INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 14, Issue 10, (2024) E-ISSN: 2222-6990

Data Analysis Using Partial Least Squares Structural Equation Modeling (PLS-SEM) in Conducting Quantitative Research

Yusuf Haji-Othman, Mohd Sholeh Sheh Yusuff, Muhammad Nasri Md Hussain

Sultan Abdul Halim Mua'dzam Shah International Islamic University (UniSHAMS), Kedah, Malaysia

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v14-i10/23364 DOI:10.6007/IJARBSS/v14-i10/23364

Published Date: 17 October 2024

Abstract

In contemporary research, Partial Least Squares Structural Equation Modeling (PLS-SEM) has emerged as a crucial statistical tool, particularly effective for analyzing complex structural models involving multiple constructs and indicators. This paper aims to elucidate the application of PLS-SEM in quantitative research, highlighting its advantages in extending theories and simultaneously estimating measurement and structural models. The methodological approach is divided into three primary stages: data screening and diagnostic tests, measurement model assessment, and structural model assessment. The data screening ensures dataset suitability by addressing missing data and outliers, while diagnostic tests fulfil normality, linearity, and multicollinearity assumptions. The measurement model assessment validates constructs through composite reliability and average variance extracted (AVE) metrics. The structural model assessment evaluates the significance and relevance of relationships between constructs, determines the coefficient of determination (R^2 and adjusted R^2), assesses mediating effects, and analyzes the moderating variables. By detailing these methodological steps, the article provides a comprehensive guide for researchers aiming to employ PLS-SEM in their studies, emphasizing its rigour and practicality in handling complex theoretical models.

Keywords: Partial Least Squares Structural Equation Modeling (PLS-SEM), data screening, measurement model, structural model.

Introduction

In contemporary research, data analysis techniques are paramount for deriving meaningful insights from complex datasets. Among these techniques, Partial Least Squares Structural Equation Modeling (PLS-SEM) has gained prominence, particularly in intricate structural models involving multiple constructs and indicators. This article aims to elucidate the application of PLS-SEM in research, underscoring its relevance and advantages in extending existing theories and simultaneously estimating measurement and structural models.

PLS-SEM is a robust statistical tool suited for scenarios where traditional covariance-based SEM might falter due to its stringent assumptions. It is advantageous in exploratory research phases and when the primary objective is prediction and theory development. This study leverages PLS-SEM due to its ability to handle complex models and its flexibility in dealing with non-normal data distributions and smaller sample sizes.

The methodological approach in this article is divided into three primary stages: data screening and diagnostic tests, assessment of the measurement model, and assessment of the structural model. Initially, the data screening process ensures the dataset's suitability for multivariate analysis by addressing missing data and outliers. Subsequently, diagnostic tests are performed to fulfil assumptions of normality, linearity, and multicollinearity.

Following data preparation, the assessment of the measurement model is conducted to validate the constructs and ensure reliability and validity through composite reliability and average variance extracted (AVE) metrics. The structural model assessment involves evaluating the significance and relevance of relationships between constructs, determining the coefficient of determination (R^2 and adjusted R^2), assessing mediating effects, and analyzing the analyzing moderating variables.

By detailing these methodological steps, this paper provides a comprehensive guide for researchers aiming to employ PLS-SEM in their studies, highlighting its methodological rigour and practical utility in handling complex theoretical models. The subsequent sections delve into each stage of the data analysis process, offering an understanding of PLS-SEM's application in research.

Data Analysis

Many researchers use PLS-SEM to analyze data. One of the reasons for using PLS-SEM is when the structural model is complex (Hair, Ringle, & Sarstedt, 2011). The complex structural model with many constructs and indicators leads to the selection of PLS-SEM for the analysis. Furthermore, according to Hair et al. (2011), the selection of PLS-SEM is more appropriate when extending an existing theory. Moreover, the advantage of PLS-SEM is that it can simultaneously estimate measurement and structural models.

The data analysis using PLS-SEM involves three stages. The first stage consists of a data screening and diagnostic tests to satisfy multivariate assumptions. This stage tests whether the data is suitable for statistical analysis. The second stage involves assessing the measurement model to identify the underlying structure of the variables involved (Hair et al., 2006). In the third stage, this study assesses the structural model, and the data is run using a partial least squares structural equation model (PLS-SEM).

The measurement model is performed using the PLS algorithm technique to validate the measurement scale of a construct (Hair et al., 2014). Variables that pass this analysis test are then used in the structural model analysis to examine the relationships between the endogenous and exogenous variables of the study.

