Utilizing PLS-SEM in Archival Research: Considerations, Implications, and Practical Guidance

Anass Bayaga
Nelson Mandela University
anass.Bayaga@mandela.ac.za, South Africa, https//orcid/0000-0003-0283-0262

Abstract

This research explores the utilization of Partial Least Squares Structural Equation Modeling (PLS-SEM) in empirical archival research. PLS-SEM is a variant of SEM that has gained attention as an under-utilized method, particularly in the context of big data and secondary data analysis. The study highlights key considerations and implications when employing PLS-SEM in archival research. It emphasizes the importance of justifying the choice of PLS-SEM based on data characteristics and research goals, particularly when dealing with non-normal data distributions and limited theoretical foundations. The article provides practical guidance on model specification, data preparation, estimation, model evaluation, and reporting in PLS-SEM analysis. Additionally, the paper discusses the significance of reporting model fit indices, variable selection, and sample size estimation. The conclusion highlights the strengths and future research directions of PLS-SEM in archival research, including methodological advancements, integration with other statistical techniques, comparative studies, application in diverse fields, and the establishment of reporting guidelines. Overall, this research aims to enhance the understanding and application of PLS-SEM in the analysis of archival data, facilitating its wider adoption and contributing to empirical knowledge in various domains.

Keywords

Partial least squares; path modeling; structural equation modeling; archival data.

Introduction

The emergence and concurrent exponential growth in the processing of secondary or archival research (used interchangeably) and big data have propelled the demand for techniques related to logging, advanced processing, and analysis of data records of the types (Angeli, Howard, Ma, Yang & Kirschner, 2017; Buczak & Guven, 2017; Ogiela, Ogiela & Ko, 2020; Xu, Jiang, Wang, Yuan & Ren, 2017). For example, over the past two decades, Lyytinen (2009) has emphasised the significance of data in developing information systems theory. Similarly, Ioannidis (2005) highlighted the prevalence of false research findings. While there have been two types of responses thus far, Rigdon (2016) emphasises the importance of selecting an appropriate modelling technique as an analytical method.

Firstly, influenced by the exponential rise, various types of research studies have evolved. For instance, Xu et al. (2017) cautioned against the potential pitfalls of big data and its mining. Another example is Ogiela et al. (2020), who focused on intelligent data management in cloud computing. Meanwhile, in the domain of computer-assisted education, Angeli et al. (2017) questioned the extent to which data mining in educational technology classroom research could significantly contribute. Additionally, Buczak and Guven (2017) highlighted, in a survey on data mining and machine learning, the need for accurate communication survey methods.

Secondly, despite the increasing use of PLS-SEM, there has been limited exploration, particularly in understanding and mining secondary or archival data (Antonakis, Bendahan, Jacquart & Lalive, 2010; Chang, Franke & Lee, 2016; Hair, Risher, Sarstedt & Ringle, 2019; Rigdon, 2016; Hampton, 2015). PLS-SEM is gradually gaining popularity in modelling structural relationships, particularly latent constructs and their observed variables. However, several factors have hindered its significant impact on big data, data mining, or secondary data. There are several reasons (Chang et al., 2016; Hair et al., 2015). For instance, although PLS-SEM has traditionally been used to establish causal-predictive relationships, the technique heavily relies on criteria to assess the explanatory power of the path model. What makes it more rigorous is the comprehensive examination of several quality criteria required by the technique. Some of the quality criteria that need better understanding include, but are not limited to, model fit, PLSpredict, cross-validated predictive ability test (CVPAT), and model selection criteria.

Angeli et al. (2017), Buczak and Guven (2017), Ogiela et al. (2020), and Xu et al. (2017) have various implications. For instance, large data repositories and the accumulation of unprocessed data result in storage space wastage and the loss of confidential and vital information. Consequently, efforts are continuously being made to refine and improve knowledge discovery through various techniques associated with archival data and mining. In response, this paper presents a detailed application of one of the PLS-SEM techniques in the context of archival, big data, and data mining research. In addition to the challenges specific to PLS-SEM, a systematic review also explores the available causal prediction criteria for PLS-SEM in archival or secondary data. The study concurrently investigates the procedures for both causal prediction criteria available for PLS-SEM and secondary data. Although the focus is partly on exploring the role of causal prediction modelling in PLS-SEM, the overall objective is to apply the quality criteria by selecting the appropriate causal.

