

On the Use of Partial Covariances in Structural Equation Modeling

Thomas D. Fletcher

University of Missouri – St. Louis

fletcher@umsl.edu

www.umsl.edu/~fletcher

Lisa M. Germano & Katherine A. Selgrade

Old Dominion University

Note. This paper is currently under revision to be submitted for publication. Please check with first author for proper reference. As presented, the reference is:

Fletcher, T. D., Selgrade, K. A., & Germano, L. M. (2006, May). On the use of partial covariances in structural equation modeling. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

ABSTRACT

The use of control variables is more common in regression than in structural equation modeling (SEM). There are theoretical and methodological reasons for this divide. Two simulations were conducted to assess the effects of using partial covariances in SEM. Results indicate structural equation models can be simplified under certain conditions.

PRESS PARAGRAPH

Researchers should always seek to use the research method and statistical analysis that best addresses their research question. However, we found a tendency to overuse one procedure, regression, as opposed to another, structural equation modeling, when analysts need to control for a set of variables. We conducted two simulation studies to determine if an often overlooked method is viable in the use of statistical controls. Partial covariances are useful in simplifying structural equation models. The results of this study could lead to an increase in the use of statistical controls with the more sophisticated structural equation modeling.

Hierarchical regression is common in the social and behavioral sciences (Cohen, Cohen, West, & Aiken, 2003). The basic steps of this procedure involves including one or variables as a set at one step and then adding one or more variables to the model as another step to determine the incremental prediction of the second set controlling for the influences of the first set. Often, the influence of the first set of variables on the outcome variable is not of particular interest to the researcher given the specific model. The use of statistical control in addition to or in lieu of experimental control is well known (Winer, Brown, & Michels, 1991). Control variables or covariates serve to hold constant the influences of variables not otherwise explicitly modeled. Linn and Werts (1969) described the use of partial regression coefficients in making causal inferences more than three decades ago. However, such practices appear to have been limited to regression analyses with directly observed variables rather than structural equation models (SEM), which utilize latent variables with manifest indicators.

To substantiate the claim that the explicit use of control variables is less common in SEM than regression, we reviewed two years (2003 and 2004) of the *Journal of Applied Psychology*. We were primarily interested in how frequently control variables were used in regression analyses and in SEM. In 2003, 91 articles were published in the *Journal of Applied Psychology*. Of these 91 articles, 43 reported regression analyses. Twenty-six or 60% of the 43 articles reporting regression analyses used control variables. Also in 2003, 13 of the 91 published articles reported SEM results. Five articles or 38% of the articles reporting SEM results also used control variables in their models. Likewise, in 2004, 86 articles were published in the *Journal of Applied Psychology*. Regression analyses were reported in 36 of the 86 articles. Control variables were included in the regression analyses in 18 or 50% of these 36 articles. Lastly, 12 of the 86 articles published in 2004 report SEM results. Three or 25% of these articles report including control

variables in their models. Totaling these across the two years yields, 44 out of 79 (56%) articles involving regression utilized control variables whereas 8 out of 25 (32%) articles involving SEM utilized control variables. The chi-square test of independence indicates this difference is significant ($\chi^2(1) = 4.26, p < .05$).

It is not clear whether researchers use control variables with less frequency because of (1) theoretical reasons such as the model to be tested does not require control variables, or (2) methodological reasons such as the difficulty in modeling variables that contribute little to the overall model. It is possible however to have it both ways. That is, to use SEM and to control for extraneous sources of variance without creating too complex a model. By first creating partial correlations (or partial covariances), researchers are able to control for covariates (extraneous sources of variance) and test the core model. This procedure has appeared in the literature previously (see Newcomb & Bentler, 1988) and more recently (Kammeyer-Mueller & Wanberg, 2003). In this paper, we address the use of partial covariances, the pros and cons of their usage, and demonstrate with two simulation studies the appropriateness of their use.