Stage 1: Data Screening Process and Diagnostic Tests
Before the multivariate analysis, all data must go through the data screening process. According to Hair et al. (2006) and Hair, Black, Babin, & Anderson (2010), before performing the assessment of the measurement model and measurement of the structural model, the process needs to go through two stages:

i) Data screening process and
ii) Diagnostic tests to fulfil multivariate statistical analysis assumptions.

The data screening process involves detecting missing data and outliers. The purpose of these diagnostic checks is to clean the data before analyzing it (Hair et al., 2010).

To manage missing data, researchers should obtain descriptive information to identify the number of questionnaires which contain missing information. If the number of questionnaires that contain complete information is enough for analysis, then the questionnaires that contain the missing data will not be used. However, if the number of samples is not enough, the remedial measures of the data are undertaken.

Researchers could use multivariate detection to detect outliers. For multivariate outliers, the Mahalanobis distances (D^2) test is used on all exogenous variables. A questionnaire is a multivariate outlier if the value of D^2/degree of freedom is 2.5 in small samples (80 or fewer observations) and 3 or 4 in larger samples (Hair et al., 2006; Hair et al., 2010). According to Hair et al. (2010), outliers should be retained unless demonstrable proof indicates that they are genuinely aberrant and not representative of any observations in the population. If they portray a representative element or segment of the population, they should be retained to ensure generalizability to the entire population. After deleting outliers, the researcher runs the risk of improving the multivariate analysis but limiting its generalizability (Hair et al., 2010)

After the data screening, the data must undergo diagnostic tests to fulfil multivariate assumptions. Normality, linearity, and multicollinearity tests are performed to fulfil multivariate assumptions. These tests must be performed before the measurement model and structural model assessment can be performed (Hair et al., 2006; Hair et al., 2010; Hair et al., 2014).

The normality test is performed to determine whether or not scores for each variable have a normal distribution. The values of skewness and kurtosis are used as a guide to measure normality.

The statistical value (z) for the skewness value is calculated as:

$$z_{\text{skewness}} = \frac{\text{skewness}}{\sqrt{\frac{6}{N}}},$$

Where N is the sample size. A z value can also be calculated for the kurtosis value as follows:

$$z_{\text{kurtosis}} = \frac{\text{kurtosis}}{\sqrt{\frac{24}{N}}}$$

If either calculated z value exceeds the specified critical value, then the distribution is not normal. Data is said to be normal at a 99 per cent confidence level if the value of skewness and kurtosis is less than ± 2.58 at the .01 significance level and ±1.96 at the 0.05 significance level (Hair et al., 2010).

Researchers should perform a linearity test to examine whether or not there is a linear relationship between two variables. To test for linearity, data can be plotted using a scatterplot and matched with a linear line (Coakes & Steed, 2003).
The last multivariate assumption is that there is no multicollinearity. Multicollinearity exists when a high correlation exists between two or more independent variables. The high correlation between two or more independent variables will reduce the predictive power of the variables (Hair et al., 2006). A method to detect multicollinearity is by examining tolerance value and Variance Inflation Factor (VIF). Based on the tolerance value and Variance Inflation Factor (VIF), multicollinearity is said to exist if the tolerance value is less than the cutoff threshold, which is 0.10. This value is equal to the value of 10 VIF, as proposed by (Hair et al., 2006)

Stage 2: Assessment of Measurement Model
After undergoing the data screening process and taking remedial measures to overcome the problems of normality, linearity and multicollinearity, the measurement model is assessed. The purposes of performing the assessment of the measurement model are:

i. To examine basic dimensions for construct variables.
ii. To validate the dimensions and
iii. To determine the number of dimensions for each of the constructs.

Assessment of the measurement model includes composite reliability and average variance extracted (AVE) to determine reliability and validity.

Composite Reliability
The purpose of assessing composite reliability is to examine a construct's internal consistency and reliability. On the other hand, the purpose of assessing the average variance extracted is to evaluate convergent validity (Hair et al. 2014).

The reliability test is essential to determine the consistency and stability of instruments with the concepts to be measured (Sekaran, 2003). A reliability test is an early indicator to assess the quality of an instrument (Churchill, 1979). Traditionally, Cronbach’s alpha procedure is used to determine the reliability of a construct. Researchers can use this procedure because it is the most basic reliability test for any research (Churchill, 1979). However, Cronbach’s alpha assumes that all items are equally reliable; all items have equal outer loadings on the construct (Hair et al. 2014). However, researchers should use PLS-SEM, which prioritizes according to their reliability. Because of the limitation of Cronbach’s alpha, researchers should choose composite reliability to measure internal consistency. Composite reliability takes into consideration the different outer loadings of the items in the construct. The formula for composite reliability is given as follows:

$$\rho_c = \frac{\left(\sum_{i=1}^{n} L_i\right)^2}{\left(\sum_{i=1}^{n} L_i\right)^2 + \left(\sum_{i=1}^{n} var\,(e_i\right)^2}$$

Where L_i stands for the standardized item i of a construct, e_i is the measurement error of item i, and var(e_i) represents the variance of measurement error, defined as $(1 – L_i^2)$.