Research objective

Given that the recurrent theme is empirical archival data, the research questions are:

- What are the critical considerations and implications of using PLS-SEM in archival research?
- How can researchers effectively apply PLS-SEM in empirical studies with limited sample sizes and underdeveloped theoretical bases?
- What are the practical guidelines and reporting standards for utilizing PLS-SEM in archival research, including model specification, data preparation, estimation, model evaluation, and reporting of findings?

Literature review

Due to their potential for solving complex problems, secondary data/archival data and big data have found significant success in various fields, including business, engineering, social media, biological science, and cyber-security. These types of data are utilized to identify patterns in complex structural information and model complex biological systems in biomedicine. However, the scope and size of data have experienced substantial growth, collectively known as big data. Consequently, the integration of secondary data/archival data and big data for decision-making requires advanced processing techniques, prompting the need for the current study. There is a demand for modelling and improving data results to provide credible and widely accepted options for long-term research data planning, particularly with regard to archival data (Davick, 2014; Hair, Ringle, and Sarstedt, 2013; Rigdon, 2016).

For example, Davick (2014) explored the utilization and misapplication of Structural Equation Modelling (SEM) in long-term research data planning. Seminal work by Hair et al. (2013) and Hair, Sarstedt, Ringle, and Gudergan (2017) emphasized the necessity for rigorous and acceptable long-term planning through big data mining. Their assertion builds upon Hampton's (2015) work on estimating and reporting credible models concerning behavioral and accounting datasets. Meanwhile, Henseler, Dijkstra, Sarstedt, Ringle, Diamantopoulos, Straub, and Calantone (2014) cautioned against common misconceptions and realities related to the application of Partial Least Squares (PLS). This is crucial when researchers decide to report results pertaining to PLS-SEM. Additionally, studies have started addressing Ioannidis's (2005) concern regarding the prevalence of false research results or insufficient statistical inference (Anderson & Gerbing, 1988; Aguirre-Urreta et al., 2018).

Over the past decades, there has been an increasing number of studies focusing on the use of PLS in various contexts, disciplinary applications, and assessing PLS quality and measurement (Becker, Rai & Rigdon, 2013; Chang, Franke & Lee, 2016; Davick, 2014; Hair, Hampton, 2015; Henseler, Dijkstra, Sarstedt, Ringle, Diamantopoulos, Straub & Calantone, 2014; Hinson & Utke, 2018; Lee, Petter, Fayard & Robinson, 2011; Rigdon, 2016; Reinartz, Haenlein & Henseler, 2009; Sarstedt, Ringle & Gudergan, 2017). For instance, Hinson and Utke (2018) employed Structural Equation Modelling (SEM) to study archival capital markets, while both Lee et al. (2011) and Licerán-Gutiérrez and Cano-Rodríguez (2020) used PLS to investigate archival accounting research, focusing on measuring earnings quality. Similarly, Becker et al. (2013) examined the predictive validity and formative measurement using SEM. Empirical comparisons between covariance-based and variance-based SEM have also been conducted (Reinartz et al., 2009). Chang et al. (2016) compared reflective and formative measures through simulations. Rigdon (2016) investigated PLS path modelling as an analytical method, while Davick (2014) explored the correct use of SEM in management research. These previous examinations and applications have shown that PLS-SEM yields superior results and higher acceptance (Hair et al., 2013; Hair et al., 2017).