Partial Regression Coefficients

Consider the following model of Y regressed onto X controlling for Z , where b and c are the regression coefficients and e is the residual error term for the model.

$$Y = bX + cZ + e \quad (1)$$

We can represent our model controlling for Z in the following manner. Consider Model 2 which contains X and Y each regressed onto Z . The error terms, or residuals in this model are denoted $X_{.z}$ and $Y_{.z}$ respectively. The coefficients d and g are of little importance in the present context.

$$\begin{aligned} X &= dZ + X_{.z} \\ Y &= gZ + Y_{.z} \end{aligned} \quad (2)$$

Now consider an equivalent representation of Model 1.

$$Y_{.z} = b_{.z}X_{.z} + e \quad (3)$$

From Model 1 and Model 3, $b = b_{.z}$. This relationship among partial regression coefficients is well known in the multiple regression literature (Linn & Werts, 1969). Of course, we do not know the explicit effect of Z on Y , c , in the second representation. This analysis of residuals is equivalent to first creating partial covariances (or partial correlations for standardized variables) and then conducting the regression analysis.

The extent to which this relationship holds is not known when using latent variables measured with multiple indicators. For example, suppose the covariate, Z , is directly observable, but Y and X have multiple indicators. The multiple indicators regressed onto Z may influence the structural coefficients in some unknown way. The purpose of this paper is to explore this possibility. We aim to determine if using partial covariances introduces bias into the structural equation model in any discernable way. We do this with two simulations, which are described below.

Pros and Cons of Analyzing Partial Covariances

The arguments for or against the use of partial covariances is not about the use of control variables per se. We believe the particular model to be tested should be guided by theory. Some models will require the use of control variables, others will not. In the cases where control variables are required, the issue is whether or not to explicitly model the controls or to statistically partial out their effects from the model variables and then test the model. There are several arguments for the use of partial covariances. Likewise there are a number of cautions against their use. We will elucidate many of those arguments here.

To partial out the controls rather than to explicitly model them, one first has to determine that they in fact do not need to know the relationship of the control to the model system – specific structural coefficients will be unobtainable if not explicitly modeled. When the use of controls complicates a model, using partial covariances will free up degrees of freedom and simplify the model. For example, Newcomb and Bentler (1988) had a core model with 10 latent variables including 32 manifest indicators. Adding 13 control variables to the already complex model would have made the model untenable given the sample size. Likewise, Kammeyer-Mueller and Wanberg (2003) controlled for seven variables, but many of them were dummy-coded with multiple categories, which would have added greatly to model complexity. One alternative is of course to resort to regression, but then one loses the benefits of SEM in the first place. Namely, regression does not account for measurement error and regression is quite cumbersome with complex models involving indirect paths.

While partialling out the effects of covariates rather than explicitly modelling them may well reduce model complexity, the degree to which the structural coefficients from the two practices are equivalent is not obvious. Further, even if the structural coefficients are equivalent, their standard errors may not be. These simulations collectively address this issue. To further complicate matters, the statistical test for the structural coefficients – coefficient/standard error – is based on a t distribution with degrees of freedom different in the two approaches. However, this is of less concern when sample sizes are at least 150 or so. At larger sample sizes the statistical test is based on a z distribution. Another concern is the possible introduction of measurement error from the controls into the model indicators. The measurement model may not be able to parse this additional variance leading to imprecise measurement of the model latent variables. Newcomb and Bentler (1988) used 11 variables measured with error as controls

(coefficient α s ranged .74 to .87). Simulation 2 addresses the possibility that measurement error of the controls could influence the results of the substantive questions.

Yet another issue is the concern of multivariate normality. Under certain conditions, the partialled variables may be closer to multivariate normality (or less so) than the original variables. Depending on how the partial covariances are created, one may or may not be able to assess the normality of the indicators. Great care should be taken with respect to measurement model assumptions when using partial covariances.

Brief Description of Simulations

We aim to conduct two simulation studies that will address the concerns listed above. The first simulation involves a model where the covariates are assumed to contain no measurement error. Examples of such covariates commonly found in research include age, gender, education and the like. Figure 1a depicts the simulated model. This model will be examined by explicitly modeling the covariates in comparison with analyzing the partial covariances having controlled for the covariates. A second simulation study will address the concerns of covariates that are themselves latent variables with multiple indicators. Examples of such covariates are limitless depending on the theoretical model being tested (e.g, social desirability, perceived workload, personality). Figure 1b depicts the simulated model. A complication in simulation 2 is that of how best to create the partial covariances. Method A involves regressing the model indicators on the covariate indicators and saving the residual variance/covariance matrix as in simulation 1 – see Figure 2a. Method B involves first creating a latent variable for the covariates and then regressing each of the model indicators on the latent covariates – see Figure 2b. The simulation studies will determine the extent to which (1)

analyzing partial covariances is a viable alternative to explicitly modeling covariates, and (2) if method of creating the partial covariances influences the outcome.

