Data Analysis Perspectives Journal, 3(3), 1-6, February 2022

Testing and controlling for endogeneity in PLS-SEM with stochastic instrumental variables

Ned Kock
Texas A&M International University, USA

Abstract

We discuss a procedure that can be used to simultaneously test and control for endogeneity in models analyzed with structural equation modeling via partial least squares (PLS-SEM). It relies on the creation of stochastic instrumental variables for endogenous latent variables, and their use as control variables. The procedure can be seen as an implementation of the Durbin–Wu–Hausman test, often referred to as the Hausman test, with stochastic instrumental variables. It can also be seen as a generalization of the two-stage least squares procedure. We illustrate the procedure with WarpPLS, a leading PLS-SEM tool.

Keywords: Endogeneity; Instrumental Variable; Analytic Composites; Path Coefficients; Structural Equation Modeling; Partial Least Squares; WarpPLS.

Introduction

Endogeneity, in the context of structural equation modeling via partial least squares (PLS-SEM), is a phenomenon whereby the structural error associated with an endogenous latent variable (LV) is correlated with one or more of the LVs pointing at the endogenous LV. If endogeneity exists with respect to an endogenous LV, the path coefficients estimated for the links pointing at the LV may be biased.

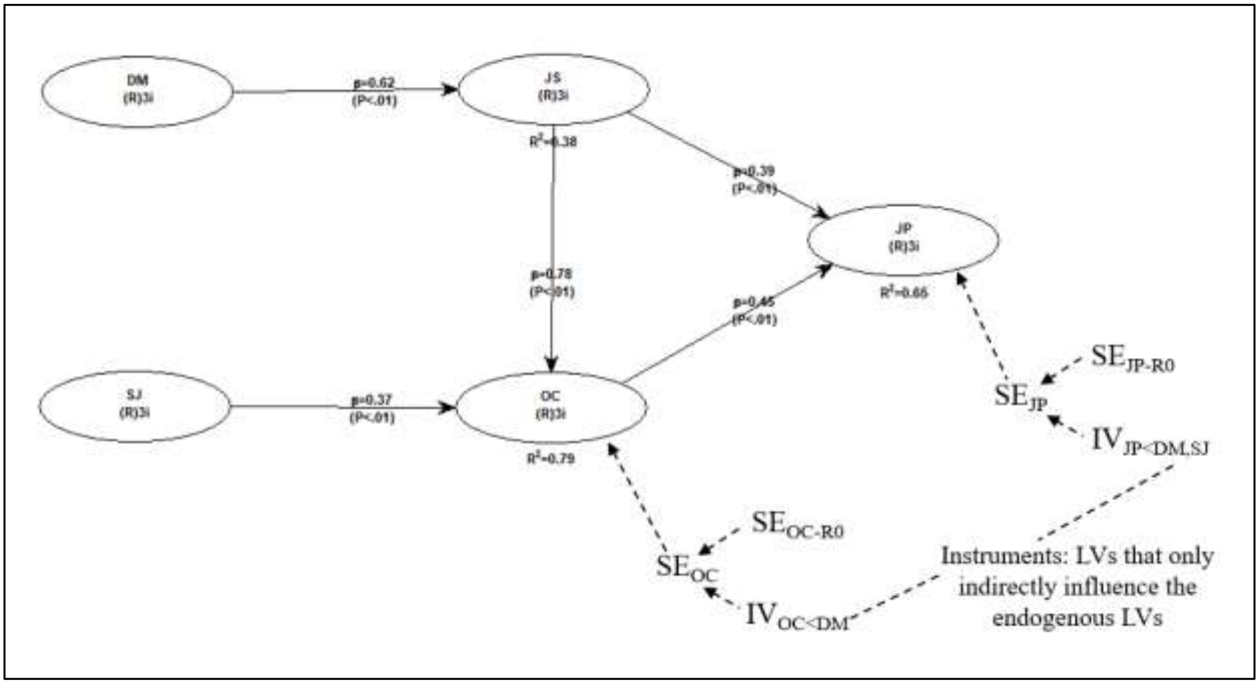
We discuss a procedure employing stochastic instrumental variables that is aimed at simultaneously testing and controlling for endogeneity. The procedure is illustrated with WarpPLS, a leading software tool that implements classic composite-based and more modern factor-based PLS-SEM algorithms (Kock, 2019a; 2019b), among other features that are useful in complex SEM analyses (Amora, 2021; Canatay et al., 2022; Hubona & Belkhamza, 2021; Kock, 2015b; 2016; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2022; Moqbel et al., 2020; Morrow & Conger, 2021; Rasoolimanesh, 2022).

