Using logistic regression in PLS-SEM: Dichotomous endogenous variables

Ned Kock
Texas A&M International University, USA

Abstract

A dichotomous endogenous variable would be impossible to occur at the population level, which an empirical sample is assumed to represent, because the structural error term associated with the endogenous variable is expected to be a random variable with many distinct values. Consequently, the endogenous variable is also expected to have many distinct values. This paper discusses how to address this problem, using logistic regression with the probit approach, in the context of structural equation modeling via partial least squares (PLS-SEM). Our discussion is based on an illustrative model analyzed with the software WarpPLS.

Keywords: Logistic Regression; Endogenous Variables; Dichotomous Variables; Structural Equation Modeling; Partial Least Squares; WarpPLS.

Introduction

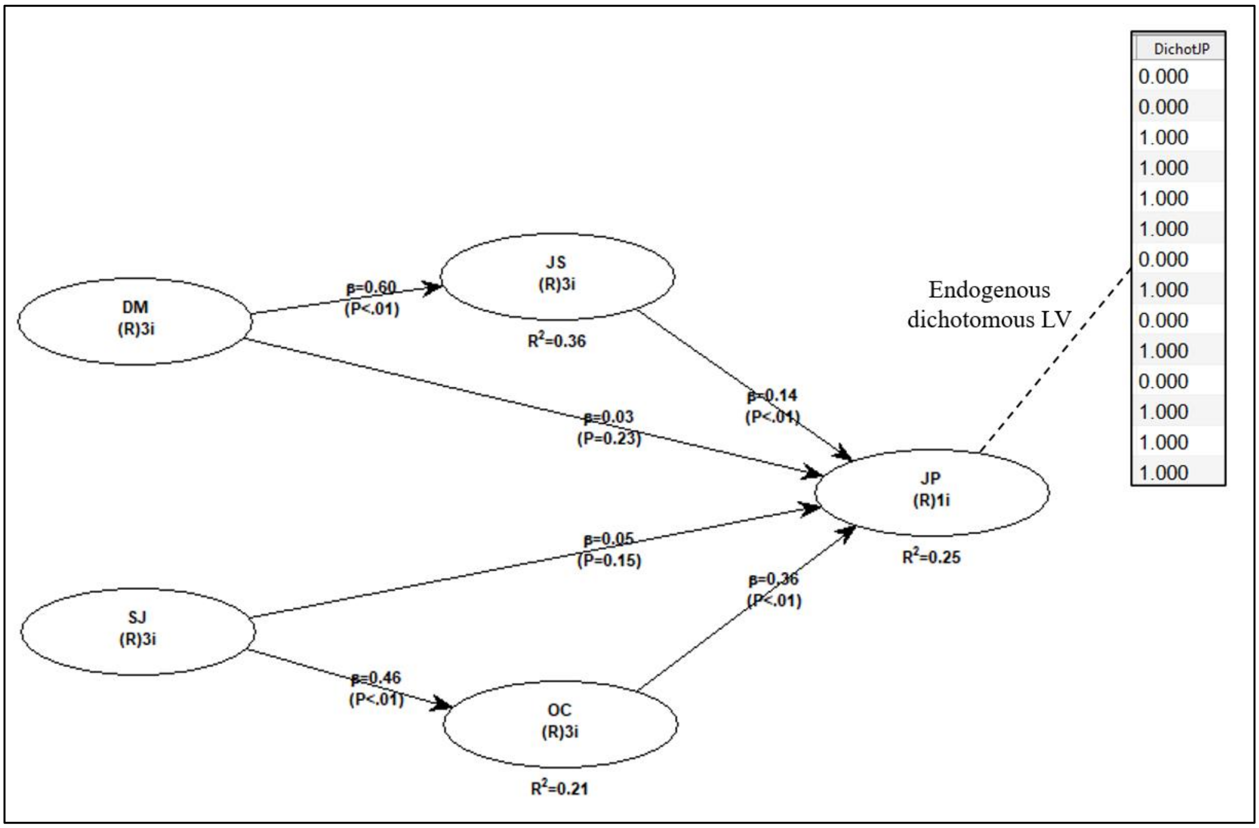
Often researchers include endogenous dichotomous variables in models aimed at analyzing empirical data. However, a dichotomous endogenous variable would be impossible to occur at the population level, which an empirical sample is assumed to represent, because the structural error term associated with the endogenous variable would be expected to be a random variable with many distinct values. Because of this, the endogenous variable would also be expected to have many distinct values (not only two), as it is an aggregation of its predictors in the model and the structural error term. This paper discusses how to address this problem using logistic regression, with the probit approach, in the context of structural equation modeling via partial least squares (PLS-SEM).

