Data Analysis Perspectives Journal, 5(1), 1-5, March 2024

Combining sub-samples for improved statistical power in PLS-SEM: A constrained latent growth approach

James Cox
Our Lady of the Lake University, USA

Abstract

Often researchers gather data that contain or can be segmented into subsamples. Therefore, sometimes a question arises as to whether the data can be treated as one sample or as several distinct samples. In this paper, I discuss how to conduct a multigroup analysis in a structural equation model with partial least squares (PLS-SEM) and demonstrate how empirical data from two different countries can be treated as one sample when using WarpPLS 8.0 to achieve higher statistical power.

Keywords: Multigroup Analysis, Latent Growth Analysis, Measurement Invariance, Structural Equation Modeling, Partial Least Squares, WarpPLS

Introduction

Often researchers gather data that contain or can be segmented into subsamples. Therefore, sometimes a question arises as to whether the data can be treated as one sample or as several distinct samples. In this paper, I discuss how to conduct a multigroup analysis in a structural equation model with partial least squares (PLS-SEM) to demonstrate how data from two different sub-samples can be treated as one sample. This analysis uses an illustrative model with empirical data collected from India and the United States that is analyzed with WarpPLS 8.0 and combined into one sample to achieve higher statistical power.

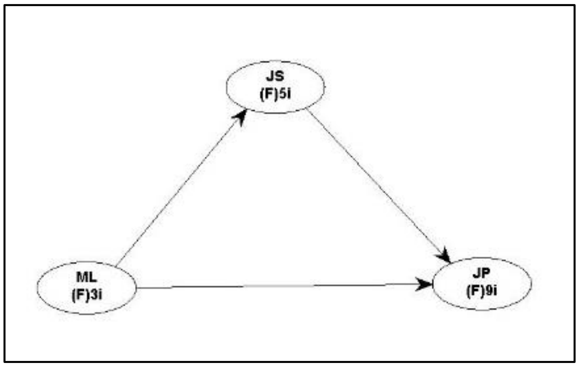
Illustrative model and data

The model that is used for this discussion is displayed in Figure 1. It contains three latent variables: the degree to which a supervisor uses motivating language (ML); the degree to which employees are satisfied with their jobs (JS); and the measure of how well employees perform their job (JP). The unit of analysis in this model is the employee.

The data were collected through online surveys using Google forms. Two hundred and fifty requests were sent to potential respondents in the United States and in India through Amazon Mechanical Turk, and 250 responses were received for each country. These requests resulted in 361 usable surveys:196 and 165 form the US and India, respectively. Both the Indian and American surveys were in English. Some responses were dropped from the study for various reasons such as not being natives of the country in which they lived, not being full-time employees, and failing an instructional manipulation check.

The model makes the following predictions: the use of ML by a supervisor leads to increased Sat and to better JP. Increased JS leads to improved JP.

Figure 1: Illustrative model



Multigroup analysis: Differences between the US and Indian sub-samples

Sometimes the data collected for an empirical study are composed of several subgroups, and there are theoretical reasons to believe that membership in one subgroup may affect the results. Splitting the sample can result in a smaller sample size that leads to lower statistical power that in turn can lead to Type 2 and capitalization errors (overestimation of a small path coefficient for a small sample) (Kock & Hadaya, 2018) as well as convergence failures (Kock, 2023). In this paper, a multigroup analysis is conducted to demonstrate that the Indian and US sub-samples are not significantly different from each other. Consequently, the samples can be treated as one, increasing the study's statistical power.

The analysis uses WarpPLS 8.0 by choosing the *explore multigroup analysis* option. The grouping by variable type uses the *unstandardized indicator*, and the *grouping by variable* option was set to the indicator *Ctry10UI*, (a categorical variable in the dataset that has values of 0 for India and 1 for the US). The analysis method was set to *constrained latent growth*. This segments the data according to the selected variable in order to analyze all possible pairings by applying the same model to each of the resulting sub-samples (Kock, 2023). This process is similar to a full latent growth analysis since "it does not "disrupt" the model in any way" (Kock, 2020).

Table 1 shows the path coefficients of the US and Indian sub-samples as well as those of the combined sample. In the initial analysis, the path coefficients of the two sub-samples appear to be mostly similar. Table 2 shows that the full collinearity of the variance inflation factors (VIFs) are below the threshold value of five for both sub-samples. This value indicates that excessive collinearity from one sub-sample is not being subsumed by the other sub-sample's lack of collinearity (Hair, Ringle, & Sarstedt, 2011; Kline, 2005; Kock, 2023). Table 3 shows the absolute difference between the full collinearity VIFs of the Indian and US models.