Method

Data Generation

Data were generated for both simulations using PRELIS 2.54 (Jöreskog & Sörbom, 1996b). All variables are created based on normal and continuous data. For both simulations, 500 replications were conducted with each having a sample size of 200. Figure 1a and 1b depicts the full simulated model with parameter specifications for Simulation 1 and Simulation 2.

Analysis of Models

Simulation 1 involves analyzing the simulated data over 500 replications by explicitly modeling the covariates and analyzing the partial covariances controlling for the covariates. Partial covariances are created by regressing the model indicators ($Y_1 - Z_3$) onto the covariates (C1 and C2) and then saving the residual covariances (i.e., freely estimated Ψ [psi] matrix in LISREL terminology). We wish this model to remain simple and have no reason to expect manipulating other factors such as sample size will impact the estimation of the structural coefficients. Likewise, Simulation 2 involves analyzing the data over 500 replications by explicitly modeling the latent covariates and analyzing the partial covariances. This simulation will address whether method of creating the partial covariances when the covariates are latent variables with multiple indicators makes a difference. That is, Method A consists of creating the partial covariances by regressing model indicators on covariate indicators ignoring the latent nature of the covariates and saving the freely estimated Ψ matrix – see Figure 2a. In Method B, we create latent variables for the covariates and then regress the model indicators (i.e., $Y_1 - Z_3$) onto the latent covariates saving the residuals for analysis – see Figure 2b. In LISREL terminology, Method B involves two latent variables, ξ (ksi) with arrows pointing to all

indicators (i.e., $Y_1 - C2_3$). The measurement error variance/covariance matrix, θ_δ (theta delta), is freely estimated for all model indicators, but fixed off diagonal for the covariate indicators. It is θ_δ in Method B that is saved for further analysis. For each of the 500 replications and for each of the analytic methods in simulation 1 and simulation 2, the model is assessed and the structural coefficients (γ , gamma) corresponding to X and Z and their corresponding standard errors are saved for examination.

Results

Simulation 1

Simulation 1 involved the analysis of two structural models:

$$Y = X + Z + C1 + C2 + \zeta \quad (4)$$

$$Y_{.C1C2} = X_{.C1C2} + Z_{.C1C2} + \zeta \quad (5)$$

Equation 4 explicitly models the covariates $C1$ and $C2$ and Equation 5 depicts the analysis of the partial covariances controlling for $C1$ and $C2$. In either case, it is not known the extent to which the structural coefficients corresponding to X and $X_{.C1C2}$ (and Z and $Z_{.C1C2}$) are equivalent. Table 1 shows the average of these coefficients over the 500 replications. The true simulated value for X was .25 and the true simulated value for Z was -.35 (see Figure 1a). Any observed bias in these models is likely due to sampling error; however, given the 500 replications, sampling error should be minimal. The bias, or difference in the average estimated parameter and its simulated population value, is very similar for the explicit method and the partial covariance method. Differences in these two methods can likely be attributed to precision and rounding. That is, when saving the partial covariances, LISREL rounds to six decimal places where, during explicit modeling, no such rounding occurs. The difference in the parameter estimates for the two methods in Simulation 1 is .001 for X and .002 for Z .

Explicitly modeling the covariates is also very similar to analyzing the partial covariances with respect to variation in estimates. Table 1 shows the average estimated standard errors, the empirical standard error, and the standard error ratios. The difference in the average estimated standard errors between the explicit and partial covariance methods is .0002 for X and .0003 for Z . The empirical standard error is the standard deviation of the 500 replications and represents the expected standard error for each replication. The standard error ratio is the ratio of the average estimated standard error to the empirical standard error. The values are all below 1.0 reflecting slight underestimation of the standard error. The standard error ratios are quite similar for the explicit method and the partial covariance method. These differences are small and may be considered trivial.

Simulation 2

Simulation 2 involved the analysis of the same structural models as depicted in Simulation 1, Equations 5 and 6. The exception is in the measurement of the covariates C1 and C2. For Simulation 2, the covariates are each measured with three indicators (see Figure 1b). A second difference in Simulation 1 and 2 is that there are at least two methods for creating the partial covariances when the covariates are measured with multiple indicators. These methods are depicted in Figure 2a and 2b respectively and described above.