The procedure discussed in this paper can be seen as an implementation employing stochastic instrumental variables of the Durbin–Wu–Hausman test, frequently referred to as simply the Hausman test. These are variables that differ from the instrumental variables normally used in econometrics and the classic Hausman test, in one key respect: they start as random uncorrelated variables that subsequently acquire variation from their instruments via the technique of variation sharing (Kock, 2019a). The instrumental variables used in econometrics and the classic Hausman test are in fact composites that aggregate instruments; as such, they tend to add massive collinearity to models if they are added to them.

Stochastic instrumental variables

The illustrative model shown in Figure 1 contains two exogenous LVs, namely DM and SJ; and three endogenous LVs, which are JS, OC and JP. Note that according to the model OC receives variation from DM only indirectly, via JS. Also, note that JP receives variation from DM and SJ only indirectly, via JS and OC.

Figure 1: Illustrative model used



Notes: DM = democratic management; SJ = scarcity of comparable jobs; JS = job satisfaction; OC = organizational commitment; JP = job performance; notation under latent variable acronym describes measurement approach and number of indicators, e.g., (R)3i = reflective measurement with 3 indicators.

If OC shares variation with DM that is not fully received via JS, the paths that point at OC may be distorted. This type of contamination can happen for various reasons, such as: the existence of hidden reciprocal relationships; common method bias; the confounding influence of omitted variables; data collection at multiple levels; and certain nonlinearity patterns (Kock, 2015b; 2020c; 2021a; 2021b; 2021c; Kock & Lynn, 2012).

The same is true for JP. If JP shares variation patterns with DM and-or SJ that are not fully received via JS and OC, then the paths that point at JP may be distorted. These problems may occur if DM and-or SJ are in fact endogenous, in a "hidden" way, even though they are hypothesized through the model as being exogenous.