Our discussion is based on an illustrative model analyzed with the software WarpPLS, Version 8.0 (Kock, 2022a). This software is a widely used SEM tool that implements both classic composite-based as well as more modern factor-based PLS-SEM algorithms (Kock, 2019a; 2019b), where latent variables (LVs) are modeled as factors, among other features that can be useful in advanced SEM analyses (Amora, 2021; 2023; Canatay et al., 2022; Hubona & Belkhamza, 2021; Kock, 2015a; 2015b; 2015c; 2016; 2020a; 2020b; 2020c; 2021a; 2021b; 2021c; 2022a; 2022b; 2022c; 2023; Kock & Gaskins, 2016; Kock & Lynn, 2012; Ma & Zhang, 2023; Moqbel et al., 2020; Morrow & Conger, 2021; Rasoolimanesh, 2022).

Illustrative model and data

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. The results shown are based on a simulated dataset, created through the Monte Carlo method (Kock, 2016). The simulated dataset has a size of 500 and was created based on the illustrative model. Two of the endogenous LVs, namely JS and OC, have many distinct values, because they aggregate multiple indicators on 7-point scales (even though each indicator stores only 7 distinct values). The variable JP is measured on a two-point scale, 0 and 1, referring to low and high job performance. That is, the variable JP is dichotomous, with only two distinct values.