Table 2 shows the results of the 500 replications for Simulation 2. The structural coefficients simulated were the same as in Simulation 1; $X = .25$ and $Z = -.35$. The bias in the estimates is much more similar for the explicit method and Method B for the partial covariance methods, than the explicit method and Method A. Figure 3 is the scatterplot matrix of the structural coefficients for each of the possible combinations of methods of estimation across the 500 replications. For X , the explicit method is correlated with the partial covariance method A, r

= .989 (top left panel) whereas the explicit method is correlated with the partial covariance method B, $r = .999$ (middle left panel). This difference in correlations is significant, $z = 18.9$. Nearly identical results are found with the Z variable (right panels of Figure 3).

With respect to variability in the estimates, Table 2 shows the average estimated standard error, the standard deviation of the estimates, and the standard error ratio. Again, the explicit method is more similar in estimation with the partial covariance method B than with method A. Figure 4 shows these similarities in distributions graphically. The plots are the densities for the estimates of the structural coefficients for the 500 replications. The density of the explicit method results are plotted with a solid line whereas the density of the partial covariance methods results are plotted over the explicit method with circles. Consistent with the differences in bias, the partial covariance method A distribution for X and Z are slightly shifted to the right (circles in density plot in upper panels of Figure 4). The lower panels of Figure 4 show the nearly identical distribution of estimates for the explicit method (solid line in density plot) and the partial covariance method B (circles in density plot). The relative efficiency of the partial covariance methods to the explicit modeling of the covariates was calculated by the ratio of the root mean square errors (RMSE; Mooney, 1997). The RMSE for the explicit method was .00441, the RMSE for the partial covariance method A was .00446, and the RMSE for the partial covariance method B was .00439. All three methods produce strikingly similar results. The relative efficiency for Method B was .996 (.4% more efficient than the explicit method). The relative efficiency for Method A was 1.01 (1% less efficient than the explicit method). Practically speaking, these differences are not large enough to warrant great attention. However, Method B does emerge as the preferred method for using partial covariances when covariates have multiple indicators.

Discussion

In the present study, we raised the questions, (1) to what extent might analyzing partial covariances be a viable alternative to explicitly modeling covariates, and (2) to what extent might the method of creating the partial covariances impact the estimates of the structural parameters? With regression we know that explicitly modeling covariates is equivalent to analyzing partial covariances (i.e., having removed the effects of the covariates). However, it was unclear if partialling the variance of covariates from a model's manifest indicators leads to equivalent structural coefficients. Simulation 1 demonstrated that the differences are minute – if any. When the covariates are measured directly with no error, partialling their effects from the indicators has little effect on the structural equation model. This provides an alternative for modeling variables that can greatly simplify structural equation models.

However, when the covariates themselves are latent variables, there are at least two methods for creating the partial covariances. That is, to partial the effects of the covariates from the model indicators, one can either (A) partial the effects of covariate indicators from the model indicators, or (B) create a latent covariate and then partial the effects of the covariates from the model indicators. There were small but noticeable differences in these methods. Method A was less similar to explicitly modeling the covariates than was method B. When measurement error is present in the covariates, one should be careful as to not introduce that error into the model indicators. The present simulations demonstrated that these differences were practically small, but present.

One thing we will note, in the simple model that we simulated based on equation 4, X had a positive influence on Y whereas Z had a negative influence on Y . The degree of bias in the estimates as well as the underestimation of the standard errors tended to be larger for the negative relationship (Z) than for the positive relationship (X). The bias is nearly twice as large

for Z than for X (.0102 vs. -.0191) in Simulation 1. Likewise, the standard error ratio is nearly 1.0 (.9970) for X but is substantially less than 1.0 (.9262) for Z . A caution in the interpretation of this result is that the magnitude of Z was simulated to be larger than X . The differences in bias and variability in estimation may be due to sign differences or simply due to magnitude differences (i.e., a percentage of original parameter).

As with all simulation research, the results are limited to the parameters of the simulated variables. In the present study, we chose to simulate two rather simple models. We did not vary factors such as sample size, model complexity, reliability of the latent variables, degree of correlation among the exogenous variables and the controls or endogenous and controls, or the model squared multiple correlation. While we feel these factors were not necessary to address the questions in this study, future research could nonetheless manipulate them to determine the degree to which each of these factors might impact the use of partial covariances.