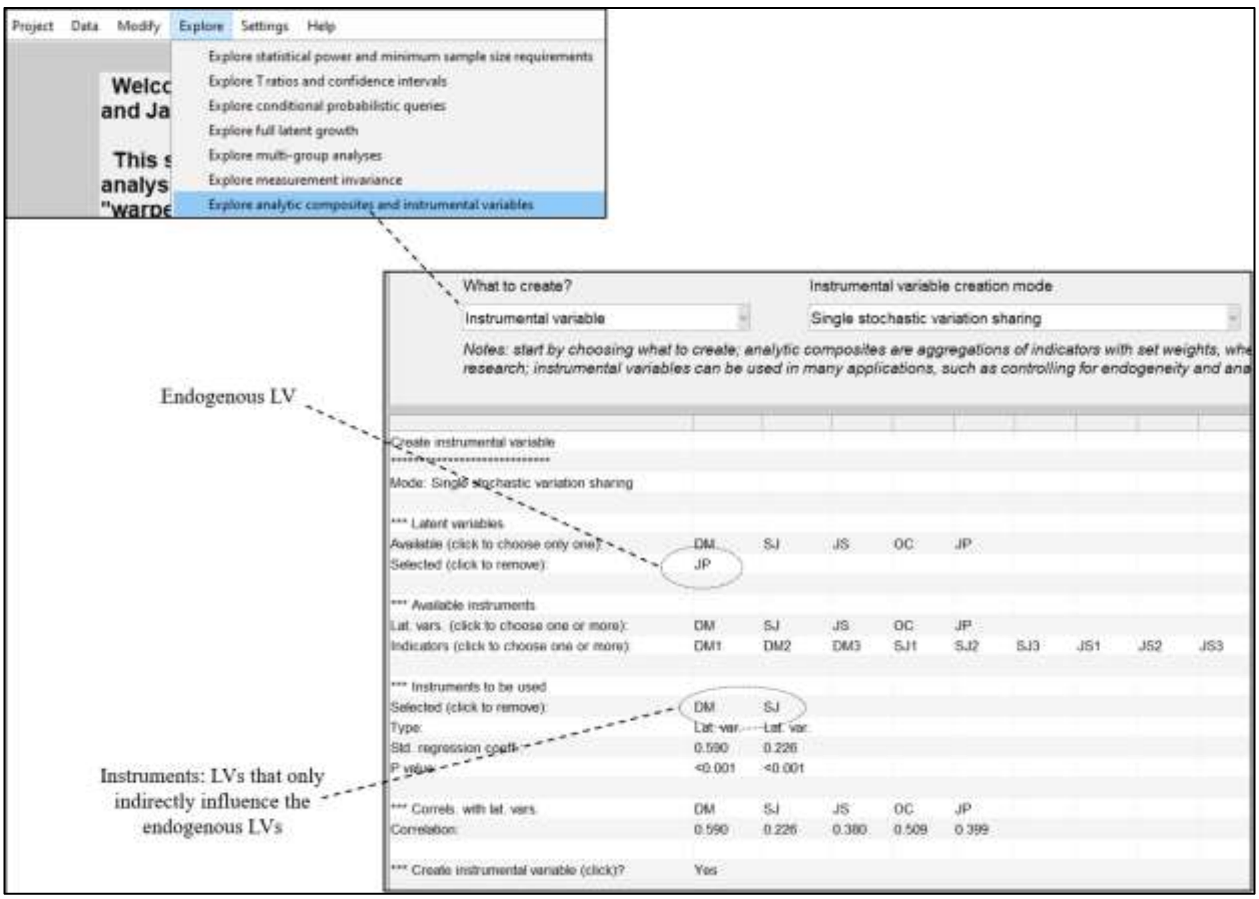
It is important to stress that it would be incorrect to add direct links among the exogenous variables DM and SJ in the model and the endogenous variables OC and JP. The reason this would be incorrect is that if these direct links do not exist in reality (i.e., at the population level), which is clearly suggested by the theory underlying the model (otherwise the direct links would be part of the model), then the inclusion of these direct links could lead to type I errors.

The structural error terms associated with OC and JP can be seen as aggregating two types of variables: instrumental variables, which contain the extraneous variation; and uncorrelated error terms. In the case of OC, the instrumental variable in question is indicated as $IV_{OC<DM}$ and the corresponding uncorrelated error term is $SE_{OC<R0}$. In the case of JP, the instrumental variable in question is $IV_{JP<DM,SJ}$ and the matching uncorrelated error term is $SE_{JP<R0}$. Here the “R0” in the subscript is meant to indicate that the term is an uncorrelated residual.

Creating instrumental variables

Instrumental variables associated with endogenous LVs can be created in WarpPLS through the menu sub-option “Explore analytic composites and instrumental variables”, under the main menu option “Explore” (see Figure 2). This sub-option becomes available after Step 5 is completed. Users should choose to create an “Instrumental variable” using the creation mode “Single stochastic variation sharing”.