Figure 1: Model with dichotomous endogenous variable



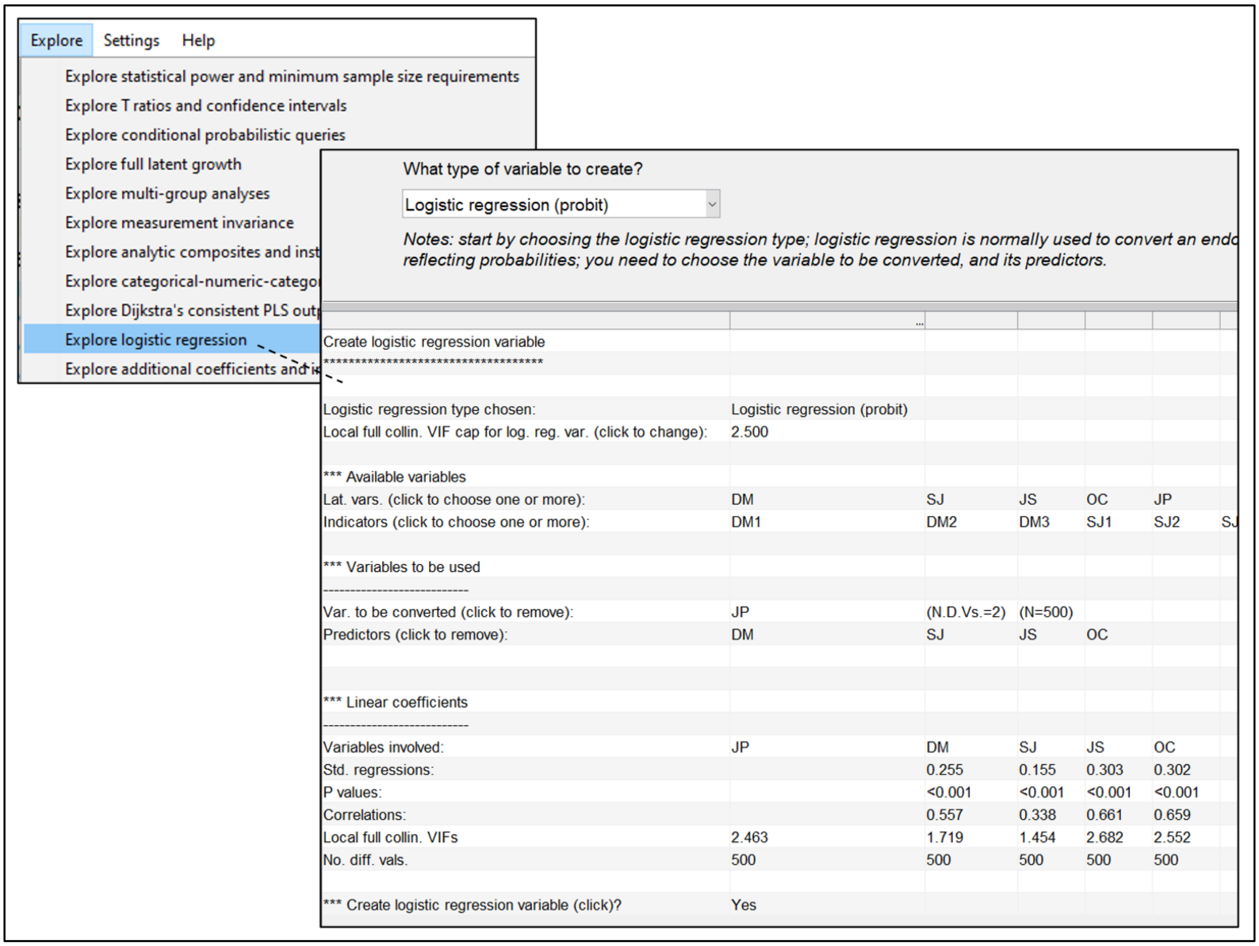
Notes: DM = democratic management; SJ = scarcity of comparable jobs; JS = job satisfaction; OC = organizational commitment; JP = job performance; notation under LV acronym describes measurement approach and number of indicators, e.g., (R)3i = reflective measurement with 3 indicators.

The outer model analysis algorithm used to generate the results in the illustrative model was "Factor-Based PLS Type CFM3". Like covariance-based SEM algorithms, this algorithm is factor-based and fully compatible with common factor model assumptions (Kock, 2019a; 2019b). The inner model analysis algorithm used was "Linear". This algorithm does not perform any warping of relationships. Both outer and inner model algorithms are fully compatible with the way in which the simulated data was created via the Monte Carlo method.

A dichotomous endogenous variable such as JP would be impossible to occur at the population level, which our sample is assumed to represent, because the structural error term associated with

JP would be expected to be a random variable with many distinct values (Kock, 2016). This also applies to situations where endogeneity exists, where the structural error term would be correlated with the endogenous variable's predictors.

