On the Use of Control Variables in PLS-SEM

Sull'Uso delle Variabili di Controllo nei PLS-SEM

Francesca De Battisti and Elena Siletti

Abstract Several authors have recently devoted more attention to the control variables methodological issue. Despite many recommendations to handle these variables more efficiently, good practices are still widely disregarded, and especially this topic has not yet been studied in depth for structural equation models. This paper suggests best research practices for researchers who deal with the use of control variables in partial least squares structural equation models.

Abstract *Recentemente diversi autori hanno dedicato attenzione al problema delle variabili di controllo. Nonostante le molte raccomandazioni suggerite per gestire queste variabili in modo piú efficiente, le buone pratiche sono ancora ampiamente ignorate, in particolare questo argomento non é stato ancora approfondito nei modelli ad equazioni strutturali. Questo lavoro propone alcune linee guida a chi deve utilizzare le variabili di controllo nei modelli ad equazioni strutturali con stima dei minimi quadrati parziali.*

Key words: Control variables, Structural equation models, Partial least squares

1 Introduction

Control variables are traditionally considered in causal models to rule out alternative explanations for findings or to reduce error terms and increase statistical power. They are variables that do not change and that can cause or be correlated with the causal variable, mediator and outcome. There are two primary means of controlling variables. The first is control by experimental design, whereby the researcher manipulates the nature of the sample. The second is statistical control, whereby

De Battisti Francesca and Siletti Elena
Department of Economics, Management and Quantitative Methods,
Universita' degli Studi di Milano, Via Conservatorio, 7 - 20122 Milan, Italy;
e-mail: francesca.debattisti@unimi.it, elena.siletti@unimi.it

the researcher measures relevant variables and includes them in the analysis. This approach mathematically partials the effect of the control from the other variables (Becker, 2005). There is no widespread agreement to handling statistical controls (Spector et al, 2000; Breaugh, 2008). Moreover, Spector and Brannick (2011) warn against the misuse of demographic variables because attention should be focused on the mechanisms that explain relations with demographics rather than on the demographics themselves. Becker et al (2016) recently recommend to handle controls more efficiently, paying attention to variables selection, to the methods for measuring and analysing them, to reporting and interpreting results. Becker (2005) considers the controls in structural equation models (SEM), underlined that this issue deserves further attention, and that authors using SEM need to explain why they are treating control variables as they are.

2 Guidelines on the Use of Control Variables in PLS-SEM

SEM allow to study the relationships among latent, not directly observable, variables. Two types of SEM can be distinguished: covariance- and variance-based models. In variance-based models, linear combinations of observed variables are first created as proxies, and then the parameters are estimated by them. Among variance-based SEM methods, partial least squares (PLS) path modelling (Wold, 1985; Lohmöller, 1989) has been called a "silver bullet" (Hair et al, 2011); it is recently used intensively and it will be discussed below. PLS is an iterative algorithm to estimate the different blocks of the measurement model separately and then, in a second step, to estimate the coefficients of the structural model; the aim is to explain the best residual variance of the latent variables and, potentially, of the manifest variables. PLS-SEM consider a sequence of equations that describe the relationships among key theoretical constructs (i.e. the structural model) and a sequence of equations that show the relations among the latent and manifest variables (i.e. the measure model). The presence of a measurement model alone represents a confirmatory factor analysis; the case of a structural model alone represents a path analysis on observed variables. Since the analysis of controls involves the presence of causal links, in the following discussion our attention will be devoted to the structural model alone, and a report of the possible relationships involving controls is presented. To simplify, the usual multiple regression notation is adopted: X is an independent (predictor/exogenous) variable, Y is a dependent (criterion/endogenous) variable, M is a mediator variable, m is a moderator variable, and C is a control. We remember that all these variables are constructs or latent variables, each of which is related to one or more manifest or measured variables. The simplest causal model has two variables: X and Y; and it can be represented mathematically with a single equation, or with the correspondent path diagram. When we introduce a control, in the system there is only one equation, but with one more variable on the right. Even with only three variables, the scenario is more complex. The third variable could be independent or dependent, a mediator, or a moderator. These options are represented by Eq. 1, 2, 3 and 4, respectively.

$$Y = \beta_0 + \beta_{X_1 Y} X_1 + \beta_{X_2 Y} X_2 + \varepsilon_Y \tag{1}$$

$$Y_1 = \beta_{0_1} + \beta_{XY_1} X + \varepsilon_{Y_1} \qquad Y_2 = \beta_{0_2} + \beta_{XY_2} X + \varepsilon_{Y_2} \tag{2}$$

$$Y = \beta_{0_1} + \beta^{'}_{XY} X + \varepsilon_{Y'}; \quad M = \beta_{0_2} + \beta_{XM} X + \varepsilon_M; \quad Y = \beta_{0_3} + \beta_{XY} X + \beta_{MY} M + \varepsilon_Y \tag{3}$$

$$Y = \beta_0 + \beta_{XY} X + \beta_{mY} m + \beta_{m*X} mX + \varepsilon_Y \tag{4}$$