Figure 2: Creating instrumental variables

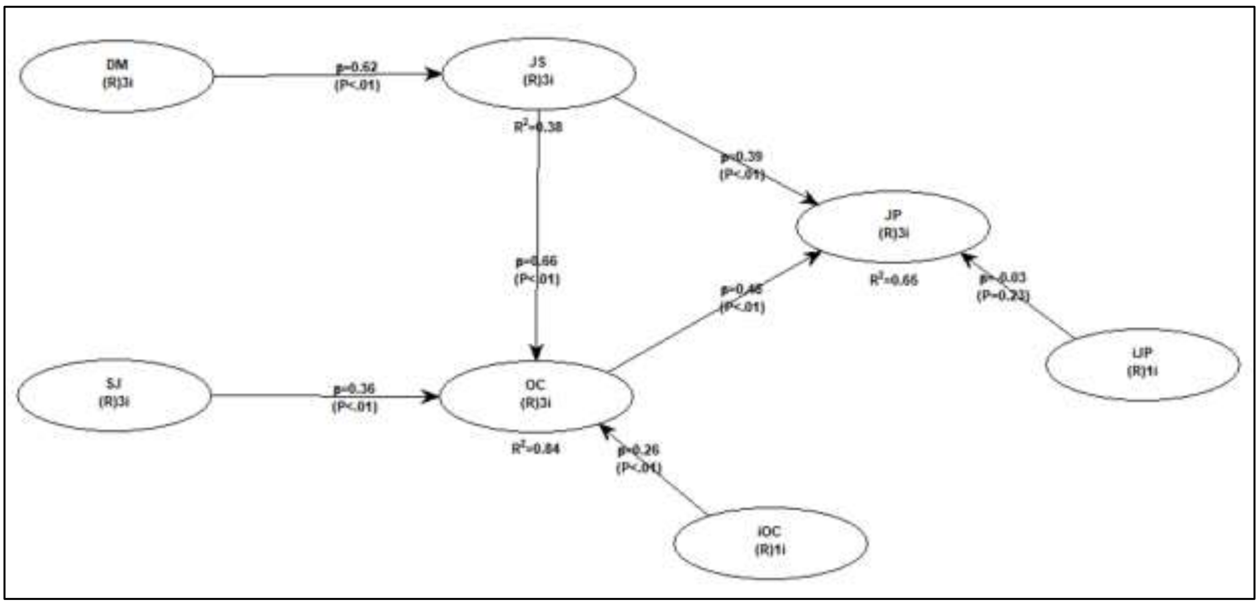


In our case, we need to create instrumental variables for JP and OC. The instrumental variable for OC will have as its instrument only DM, since this is the only LV that causes it indirectly. The instrumental variable for JP will have as its instruments DM and SJ. It should be noted that, for completeness, instruments should be all LVs that influence an endogenous LV indirectly, even if the instruments themselves are endogenous LVs. This makes the use of instrumental variables, as proposed here, a generalization of the two-stage least squares procedure.

Testing and controlling for endogeneity

Once instrumental variables are created, they should be added to the model, each pointing at the endogenous LV for which they were created (see Figure 3). If the path coefficient for the link between an instrumental variable and its endogenous LV is statistically significant, then that is an indication of endogeneity, which is nevertheless controlled for by the instrumental variable link.

Figure 3: Testing and controlling for endogeneity



As we can see, the path coefficient for the link iOC > OC is statistically significant (β=0.26, P<0.01), which means that endogeneity seems to exist with respect to OC. This endogeneity is controlled for via iOC. On the other hand, the path coefficient for the link iJP > JP is not statistically significant (β=0.03, P=0.23), which means that endogeneity does not seem to exist with respect to JP.

Since the instrumental variables are control variables, for this type of test one should arguably use two-tailed P values (Kock, 2015a), which are twice the one-tailed values shown on graphs – e.g., P=0.46 for iJP > JP, which is twice the P=0.23 shown on the graph. The recommended threshold for statistical significance is still the generally used value of 0.05 (Kock, 2015a). That is, a two-tailed P value lower than 0.05 would suggest the presence of statistically significant endogeneity.

If we compare this model including iOC and iJP with the earlier model without these variables, we can see that the path coefficients for the links JS > OC and SJ > OC are noticeably different – particularly JS > OC. They are respectively 0.78 and 0.37 for the prior model, and 0.66 and 0.36 for this model including iOC and iJP. The reason for the difference is the statistically significant endogeneity with respect to OC.

On the other hand, the path coefficients for the links pointing at JP, namely JS > JP and OC > JP, are virtually the same for both models. They are respectively 0.39 and 0.45 for the prior model, and 0.39 and 0.46 for this model including iOC and iJP. This similarity is what one would expect, since the test suggested that there is no statistically significant endogeneity with respect to JP.

Conclusion

As noted earlier, the stochastic instrumental variable procedure discussed here can be seen as an implementation of the Durbin–Wu–Hausman test, often referred to as simply the Hausman test, with stochastic instrumental variables. It can also be seen as a generalization of the two-stage least squares procedure. One key difference is that the procedure discussed here allows one to simultaneously test *and* control for endogeneity. Instrumental variables associated with nonsignificant endogeneity can be either kept in or excluded from the model, as they do not affect the path coefficients for links going into their respective LVs.