However, existing literature lacks comprehensive exploration of the various forms in which PLS is used in archival data and big data (Licerán-Gutiérrez & Cano-Rodríguez, 2020). Consequently, despite ongoing research on PLS-SEM, several questions remain unanswered. For instance, the role of secondary data and archival material in PLS-SEM is not fully understood, nor do we have a complete understanding of when and why to utilize secondary data in PLS-SEM. While some researchers have started investigating the use of PLS-SEM (Licerán-Gutiérrez & Cano-Rodríguez, 2020), there are still several under-researched questions. These include, but are not limited to, accessing and using published or archival datasets, comprehending big data sets, and understanding the integration of secondary data within the PLS technique. In light of these unanswered questions, the present study aims to examine the application of PLS in archival education research.

Use and drawbacks of secondary data and archival material in PLS-SEM

The utilization of secondary data and archival material in PLS analysis depends on their nature and characteristics. Primary data is collected directly through interactions with participants or sources, such as interviews, focus groups, surveys, or participant observation. In contrast, secondary data refers to existing data collected for a specific purpose, without direct involvement of the researcher. Primary resources are firsthand accounts or direct evidence of historical data, typically collected during the event and not subjected to secondary analysis or interpretation. Examples include official surveys on education, economic reviews, labor market data, or general household surveys.

However, there are drawbacks to using secondary data. These include the cost of acquiring the dataset, the researcher's familiarity with the data, the potential mismatch between the data and the research question, gaps in the data, or data collected for a different purpose. Moreover, measures may not be directly comparable, and researchers have limited control over data quality in terms of rigor and reliability. Commercially sensitive data may also be challenging to access from company archives, departments, or intranets. Different types of secondary data include censuses, repeated surveys, ad hoc surveys, and time series data, each with their own research considerations such as relevance, reliability, and currency.

Archival or secondary documentary data and records are valuable resources left as a byproduct of everyday activities, and they are utilized by historians and business researchers alike. They provide insight into management decisions beyond the scope of interviews and allow for a historical perspective in research. Archival research can be used to triangulate with other qualitative methods or as exploratory research prior to a main study. Examples of archival data include organizational records (e.g., human resources, accounts, and payroll data), sales data, project files, correspondence, meeting minutes, reports, diaries, sales literature, non-textual materials (e.g., maps, videos, photographs), and data held in management information systems (MIS) related to recruitment and management training.

In summary, secondary data and archival material play a vital role in PLS analysis, offering researchers access to existing information and historical perspectives. However, careful consideration should be given to their relevance, reliability, and compatibility with the research question.

Processing of archival data through PLS

With regards to processing and analyzing archival data, additional techniques such as moderation and mediation have become necessary. Moderation and mediation aim to describe causal relationships within the data. The moderator variable strengthens or weakens the cause-effect relationships, while the mediator variable acts as a third variable that influences the cause-effect relationship through an intermediary process. In multivariate regression, interaction terms are commonly used to estimate moderation effects, unlike in PLS-SEM where the sums approach is used. For mediating effects, separate multi-step processes like the Barron and Kenny method can be employed. Although these techniques are beyond the scope of the current study, it is worth noting that PLS-SEM does not strictly require a larger sample size, especially in relatively new research fields (minimum sample size should be 200). The decision to use PLS-SEM is influenced by various factors, including exploratory or predictive research objectives, non-normality of data distribution, analysis of constructs (both formative and reflective), the number of interaction terms, and the inclusion of mediated models. SmartPLS3 software is commonly used for analyzing the moderating effect of a latent variable in PLS-SEM noting that Table 0 shows both first and second-generation statistical tools, along with their associated technologies or software.

Table 0: Examples of first/second-generation statistical tools, along with their technologies (own Table)