With two independent variables the model is represented by only one equation (Eq.1). With two dependent variables we have a system with two equations (Eq.2), and the control can be introduced in three different ways: only in the first, in the second, or in both equations. Dealing with a mediation, we refer to the mathematical representation proposed by Judd and Kenny (1981), as the system in Eq.3, where in the first equation, $\beta^{'}_{XY}$, represents the total effect of variable X on Y; in the second equation, β_{XM} is the effect of X on mediator M; in the last equation, β_{XY} is the direct effect of X on Y, and β_{MY} is the effect of M on Y. Notably, with ordinary least squares (OLS) regression the first and third equations are fitted as separate regression models, but in PLS-SEM they are fitted simultaneously. Based on the coefficients, the indirect effect can be computed as the product of the β_{XM} and β_{MY} paths or as the difference between $\beta^{'}_{XY}$ and β_{XY} (i.e. $\beta^{'}_{XY} - \beta_{XY}$). Furthermore, the proportion mediated can be calculated as $\frac{\beta_{XM}\beta_{MY}}{(\beta_{XM}\beta_{MY}+\beta_{XY})}$, $\frac{\beta_{XM}\beta_{MY}}{\beta^{'}_{XY}}$ or $1 - (\frac{\beta_{XY}}{\beta^{'}_{XY}})$. Traditionally, we refer to this kind of link as *partial mediation*, while if the direct effect, β_{XY}, is not significant the mediation is said to be *full*. In mediation analysis, the latter case, in which the total effect, $\beta^{'}_{XY}$, is equal to the indirect effect, $\beta_{XM}\beta_{MY}$, is the most interesting because the link between the dependent and independent variables is significant only through the mediator. In this kind of model, the controls can be expressed in different ways: on only the dependent variable Y, on only the mediator variable M or on both of the variables. The way to deal with controls does not change for partial or full mediation models. With a moderator (Eq.4), we need to take its nature into account. Baron and Kenny (1986) defined four cases of moderation by predictor and moderator scales. In the context of PLS-SEM, the predictor is latent; for this reason, only two cases must be considered, where moderator effects are indicated by the interaction of X and m in explaining Y and are measured by β_{m*X}. With continuous moderators, the global effect of X on Y is, therefore, defined as the sum of the simple effect, β_{XY}, and the product, $\beta_{m*X}m$, since the effect of X on Y depends on the value of m (the product term approach (Chin et al, 2003), see Eq.4); the corresponding model with a control is represented as:

$$Y = \beta_0 + \beta_{XY} X + \beta_{mY} m + \beta_{m*X} mX + \beta_{CY} C + \varepsilon_Y \tag{5}$$

If the moderator is non-metric, or if we categorize a quantitative variable, we apply the multi-group analysis (MGA; Henseler et al (2009)). Following this method, the moderator effect is measured by replicating the analysis on subgroups that differ by moderator level. When we conduct an MGA, if there is a moderator with *k* levels, this technique considers *k* models. Increasing the number of variables, the number of links among them also increases and their nature differs, yielding greater complexity. As we discussed for moderators, we also need to reflect on the nature of controls. Considering the case of only one control, if it is continuous, we simply introduce a new latent variable with the control as a unique indicator. However, this single-item approach is not free of criticism (Diamantopoulos et al, 2012). When the control is categorical, we can adopt the MGA using the levels to generate groups. Then, the overall model is compared in subgroups, and all the relationships are involved in the comparison, making it impossible to set the control only on a single link. The use of the MGA relies on the sample size; also when using PLS-SEM, which is applicable to a small sample size, a fitted issue could occur when we split in many subgroups, and it increases with a moderator. Considering Eq.2, the previously possible proposed solutions refer only to the single-item approach, which examines only one link at a time. In contrast, applying the MGA disables the choice of a solution since the MGA offers a complete picture of the control's influence on the system, considering the impact of the control on all the links. When theoretical assumptions require assessment with several controls, more than one approach is available. We can deal with controls separately, in this way we go back to previous cases, or conjointly. In the latter approach, we can alternatively apply the MGA, simultaneous single-item variables or one latent variable approach (also called a multi-item approach). With the MGA, we compare several subgroups identified by different mixtures of levels, considering the worsened issue of sub-sample dimensions discussed above. With the simultaneous single-item approach, we input as many latent variables as controls in the same model, with consideration to the worsened issue of single-item measure. Using single-item variables simultaneously or separately, as above, implies the same differences as in regression models when considering a unique multiple regression with all the controls together or several regressions that differ only by the control. Finally, we can identify a new construct using one latent variable measured by several controls. This approach requires accuracy and a choice between a reflective and formative shape to define the new construct. Either way, it is not easy to imagine a latent variable which summarizes different controls. To better explain the different options described, we introduce an illustrative example with *customer satisfaction* modelled as a mediator between *corporate reputation* and *customer loyalty* (see Eq.3). Supposing to control for *gender* and *age*, two research strategies are possible. In the first, the controls are separately considered. Treating *age* as continuous, a new latent variable (with single-item approach) must be added in the model, only on the dependent variable *loyalty* or on the *satisfaction* too. The same holds for *gender*, if it is handled as dichotomized. Considering *age* or *gender* as categorical, MGA has to be applied; in this case the overall model is controlled. A second research strategy analyses the controls simultaneously. With both the controls categorical, MGA implies to split the dataset in a number of sub-samples equal to the product of their levels (e.g. for three *age* levels and two *gender* levels, six sub-samples are considered). Taking into account *age* as continuous and *gender* as categorical, two sub-samples are created, for female and male, and with *age* treated in the way described above. If *age* and *gender* are expressed as continuous or dichotomized, two different latent variables can be inserted simultaneously in the same model, both of them or alternatively only on *loyalty* or on *satisfaction* too. Finally, the multi-item approach, with an unique latent variable summarizing all the controls, in the specific case of *gender* and *age* is not suitable.