These simulations demonstrate that one can in fact use structural equation modeling and still statistically control for variables without explicitly adding them to the model. Many theories require the use of statistical controls, especially when those data were not collected under experimental conditions as in applied research. We reiterate three cautions when using partial covariances. First, the researcher should not be interested in the direct effects of the covariates to any of the model variables, because their direct influence would not be known if using partial covariances. Second, researchers should make clear the interpretation of the structural parameters when the effects of covariates are partialled. Third, researchers should not lose sight of the measurement model and the influences of covariates. When the covariates themselves have multiple indicators, one should be careful to not reintroduce measurement error into the model indicators.

References

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed.* Mahwah, NJ: Lawrence Erlbaum Associates.
- Jöreskog & Sörbom (1996a). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.
- Jöreskog & Sörbom (1996b). *PRELIS 2: User's reference guide.* Chicago: Scientific Software International.
- Kammeyer-Mueller, J. D. & Wanberg, C. R. (2003). Unwrapping the organizational entry process: Disentangling multiple antecedents and their pathways to adjustment. *Journal of Applied Psychology, 88*, 779-794.
- Linn, R. L. & Werts, C. E. (1969). Assumptions in making causal inferences from part correlations, partial correlations, and partial regression coefficients. *Psychological Bulletin, 72*, 307-310.
- Mooney, C. Z. (1997). *Monte Carlo simulation.* Thousand Oaks: Sage Publications.
- Newcomb, M. D. & Bentler, P. M. (1988). Impact of adolescent drug use and social support on problems of young adults: A longitudinal study. *Journal of Abnormal Psychology, 97*, 64-75.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design (3rd ed.)*. Boston: McGraw-Hill.

Table 1

Summary of Simulation 1 Results

	Explicit Covariate Model		Partial Covariance Model	
	X	Z	X	Z
Average of Estimates	.2616	-.3714	.2602	-.3690
Bias	.0116	-.0214	.0102	-.0191
Average Standard Error	.0993	.1016	.0990	.1012
Standard Deviation of Estimates	.0995	.1100	.0993	.1093
Standard Error Ratio	.9975	.9232	.9970	.9262

Note. For all models, N = 200, number of replications = 500. The ‘true’ parameter value for X is .25 and for Z is -.35. The standard error ratio is the average standard error computed from LISREL divided by the empirical standard error (standard deviation of the distribution of parameter estimates).

Table 2

Summary of Simulation 2 Results

	Explicit Covariate Model		Partial Covariance Method A		Partial Covariance Method B	
	X	Z	X	Z	X	Z
Average of Estimates	.2623	-.3728	.2662	-.3676	.2613	-.3722
Bias	.0123	-.0228	.0162	-.0176	.0113	-.0222
Average Standard Error	.0956	.0974	.0949	.0967	.0947	.0967
Standard Deviation of Estimates	.0980	.0980	.0984	.9980	.0976	.0981
Standard Error Ratio	.9755	.9937	.9641	.9691	.9699	.9857

Note. Partial covariance method A is a model in which the covariates are partialled out using only the manifest indicators. Partial covariance method B is a model in which the latent covariates are partialled out of the manifest indicators for the model variables.

Figures

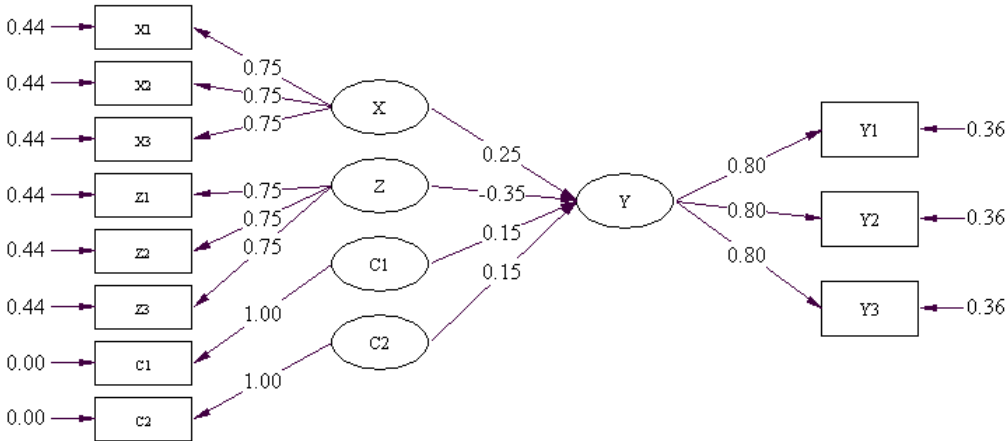


Figure 1a. Model parameters simulated for Simulation 1.