1st-Generation Statistical Tools	Technologies/Software	2nd-Generation Statistical Tools	Technologies/Software
T-tests	Statistical software (e.g., SPSS)	Structural Equation Modeling (SEM)	SEM software (e.g., AMOS, SmartPLS)
Analysis of Variance (ANOVA)		Bayesian Statistics	Bayesian software (e.g., JAGS, Stan)
Chi-square test		Machine Learning	Python (scikit-learn, TensorFlow)
Regression analysis		Data Mining	Data mining software (e.g., RapidMiner, KNIME)
Pearson correlation		Deep Learning	Python (Keras, PyTorch)
Mann-Whitney U test		Multilevel Modeling	Statistical software (e.g., R, Mplus)
Kruskal-Wallis test		Latent Class Analysis	Statistical software (e.g., R, Mplus)
Analysis of Covariance (ANCOVA)		Geospatial Analysis	Geographic Information System (GIS) software (e.g., ArcGIS, QGIS)
Time series analysis		Text Mining	Text mining software (e.g., Python NLTK, R tm)
Factor analysis		Network Analysis	Network analysis software (e.g., Gephi, Pajek)
Student's t-test			
Wilcoxon signed-rank test			
Friedman test			
Log-linear analysis			
Discriminant analysis			
Survival analysis			
Cluster analysis			
Probit regression			
Logit regression			
Poisson regression			

In the context of archival data, other important considerations for PLS-SEM include internal consistency, reliability assessment, composite reliability, indicator reliability, convergent validity, average variance extracted (AVE), heterotrait-monotrait ratio (HTMT), evaluation of the inner model fit, predictive relevance (Q2), coefficient of determination (R2), and standardized root mean square residual (SRMR). These factors contribute to the understanding and assessment of PLS features when applied to archival data (Garson, 2016).

PLS features in archival data

While this section does not aim to provide an exhaustive review of the extensive literature on PLS-SEM, it is important to address various considerations. Table 1, derived from Henseler et al. (2014, p. 2), provides an overview of the evolving understanding of PLS-SEM, encompassing both traditional and contemporary perspectives. For a comprehensive examination of PLS features, readers are encouraged to consult other works such as Dijkstra and Henseler (2012), Garson (2016), Hair et al. (2019), Bentler and Yuan (1999), Becker et al. (2013), and Reinartz et al. (2009).

Dijkstra and Henseler (2012) focused on investigating consistent and asymptotically normal PLS estimators for linear structural equations, while Garson (2016) explored PLS through regression and SEM methodologies. Hair et al. (2019) provided insights on when and how to report the results of PLS-SEM more generally. Additionally, studies of significance include Bentler and Yuan's (1999) examination of SEM with small samples and its implications for test statistics in behavioral research. Becker et al. (2013) explored topics such as predictive validity and formative measurement, while Reinartz et al. (2009) conducted a comparative analysis of the effectiveness of covariance-based and variance-based SEM.

For a deeper understanding of these topics, readers are encouraged to refer to the aforementioned works, as they provide valuable insights into the theoretical and methodological aspects of PLS-SEM.

Table 1. Significant changes in understanding PLS-SEM (Source: Henseler et al., 2014, p. 2).