Figure 2: Creating a logistic regression variable

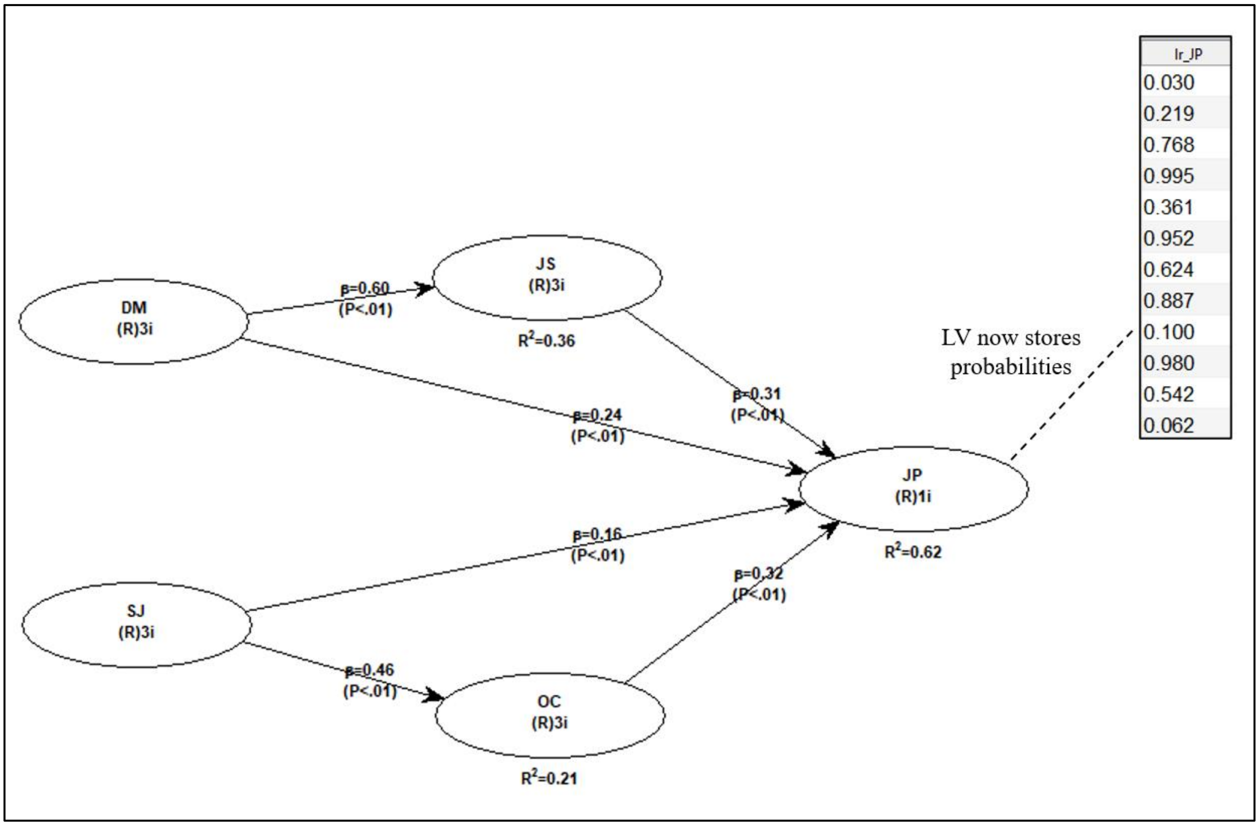


Because the structural error term is expected to be a random variable with many distinct values, JP would also be expected to have many distinct values, as it incorporates that structural error term. Since JP is the main dependent variable in our model, it would be particularly problematic to keep it on a two-point scale. Among other problems, this could suppress the path coefficients associated with JP's predictors in the model, possibly causing type II errors (false negatives), and leading to a corresponding suppressed R-squared value for JP.

Creating a logistic regression variable

To address the problem above, the variable JP was modeled as a logistic regression variable, estimated via the probit approach (Kock, 2022a). After the logistic regression modeling, JP stored the probabilities that the job performance would be high. The menu option "Explore logistic regression" allows one to create a logistic regression variable as a new indicator that has both unstandardized and standardized values (see Figure 2). This new indicator was used to measure JP, replacing the original dichotomous indicator.

Figure 3: New model with logistic regression variable



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

Two logistic regression approaches, or algorithms, are available: probit and logit. The former, namely probit (which we used here), is recommended for dichotomous variables; the latter for variables where the number of different values (a.k.a. “distinct observations”) is greater than 2 but still significantly smaller than the sample size; e.g., 10 different values over a sample size of 100.