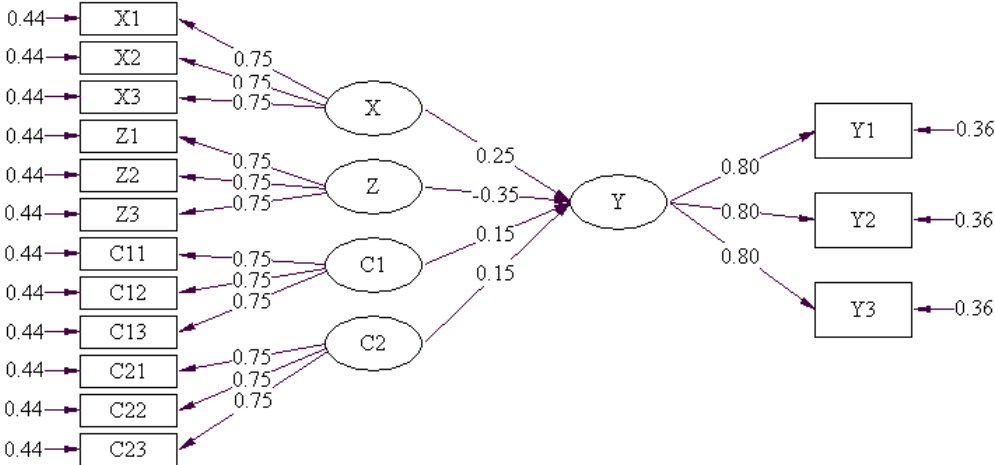


Figure 1b. Model parameters simulated for Simulation 2.

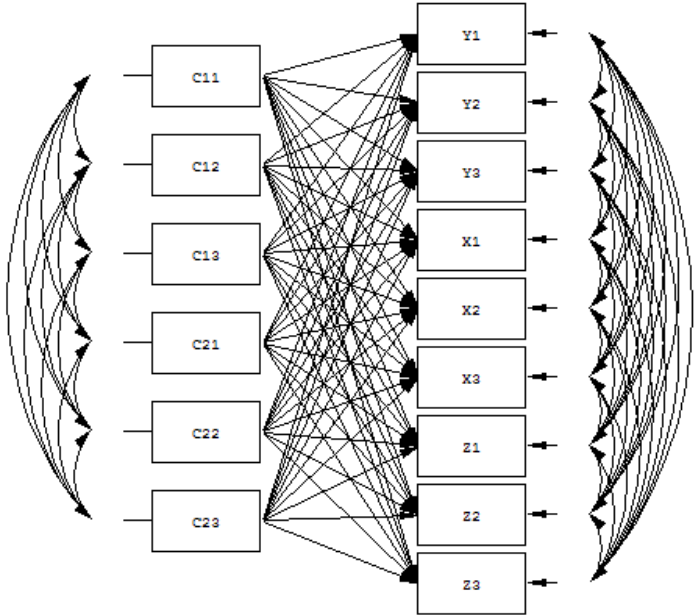


Figure 2a. Method A for Partialing out covariates measured with multiple indicators.

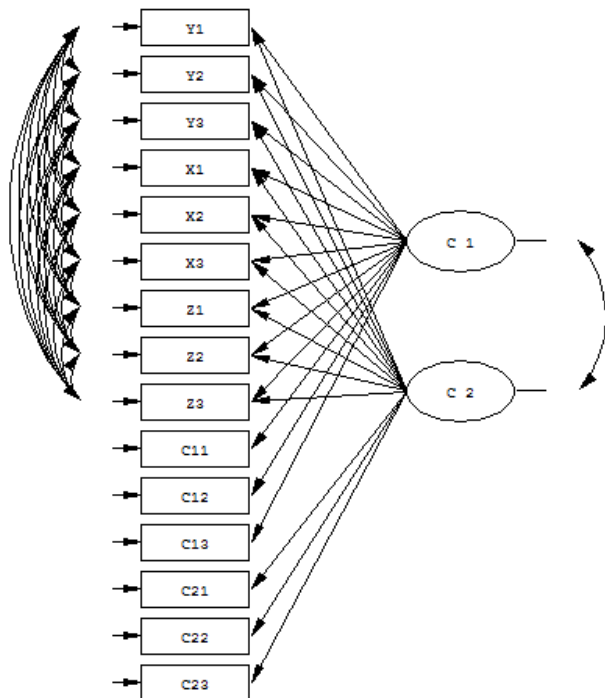


Figure 2b. Method B for Partialing out covariates measured with multiple indicators.