Traditional standpoint	**Current view**
1. Primarily for exploratory as well as early-stage research	Various types of research, including but not limited to confirmatory, explanatory, or predictive is, applicable
2. Advantageous than covariance-based (CB-SEM) with small size	Although generally produced, estimates with small sizes can be far less accurate when large sample sizes are used. Accordingly, the justification for a small sample size must be cautiously considered.
3. Only used in estimating recursive structural models	With the use of two or three -stage-least square (2SLS or 3SLS) instead of ordinary least square (OLS), Dijkstra and Henseler (2015b) opine it could equally be used in estimating non-recursive structural models
4. When using PLS-SEM, model identification is not needed	Generally, PLS is used in estimating the underlying composite model. Whether PLS is composed of latent variables, composite model rules still need to be applied. Accordingly, model identification remains paramount.
5. Unlike the maximum-likelihood (ML) estimator, PLS does have greater statistical power	Sadly, the assertion is rooted in inconsistent parameter estimates and has been revealed as inaccurate or invalid. Additionally, estimators do not have statistical power. Instead, it is prudent to refer to efficiency. One could also refer to its accuracy when estimating parameters customarily expressed in standard error terms. Instead, only a statistical test is assessed in terms of its statistical power
6. Mode types say 'A' in PLS is consistently used in estimating reflective measurement models	Irrespective of the mode applied, PLS can create linear combinations in the form of observed indicators. These form proxies in theoretical concepts. For that reason, Dijkstra and Henseler (2015b), for estimating models involving latent variables, correction of attenuation of construct scores correlations are essential- such procedure is termed as PLSc.
7. Mode type say 'B' in PLS is consistently used in estimating causal–formative measurement models	Mode B is used in obtaining weights to build composites. Thus, only sometimes used in estimating causal-formative measurement models. Nevertheless, causal-formative measurement models may be estimated through the MIMIC model.
8. Usually, the estimated overall fit of models for PLS is not assessable	Dijkstra and Henseler (2015b) indicated two non-exclusive forms of assessment. The first is bootstrap-based tests used for overall model fit. The second is measures of overall model fit. Generally, both are used to assess the discrepancy between empirical and the model-implied indicator variance called covariance matrix. However, whereas the latter is rooted in heuristic rules, the former is grounded on statistical inferences.
9. Reliability of construct scores obtained by PLS-SEM should be assessed using the two fundamental forms of assessing reliability constructs Cronbach's α as well as Dillon Goldstein's ρ, sometimes referred to as Jöreskog's ρ, or even composite reliability	Dijkstra-Henseler's ρ_A tests the reliability coefficient for PLS construct scores consistently. While Dillon-Goldstein's ρ, as well as Cronbach's α, do indicate the reliability of sum scores. Meanwhile, Cronbach's α is anchored on the indicator variance-covariance matrix. On the other hand, Dillon-Goldstein's ρ is rooted in

	the factor loadings. Consequently, in estimating Dillon-Goldstein's ρ, consistent factor loading estimates are used. Likewise, Cronbach's α does assume equal population covariances within the indicators of a single block - an assumption that is not frequently met in empirical data. Nevertheless, it can be used as a lower bound for reliability
10. Fornell-Larcker criterion is used to examine discriminant validity	HTMT should be considered in assessing discriminant validity

Nevertheless, a recap or summary of PLS features in archival data suggests that while internal consistency reliability is assessed using Cronbach's alpha, its threshold must be > 0.8. Meanwhile, composite reliability, which measures the sum of the latent variable's factor loading relative to the sum of the factor loadings plus error variance, should have a threshold > 0.6. Finally, the indicator reliability assesses' the contribution of the indicators to latent variables with values >0.6. As reflected in Table 2, the convergent validity through the average variance extracted has a value > 0.5.

Table 2: Threshold with cut-off points or values and model modification.

Categories	Fitness index	Recommended values
Absolute Fit Index	Chi-square	P > 0.05
	Root mean square error approximation (RMSEA)	< 0.08
	Goodness of Fit	> 0.90
Incremental Fit Index	Adjusted goodness of Fit (AGFI)	> 0.90
	Comparative Fit Index (CFI)	> 0.90
	Tucker-Lewis Index (TLI)	> 0.90
	Normed Fit Index (NFI)	> 0.90
Parsimonious Fit Index	Chi-square/degree of freedom	> 5.0

The evaluation of model fit and validity criteria plays a crucial role in assessing the strength and significance of the hypothesized relationships in PLS-SEM. Several factors are considered in this process.

Convergent validity, which examines the extent to which individual items reflect construct coverage, is typically assessed using average variance extracted (AVE). AVE values higher than 0.5 indicate good convergent validity. Additionally, the heterotrait-monotrait ratio (HTMT) assesses the correlations between indicators measuring different constructs versus indicators within the same construct. Threshold values below one indicate acceptable discriminant validity.

Inner model fit is evaluated through various measures. The coefficient of determination (R2) represents the proportion of variance in the dependent variable explained by the independent variables. An R2 value above 0.1 is considered substantial. Predictive relevance (Q2) is assessed through blindfolding, which measures the model's predictive strength. Q2 values of 0.02, 0.15, and 0.35 indicate small, medium, and considerable predictive relevance, respectively. The standardized root mean square residual (SRMR) is used to evaluate model fitness, with values below 0.1 indicating good fit.