The composite reliability values range between 0 and 1. The higher the composite reliability, the higher the level of reliability. According to Hair et al. (2014), if composite reliability values are 0.60 to 0.70, then it is acceptable. Composite reliability values of less than 0.60 show a lack of internal consistency reliability.

Convergent Validity

Convergent validity refers to the extent to which an item correlates positively with alternative items of the same construct. The items of a specific construct should converge, which means they share a high proportion of variance (Hair et al., 2014). To evaluate convergent validity, researchers should assess the outer loadings of the items, together with the average variance extracted (AVE).

If the outer loadings of items in a specific construct are high, then the items have much in common captured by the construct. This situation is called indicator reliability. All outer loadings of all items should be statistically significant and at least 0.708 (Hair et al. 2014).

If the outer loadings are less than 0.708, researchers should examine the effect of removing the item on composite reliability. Hair et al. (2014) suggested that items having outer loadings between 0.40 and 0.70 should be removed only if the removal increases composite reliability and average variance extracted (AVE). Items with outer loadings of less than 0.40 must be deleted from the construct (Hair, Ringle, & Sarstedt, 2011).

Research could use average variance extracted (AVE) to establish convergent validity, as suggested by Hair et al. (2014). AVE is defined as the mean value of the squared loadings of the items associated with a specific construct. It measures the sum of the squared loadings divided by the number of items in the construct.

The average variance extracted (AVE) is calculated as the mean-variance extracted for the items loading on a construct. AVE is calculated using the following formula:

$$AVE = \frac{\sum_{i=1}^{n} L_i^2}{n}$$

Where L_i is the standardized factor loading, and i is the number of items. An AVE of 0.5 or higher shows adequate convergence.

The minimum acceptable value of AVE is 0.50 because an AVE of 0.50 or higher means that the construct explains more than half of the variance of its items. If AVE is less than 0.50, it means that, on average, more errors remain in the items than the variance explained by the construct (Hair et al. 2014).

The rules for outer loading testing are summarized as follows:

1. If outer loading is less than 0.40, delete the item.
2. If outer loading is greater than 0.40 but less than 0.70, then analyze the effect of deleting the item on AVE and composite reliability. Delete the item if deletion increases AVE and composite reliability above the threshold. However, if item deletion does not increase AVE and composite reliability above the threshold, retain the item.
3. If outer loading is greater than 0.70, retain the item.

Based on those criteria, researchers decide on the dimensions/factors to be included in the study. The next step is to name the items. The names given to the items must be related to the components they represent.

The items that researchers retain according to the rules of outer loading testing are then subjected to factor analysis validation. The purpose is to evaluate the generalizability and stability of the structure of data from the sample with a population (J.F. Hair et al., 2006).

Stage 3: Assessment of Structural Model
After confirming that the items in the construct are reliable and valid, researchers should assess the structural model. The structural model assessment procedure is performed in four steps as follows:

Assess the Significance and Relevance of the Structural Model Relationships
By running PLS-SEM algorithm, this study obtains the structural model relationships which indicate hypothesized relationships among constructs in the theoretical framework of the study. It shows path coefficients which have standardized values between -1 and +1. Estimated path coefficients close to +1 represents strong positive relationships. On the other hand, estimated path coefficients close to -1 represents strong negative relationships. If the estimated coefficients are closer to 0, it shows weaker relationships. If the estimated coefficients are very close to 0, the relationships are nonsignificant.

Assess Coefficient of Determination (R^2 and Adjusted R^2)
The coefficient of determination (R^2) is the most commonly used measure to evaluate structural models (Hair et al., (2014). R^2 is calculated as the squared correlation between a construct's actual and predicted values. It measures the model's predictive accuracy. The R^2 value ranges from 0 to 1. The higher the level of $R^{2,}$ the higher the levels of predictive accuracy. One problem with using R^2 as the model's predictive accuracy is that adding nonsignificant constructs to a structural model will increase R^2.