Table 1: Path coefficients

	India		United States		Combined Sample	
	ML	JS	ML	JS	ML	JS
JS	0.816*		0.713*		0.760*	
JP	0.267*	0.344*	0.150*	0.212*	0.260*	0.200*

*P < 0.01

Table 2: Full collinearity VIFs

	ML	JS	JP
IN	2.894	3.069	1.265
US	2.062	2.045	1.089

Table 3: Absolute differences in the full collinearity VIFs

ML	JS	JP
0.832	1.025	0.177

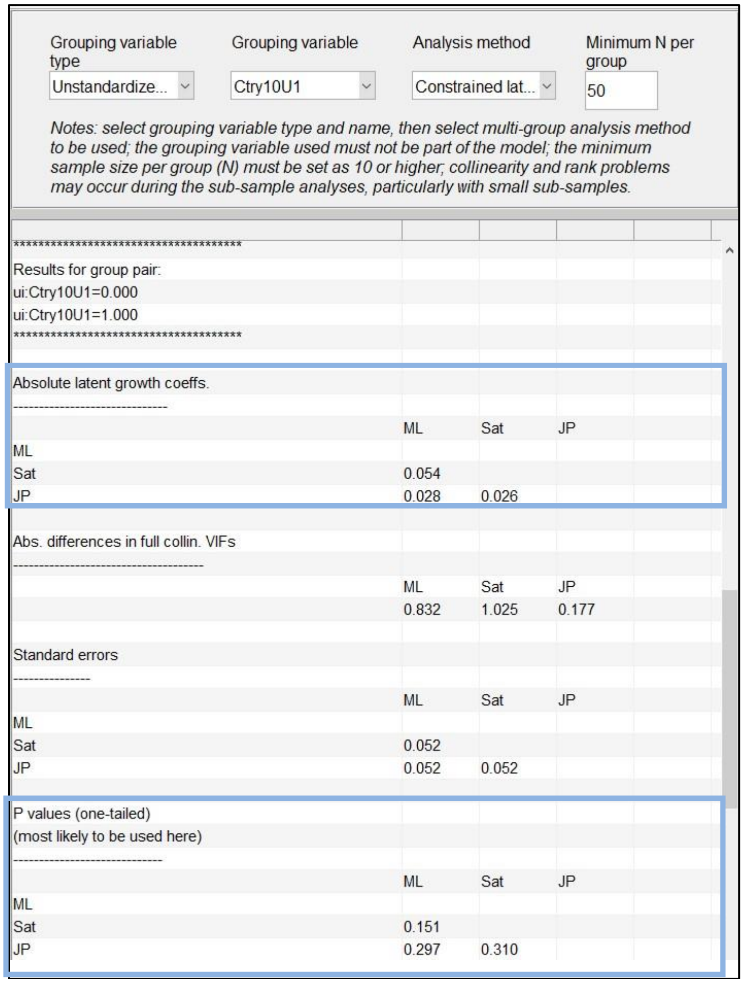
Figure 2 shows that the absolute latent growth coefficients are quite small. It also shows their p-values. As mentioned above, the method used for this analysis is constrained latent growth, which treats the segmenting variable or indicator as a moderating variable by estimating the interaction effects between it and all the paths in the model at once without including any links in the model. Therefore, the absolute latent growth coefficients are akin to the moderating effect that the country of origin of the respondent has on the paths in the model (Kock, 2023). According to this analysis, there are no absolute latent growth coefficients that are statistically significant. The above results indicate that there is no meaningful statistical difference between the US and Indian models, therefore both sub-samples can be treated as one.

A power analysis was conducted for the combined sample (Figure 3) as well as the India and US sub-samples. As in Ezeugwa, et al (2022) this was done by choosing the “Explore statistical power and minimum sample requirements” option from the “Explore” menu. The “Minimum absolute significant path coefficient” was set with the value of the smallest path coefficient in the model (0.200). Next, the “Power level required” was manually adjusted until the value of “minimum required sample size” reached 361, which is the number of observations in the combined sample.

The power analysis was repeated for the Indian and US sub-samples. It is important to note that this power and minimum sample size requirement tool in WarpPLS 8 is independent of the model. That is: this power analysis can be conducted regardless of whether the model is being constrained for a multigroup analysis or not.