The unstandardized values of a logistic regression variable are probabilities; going from 0 to 1. Since a logistic regression variable can be severely collinear with its predictors, one can set a local full collinearity VIF cap for the logistic regression variable. The software’s default is 2.5, set as such so that the provision of shared variance by the endogenous LV’s “mini-model” does not unduly raise the full collinearity VIFs for the whole model beyond the conservative threshold of 3.3 (Kock, 2015b). Predictor-criterion collinearity, or lateral collinearity (Kock & Lynn, 2012), is rarely assessed or controlled in classic logistic regression algorithms.

New model with logistic regression variable

If several predictors are available, the new logistic regression variable will typically incorporate much more variation than the endogenous variable on which it is based, which will typically be reflected in larger absolute coefficients of association (e.g., path coefficients). This can be seen in Figure 3, which shows increases in the path coefficients associated with the predictors of JP, and consequently a much higher R-squared for that endogenous variable. The exception is the link OC > JP, whose path coefficient was reduced.

The situation with the link OC > JP is not uncommon with respect to some paths associated with predictors of the endogenous LV, even though overall the paths are likely to increase in strength - i.e., being greater in absolute terms, whether the path coefficients are negative or positive.

It is important to stress that, for a logistic regression variable to be created, the original variable (in this example, the dichotomous version of JP) must be available. Also, predictors must exist and be selected. Note that the logistic regression variable was created assuming four predictors: DM, SJ, JS and OC. These are the predictors of JP in the model, which is presumably based on a theoretical framework that the structural model is meant to reflect.

Conclusion

Often researchers include endogenous dichotomous variables in their SEM models. However, a dichotomous endogenous variable would be impossible to occur at the population level, which an empirical sample is assumed to represent, because of a property of the structural error term associated with the endogenous variable, which (i.e., the error term) explains the variance in the endogenous variable that is not explained by the variable's predictors in the SEM model. The property in question is that the structural error term is expected to be a random variable with many distinct values. Because of this, the endogenous variable is also expected to have many distinct values, as it incorporates variation from that structural error term. This paper discussed how to address this problem by using logistic regression with the probit approach.

Acknowledgments

The author is the developer of the software WarpPLS. He is grateful to WarpPLS users for questions, comments, discussions, and continued use. This article contains revised text, originally written by the author, from a recent edition of the WarpPLS User Manual.

References

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

Amora, J. T. (2023). On the validity assessment of formative measurement models in PLS-SEM. *Data Analysis Perspectives Journal*, 4(2), 1-7.

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. (2015c). How likely is Simpson's paradox in path models? *International Journal of e-Collaboration*, 11(1), 1-7.

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). Full latent growth and its use in PLS-SEM: Testing moderating relationships. *Data Analysis Perspectives Journal*, 1(1), 1-5.

Kock, N. (2020b). Multilevel analyses in PLS-SEM: An anchor-factorial with variation diffusion approach. *Data Analysis Perspectives Journal*, 1(2), 1-6.

Kock, N. (2020c). 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. (2022a). *WarpPLS User Manual: Version 8.0*. Laredo, TX: ScriptWarp Systems.

Kock, N. (2022b). Testing and controlling for endogeneity in PLS-SEM with stochastic instrumental variables. *Data Analysis Perspectives Journal*, 3(3), 1-6.

Kock, N. (2022c). Using causality assessment indices in PLS-SEM. *Data Analysis Perspectives Journal*, 3(5), 1-6.

Kock, N. (2023). Assessing multiple reciprocal relationships in PLS-SEM. *Data Analysis Perspectives Journal*, 4(3), 1-8.

Kock, N., & Gaskins, L. (2016). Simpson's paradox, moderation, and the emergence of quadratic relationships in path models: An information systems illustration. *International Journal of Applied Nonlinear Science*, 2(3), 200-234.

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.

Ma, K. Q., & Zhang, W. (2023). Assessing univariate and multivariate normality in PLS-SEM. *Data Analysis Perspectives Journal*, 4(1), 1-7.

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.