still missing (i.e., the variance and covariances with ZX). To obtain the covariance between X and XZ, we observe

I haven't entirely figured out why this is. That's just what the simulation tells me. If selection occurs on Z, I can see why xz would be needed, but why x?

$$cov(X, XZ) = E(XXZ) - E(X)E(XZ) \tag{6}$$

Recall

$$E(XXZ) = \int\limits_X \int\limits_Z x^2 z f(XZ) dz dx$$

And

$$f(XZ) = f(X|Z)f(Z)$$

So

$$\begin{aligned} E(XXZ) &= \int\limits_X \int\limits_Z x^2 z f(X|Z) f(Z) dz dx \\ &= \int\limits_Z Z f(Z) \int\limits_X x^2 f(X|Z) dz dx \end{aligned} \tag{7}$$

Recall

$$E(X^2|Z) = \int\limits_Z X^2 f(X|Z)dz = V(X|Z) + E(X|Z)^2$$

Therefore

$$E(XXZ) = \int\limits_Z Zf(Z)[V(X|Z) + E(X|Z)^2]dz$$
$$E(XXZ) = \int\limits_Z Z[V(X|Z)]f(Z)dz + \int\limits_Z Z[E(X|Z)^2]f(Z)dz$$

Recall that, under normality, $v(X|Z)$ is constant and its value is $v(e)$. Therefore

$$\begin{aligned} E(XXZ) = &\int\limits_Z Z[V(X|Z)]f(Z)dz + \int\limits_Z Z[E(X|Z)^2]f(Z)dZ \\ &= v(e)E(Z) + \int\limits_Z Z(\beta_0 + \beta_1 Z)^2 f(Z)dZ \\ &= v(e)E(Z) + \beta_0^2 E(Z) + 2\beta_0\beta_1 E(Z^2) + \beta_1^2 E(Z^3) \end{aligned}$$

Pluggin back in to Equation 8, we get

$$cov(X, XZ) = v(e)E(Z) + \beta_0^2 E(Z) + 2\beta_0\beta_1 E(Z^2) + \beta_1^2 E(Z^3) - (\beta_0 + \beta_1\bar{Z})[E(XZ)] \quad (8)$$

An alternative;

According to West and Aiken, then the covariance between ZX and X under symmetry is

Cov(xz,x) = V(x)*Mz + cov(x,z)*Mx.

All of which can be estimated from sample data/population values of z.

Author Note

We wish to thank the editor and our anonymous referees for their insightful feedback during the preparation of this manuscript. Their suggestions have helped clarify the purpose, results, and impact of our manuscript.

Table 1

Simulated Dataset of SAT, High School GPA, and First Year GPA Scores. Missing Cells Indicate Those Scores that Fall Below the Median of SAT.

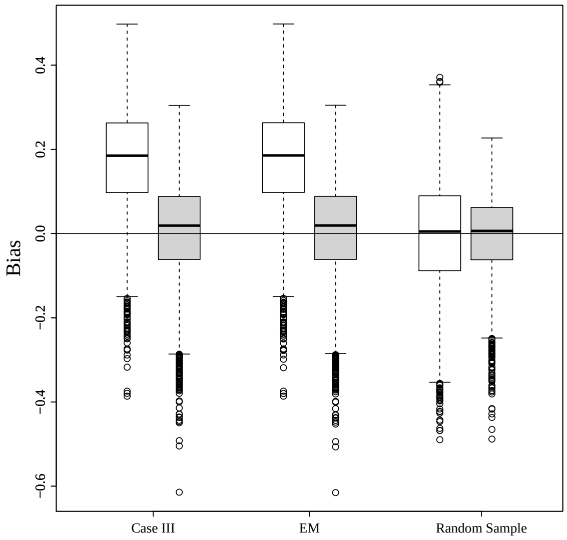
SAT	HSGPA	SAT × HSGPA	First Year GPA
565.51	3.02	9.11	3.17
612.75	2.85	8.13	2.87
527.39			
556.31	3.06	9.35	3.56
359.03			
660.74	3.74	13.98	3.23
578.58	2.98	8.89	2.76
649.78	2.71	7.36	2.41
565.17	2.89	8.33	3.28
316.50			
440.26			
558.56	3.43	11.74	3.55
455.61			
391.18			
460.10			
336.51			
536.26	2.86	8.17	3.08
444.54			
452.57			
532.63	2.85	8.14	3.15

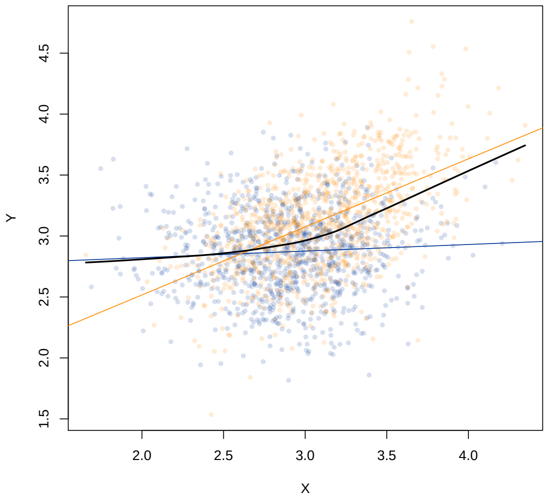
	SAT	GPA	c(covz.xz, covx.xz, var.xz)
SAT	10000.00	4.26	35120.36
GPA	4.26	0.10	110.72
3	35120.36	110.72	162427.03

Figure Captions

Figure 1. Boxplots showing the distribution of two correction procedures (Case III and the EM algorithm) relative to a random sample. The distribution is across 10,000 iterations, each with a net sample size of 50 after selection

Figure 2. Boxplots showing the distribution of two correction procedures (Case III and the EM algorithm) relative to a random sample. The distribution is across 10,000 iterations, each with a net sample size of 50 after selection


Bias
0.4
0.2
0.0
−0.2
−0.4
−0.6
Case III
EM
Random Sample


Y
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.0
2.5
3.0
3.5
4.0
X