The analysis indicates that the US model has a power of 0.696 and the Indian model a power of 0.976, while the combined sample model has a power of 0.988. Power analyses were conducted at the 5% significance level and are based on the more accurate gamma-exponential method (Kock & Hadaya, 2018). As Table 4 shows, the power of the US sub-sample is below the minimum requirement of 0.800 (Kock & Hadaya, 2018). The minimum sample size for the US as indicated by the gamma exponential method in WarpPLS 8.0 of 262. The minimum sample sizes were calculated using a minimum power of 0.800 for the 3 models.

Figure 2: Absolute latent growth coefficients and their P values



Grouping variable type: Unstandardize...

Grouping variable: Ctry10U1

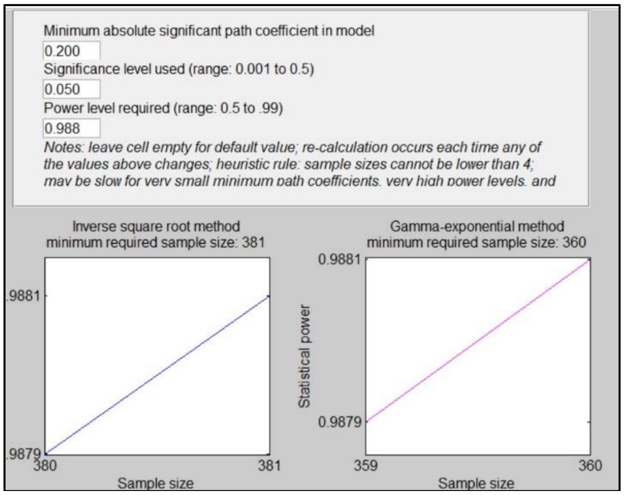
Analysis method: Constrained lat...

Minimum N per group: 50

Notes: select grouping variable type and name, then select multi-group analysis method to be used; the grouping variable used must not be part of the model; the minimum sample size per group (N) must be set as 10 or higher; collinearity and rank problems may occur during the sub-sample analyses, particularly with small sub-samples.

```
*************************************
Results for group pair:
ui:Ctry10U1=0.000
ui:Ctry10U1=1.000
*************************************

Absolute latent growth coeffs.
------------------------------
                         ML      Sat     JP
ML
Sat                      0.054
JP                       0.028   0.026

Abs. differences in full collin. VIFs
-------------------------------------
                         ML      Sat     JP
                         0.832   1.025   0.177

Standard errors
---------------
                         ML      Sat     JP
ML
Sat                      0.052
JP                       0.052   0.052

P values (one-tailed)
(most likely to be used here)
-----------------------------
                         ML      Sat     JP
ML
Sat                      0.151
JP                       0.297   0.310
```

Figure 3: Power analysis for the combined sample



Minimum absolute significant path coefficient in model: 0.200

Significance level used (range: 0.001 to 0.5): 0.050

Power level required (range: 0.5 to .99): 0.988

Notes: leave cell empty for default value; re-calculation occurs each time any of the values above changes; heuristic rule: sample sizes cannot be lower than 4; mav be slow for verv small minimum path coefficients. verv hiah power levels. and

Inverse square root method minimum required sample size: 381

Gamma-exponential method minimum required sample size: 360

Table 4: Power Analysis

Sample	Size	Power	Minimum Size (0.800 power)
India	165	0.976	74
US	196	0.696	262
Combined	361	0.988	142

Conclusion

Often researchers gather data that contain or can be segmented into sub-samples, but a researcher can encounter situations in which treating the sub-samples as one sample is preferable. In this paper, I show how to conduct multigroup to support the use of one sample instead of sub-samples to increase statistical power.

Acknowledgments

The author is thankful to the editor of this journal, Dr. Ned Kock for his invaluable guidance and his dedication to the continued development of WarpPLS, as well as to the anonymous reviewers who provided valuable feedback for this article.

References

Ezeugwa, B., Talukder, M. F., Amin, M. R., Hossain, S. I., & Arslan, F. (2022). Minimum sample size estimation in SEM: Contrasting results for models using composites and factors. *Data Analysis Perspectives Journal, 3*(4), 1-7.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice, 19*(2), 139-152.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed. ed.). New York: Guilford Press.

Kock, N. (2020). Full latent growth and its use in PLS-SEM: Testing moderating relationships. *Data Analysis Perspectives Journal, 1*(1), 1-5.

Kock, N. (2023). *WarpPLS 8.0 User Manual*. Laredo, TX USA.

Kock, N., & Hadaya, P. (2018). Minimum Sample Size Estimation in PLS-SEM: The Inverse Square Root and Gamma-Exponential Methods. *Information Systems Journal, 28*(1), 227-261.