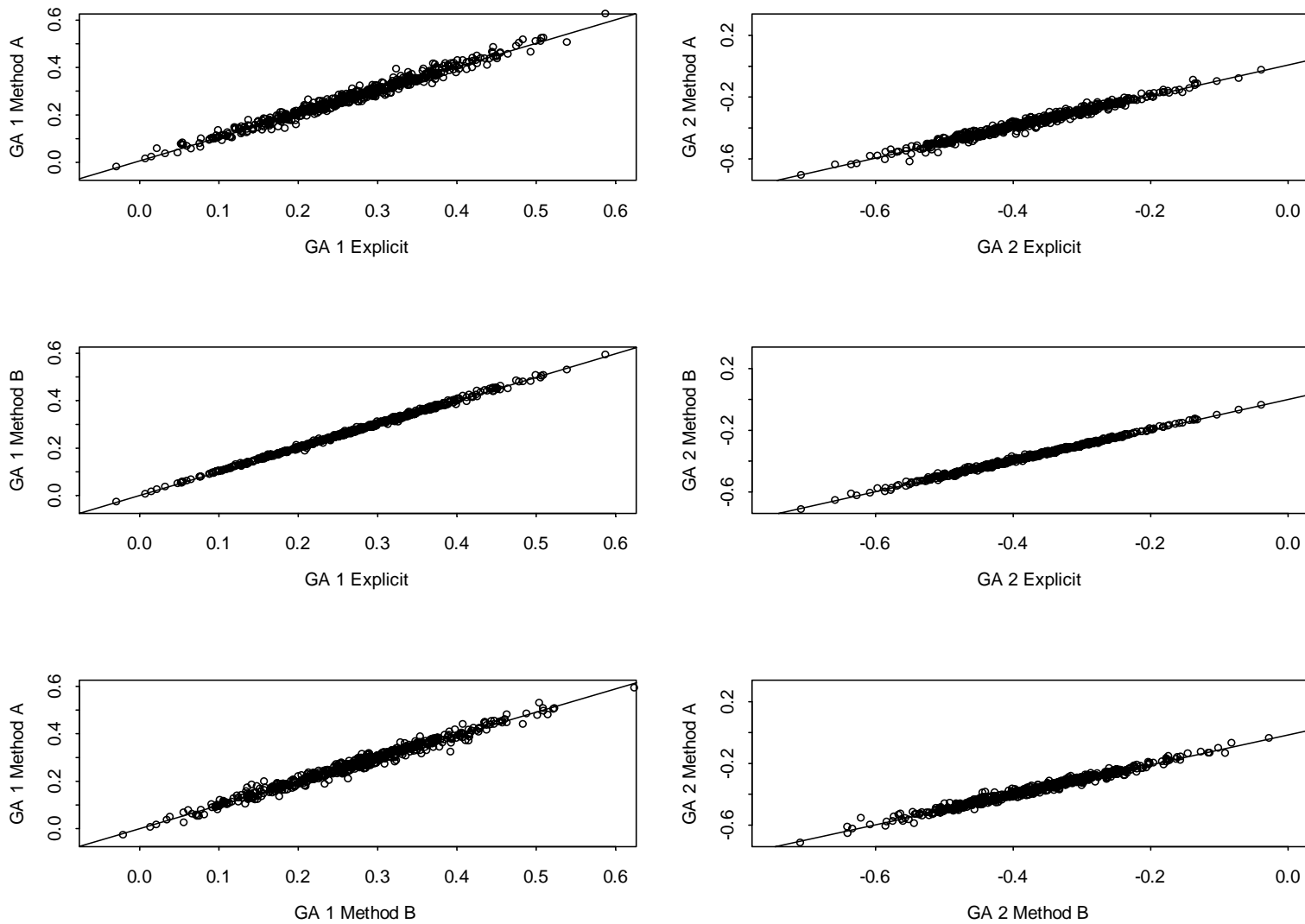


Figure 3. Scatterplots with regression lines for parameter estimates from Simulation 2.