3 Discussion

Guidelines to consider control variables in PLS-SEM could contribute to the literature debate. First of all, it is essential to pay attention to the recommendations of Becker et al (2016) about controls selection: to provide a brief explanation for why each control was selected, to avoid proxies for them, to include them in hypotheses and to describe why and how each covariate was measured in the analysis, noting that a clear description allows for effective replication and extension of the findings. Since in PLS-SEM preliminary analysis we usually evaluate the constructs by applying exploratory factor analysis, it could be helpful to test the correlations between the first factors of the considered constructs with the hypothesised theoretical controls. In this way, an initial choice could be made. After this first step, we can input controls into the model in different ways. In PLS-SEM we have variables that can be simultaneously dependent and independent. Therefore, there is the issue of choosing the variables to which the controls must be related. If we use the MGA, this problem does not arise since we are analysing the model as a whole. This approach allows us to simultaneously consider more than one control, but it could be unrealisable because of the sample size; moreover, the results could be difficult to interpret. As an alternative to the MGA, we can adopt the single- or multi-item approach. With many controls, the single-item approach can be implemented, examining the controls one by one or simultaneously. If several controls are summarized in a latent variable, we have to check the reliability and validity of the construct. In all of these cases, we have to initially choose the relationships between the controls and the other variables; for this purpose, we need to take the theory into account, and we can conduct a comparative test to understand the role of the controls. Their effects can be evaluated by the intensity and significance of the estimated paths. The criterion of parsimony always applies to avoid inserting too many controls, which makes interpretation difficult. Referring to the advice of a report on the shared variance between original and residual predictor variables (Breaugh, 2008), while recognizing its remarkable benefits, we highlight that in PLS-SEM because of their complexity it is difficult to provide overall suggestions. This is still an open issue for future development. After testing the results with and without controls, if the results do not differ, only the analysis without controls should be displayed, along with a motivation of the removal. Moreover, statistics of the controls must be included in the model results. Finally, researchers must be cautious with result generalizations involving controls. Indeed, by using them, a statistical sample of individuals is created for which predictors that are correlated are forced to be orthogonal or independent. Therefore, it is questionable to make inferences on the population.

References

Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J PERS SOC PSYCHOL 51(6):1173–1182

Becker T (2005) Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. ORGAN RES METHODS 8(3):274–289

Becker T, Atinc G, Breaugh J, Carlson K, Edwards J, Spector P (2016) Statistical control in correlational studies: 10 essential recommendations for organizational researchers. J ORGAN BEHAV 37(2):157–167

Breaugh J (2008) Important considerations in using statistical procedures to control for nuisance variables in non-experimental studies. HUM RESOUR MANAGE R 18(4):282–293

Chin W, Marcolin B, Newsted P (2003) A partial least squares latent variable modeling approach for measuring interaction effects: Results from a monte carlo simulation study and an electronic-mail emotion/adoption study. INFORM SYST RES 14(2):189–217

Diamantopoulos A, Sarstedt M, Fuchs C, Wilczynski P, Kaiser S (2012) Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. J ACAD MARKET SCI 40(3):434–449

Hair J, Sarstedt M, Ringle C, Mena J (2011) An assessment of the use of partial least squares structural equation modeling in marketing research. J ACAD MARKET SCI 40(3):414–433

Henseler J, Ringle C, Sinkovics R (2009) The use of partial least squares path modeling in international marketing. In: Rudolf R Sinkovics PNG (ed) New Challenges to International Marketing, vol 20, Emerald Group Pub. Lim., pp 277–319

Judd C, Kenny D (1981) Process analysis: Estimating mediation in treatment evaluations. Evaluation Review 5(5):602–619

Lohmöller J (1989) Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag Heidelberg

Spector P, Brannick M (2011) Methodological urban legends: The misuse of statistical control variables. ORGAN RES METHODS 14(2):287–305

Spector P, Zapf D, Chen P, Frese M (2000) Why negative affectivity should not be controlled in job stress research: don’t throw out the baby with the bath water. J ORGAN BEHAV 21(1):79–95

Wold H (1985) Partial least squares. In: Kotz S, Johnson N (eds) Encyclopedia of Statistical Sciences, John Wiley, New York, pp 581–591