Since the procedure relies on the creation of stochastic instrumental variables, and their inclusion in the model, it will usually increase full collinearity variance inflation factors (FCVIFs). Stochastic instrumental variables start as random uncorrelated variables that then acquire variation from their instruments via the technique of variation sharing (Kock, 2019a). As such, they do not increase FCVIFs as much as would instrumental variables traditionally used in econometrics for the classic Hausman test. The latter are in fact composites that aggregate instruments.

Nevertheless, it is recommended that FCVIFs be estimated *prior* to the implementation of the procedure and reported as the actual model FCVIFs for various tests – e.g., the widely used common method bias test using FCVIFs (Kock, 2015b). This can easily be done by saving the WarpPLS project file prior to the procedure, and creating a new project file to store the results after the procedure. If users decide to use the FCVIFs obtained after the implementation of the procedure, with the model including the new stochastic instrumental variables, and the FCVIFs indicate that the model passes various tests, those users should interpret the results as more conservative than they would normally be.

Acknowledgments

The author is the developer of the software WarpPLS. He is grateful to WarpPLS users for questions, comments, discussions, and continued use.

References

Amora, J. T. (2021). Convergent validity assessment in PLS-SEM: A loadings-driven approach. *Data Analysis Perspectives Journal*, 2(3), 1-6.

Canatay, A., Emegwa, T., Lybolt, L. M. & Loch, K. D. (2022). Reliability assessment in SEM models with composites and factors: A modern perspective. *Data Analysis Perspectives Journal*, 3(1), 1-6.

Hubona, G., & Belkhamza, Z. (2021). Testing a moderated mediation in PLS-SEM: A full latent growth approach. Data Analysis Perspectives Journal, 2(4), 1-5.

Kock, N. (2015a). One-tailed or two-tailed P values in PLS-SEM? *International Journal of e-Collaboration*, 11(2), 1-7.
Kock, N. (2015b). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration*, 11(4), 1-10.
Kock, N. (2016). Non-normality propagation among latent variables and indicators in PLS-SEM simulations. *Journal of Modern Applied Statistical Methods*, 15(1), 299-315.
Kock, N. (2019a). From composites to factors: Bridging the gap between PLS and covariance-based structural equation modeling. *Information Systems Journal*, 29(3), 674-706.
Kock, N. (2019b). Factor-based structural equation modeling with WarpPLS. *Australasian Marketing Journal*, 27(1), 57-63.
Kock, N. (2020a). *WarpPLS User Manual: Version 7.0*. Laredo, TX: ScriptWarp Systems.
Kock, N. (2020b). Full latent growth and its use in PLS-SEM: Testing moderating relationships. *Data Analysis Perspectives Journal*, 1(1), 1-5.
Kock, N. (2020c). Multilevel analyses in PLS-SEM: An anchor-factorial with variation diffusion approach. *Data Analysis Perspectives Journal*, 1(2), 1-6.
Kock, N. (2020d). Using indicator correlation fit indices in PLS-SEM: Selecting the algorithm with the best fit. *Data Analysis Perspectives Journal*, 1(4), 1-4.
Kock, N. (2021a). Harman’s single factor test in PLS-SEM: Checking for common method bias. *Data Analysis Perspectives Journal*, 2(2), 1-6.
Kock, N. (2021b). Common structural variation reduction in PLS-SEM: Replacement analytic composites and the one fourth rule. *Data Analysis Perspectives Journal*, 2(5), 1-6.
Kock, N. (2021c). Moderated mediation and J-curve emergence in path models: An information systems research perspective. *Journal of Systems and Information Technology*, 23(3), 303-321.
Kock, N. (2022). *WarpPLS User Manual: Version 8.0*. Laredo, TX: ScriptWarp Systems.
Kock, N., & Lynn, G.S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.
Moqbel, M., Guduru, R., & Harun, A. (2020). Testing mediation via indirect effects in PLS-SEM: A social networking site illustration. *Data Analysis Perspectives Journal*, 1(3), 1-6.
Morrow, D. L., & Conger, S. (2021). Assessing reciprocal relationships in PLS-SEM: An illustration based on a job crafting study. *Data Analysis Perspectives Journal*, 2(1), 1-5.
Rasoolimanesh, S. M. (2022). Discriminant validity assessment in PLS-SEM: A comprehensive composite-based approach. *Data Analysis Perspectives Journal*, 3(2), 1-8.