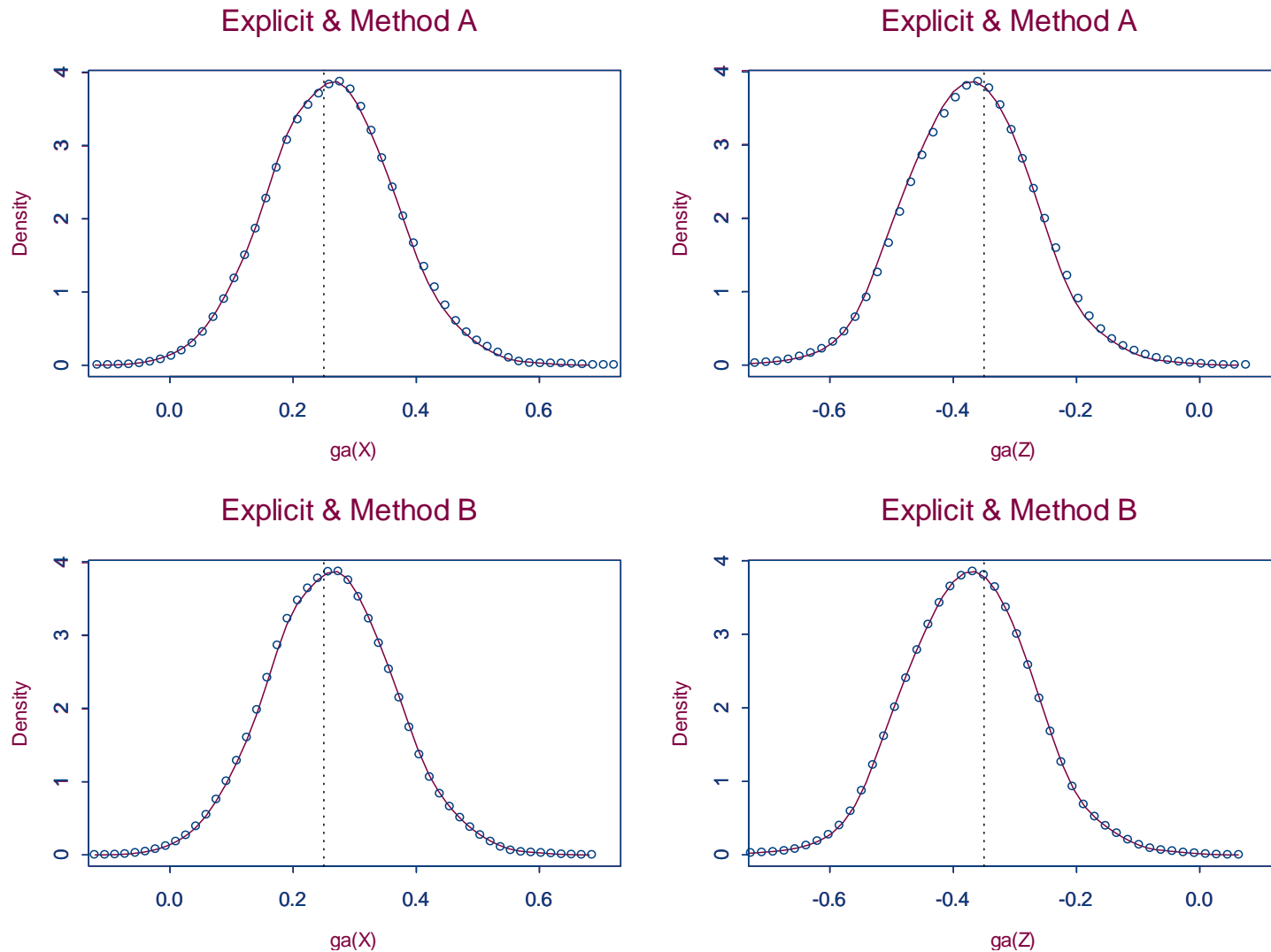


Figure 4. Distributions of parameter estimates from Simulation 2. Vertical dotted line is the population parameter estimate for X and Z respectively. The partial covariance methods are depicted with circles and the explicit method is depicted with a line.