Therefore, using R^2 as a measure of the goodness of a model is not a good approach (Hair et al., 2014). This is because any addition of a nonsignificant construct to a structural model will always lead to higher R^2. Therefore, researchers should use the adjusted R^2 (R^2_{adj}) to avoid bias toward selecting models with many constructs. The formula for calculating the R^2_{adj} is given as follows:

$$R^2_{adj} = 1 - (1 - R^2).\frac{n-1}{n-k-1},$$

Where n is the sample size and k is the number of constructs. Therefore, the R^2_{adj} decreases R^2 by the number of constructs and the sample size and thus penalizes the addition of nonsignificant construct.

The Assessment of Mediating Effect

Researchers can follow the work of Baron & Kenny (1986), and Kenny et al (1998), to assess the mediating effect. There are four steps involved in assessing the mediating effect. In the first step, researchers assess whether there is a significant relationship between the independent and dependent variables. In the second step, researchers should assess whether there is a significant relationship between the independent and mediating variables. In the third step, researchers should assess whether the mediating variable is significantly related to the dependent variable when both the independent and mediating variables are predictors of the dependent variable. In the fourth step, researchers should determine that the coefficient relating the independent variable to the dependent variable must be larger (in absolute value) than the coefficient relating the independent variable to the dependent variable in the model, with both the independent variables and the mediating variable predicting the dependent variable. These steps of assessing mediation have been the most widely used method to assess mediation.

Assess the Effect Size f^2

The f^2 effect size measures the change in the R^2 when a specific exogenous construct is omitted from a model to evaluate whether the omitted exogenous construct has a substantive effect on the endogenous construct. The formula for the f^2 effect size is as follows:

$$f^2 = \frac{R^2 included - R^2 excluded}{1 - R^2 included}$$

Where $R^2_{included}$ and $R^2_{excluded}$ are the R^2 of the endogenous construct when a selected exogenous construct is included or excluded from the model. As a rule of thumb, f^2 of 0.02, 0.15 and 0.35 are considered as small, medium and substantive effects, respectively (Hair et al., 2014).

The Analysis of Moderating Variable

Moderating construct refers to the third construct which changes the relationship between two constructs from not significant or less significant to become significant relationship (Baron & Kenny, 1986). The significance of moderating variable could be tested in the structural equation model (Ping, 1995).

Based on Ping (1995), there are four steps which need to be taken as follows:

i. Estimate the model without the interacting variable.
ii. Calculate factor loading and residual variance for the interacting variable based on the first step.

iii. Construct moderating model by including factor loading and residual variance which was estimated in the second step, and
iv. Determine whether the moderating effect is significant.

Any construct is valid as a moderating construct if the interacting variable has a significant relationship with endogenous variables (Ping, (1995); Joseph F. Hair et al., (2010)). However, the moderating effect is not supported if the interacting variable has no significant relationship with the endogenous variables.

Conclusion

This paper has explained the complex methodology and practical uses of Partial Least Squares Structural Equation Modelling (PLS-SEM) in quantitative research. PLS-SEM is a powerful statistical tool that is highly effective in analyzing complex models. It is beneficial in both exploratory and predictive research situations. This paper thoroughly studied the data analysis process, including the key steps: data screening and diagnostic tests, measurement model assessment, and structural model assessment.

Data screening is a process that verifies the appropriateness of a dataset by dealing with missing data, outliers, and meeting multivariate criteria, including normality, linearity, and multicollinearity. The subsequent evaluation of the measurement model, using composite reliability and average variance extracted (AVE), verifies the constructs and affirms the reliability and validity of the data. The structural model assessment is a process that examines the importance and pertinence of connections between different elements, takes into account the influence of mediating and moderating factors, and gauges the predictive capability of the model using metrics such as R^2 and adjusted R^2.

By offering a structured guide for researchers, this paper underscores PLS-SEM's methodological rigour and practical utility in extending theories and estimating complex models. As the demands of contemporary research evolve, PLS-SEM provides a versatile and powerful approach for researchers aiming to derive meaningful insights from multifaceted datasets.

References

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations. *Journal of Personality and Social Psychology*, *51*(1), 173–182.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64–73.

Coakes, S. J., & Steed, L. G. (2003). *SPSS Analysis Without Anguish*. John Wiley & Sons Australia, Ltd.

Hair, J.F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis* (Sixth Edit). Pearson International Edition.

Hair, F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *The Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Hair, F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis A Global Perspective* (Seventh Ed). Pearson Global Edition.

Hair, Joseph F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modelling (PLS-SEM)*. SAGE Publications.

Ping, R. A. (1995). A Parsimonious Estimating Technique for Interaction and Quadratic Latent Variables. *Journal of Marketing Research*, *32*(3), 336–347.

Sekaran, U. (2003). *Research Methods for Business: A Skill-building Approach*. John Wiley & Sons.