Significance tests and p-values for path coefficients are estimated through bootstrapping. Critical t-values for two-tailed tests are typically set at 1.65 (10% significance level), 1.96 (5% significance level), and 2.58 (1% significance level).

In conclusion, the evaluation of PLS-SEM models involves assessing various validity criteria. This analysis provides valuable insights into the strength and significance of the hypothesized relationships, contributing to a better understanding of the model under investigation.

Discussions

When to use PLS-SEM in archival-based research

In empirical research, PLS-SEM (a variant of SEM) has been identified as an under-utilized method, particularly in the context of big data, archival data, and secondary data. Both multivariate regression and PLS-SEM aim to maximize the prediction of raw scores and are predictive-oriented methods. Therefore, PLS-SEM can be seen as a viable alternative to multivariate regression in predicting causal relationships. However, it is crucial to justify the choice of PLS-SEM for estimating cause-effect relationships based on data characteristics and research considerations. Table 1 provides an overview of PLS-SEM, highlighting its suitability for empirical-historical data that may not meet the parametric assumptions of multivariate regression, such as normality. Since archival data is often documented historical data, it is challenging to alter or enhance it to meet these assumptions. Nevertheless, logarithm transformation is commonly employed to correct skewed distributions of variables in archival data. This transformation helps address biased estimates and interpretation issues that may arise in multivariate regression. PLS-SEM is preferred in part because it is a non-parametric technique while still adhering to the predictive-oriented methods of multivariate regression. Recent studies have shown that PLS-SEM estimation remains robust even with skewed data and thus tend to respond variety of questions and themes as illustrated in Table 3. By addressing these research questions and exploring the key themes, the selected sources contribute to the theoretical understanding of the application and interpretation of PLS-SEM.

Table 3: Research questions and exploring the key themes, the selected sources contribute to the theoretical understanding

Source	Research Questions Addressed	Key Themes and Questions
Hair et al. (2019)	How can PLS path modeling be used as an analytical method in the context of E-learning?	- Application of PLS path modeling in E-learning
Rigdon (2016)	What are the advantages and limitations of PLS path modeling as an analytical method in E-learning?	- Advantages and limitations of PLS path modeling in E-learning
Reinartz et al. (2009)	How does covariance-based SEM compare to variance-based SEM in the context of E-learning research?	- Comparison of covariance-based SEM and variance-based SEM in E-learning research
Dijkstra & Henseler (2012)	What are the properties and estimation methods of consistent and asymptotically normal PLS estimators?	- Properties and estimation methods of consistent and asymptotically normal PLS estimators
Garson (2016)	How can PLS regression and SEM be applied in the analysis of E-learning data?	- Application of PLS regression and SEM in analyzing E-learning data
Hair et al. (2019)	When should PLS-SEM be used, and how should its results be reported in E-learning research?	- Guidelines for using PLS-SEM
Bentler & Yuan (1999)	How can SEM be applied with small samples and test statistics in the context of behavioral E-learning research?	- Application of SEM with small samples and test statistics in behavioral E-learning research
Becker et al. (2013)	What is the predictive validity of different measurement models in E-learning research?	- Evaluation of predictive validity of different measurement models
Henseler et al. (2014)	What are the significant changes and advancements in the understanding of PLS-SEM in E-learning research?	- Significant changes and advancements in the understanding of PLS-SEM
Ringle et al. (2018)	How can the coefficient of determination, predictive relevance, and path coefficients be interpreted in PLS-SEM?	- Interpretation of coefficient of determination, predictive relevance, and path coefficients in PLS-SEM

This indicates that PLS-SEM has higher statistical power in estimating non-linear relationships among predictor constructs in empirical studies. Unlike multivariate regression, which requires normally distributed data, PLS-SEM allows for modeling non-linear terms, even in heterogeneous data. Additionally, PLS-SEM accommodates the inclusion of single-item constructs, along with multi-item constructs, providing more flexibility compared to covariance-based SEM (CB-SEM). PLS-SEM is particularly suitable for archival data as it helps conceptualize precise and concrete attributes. It facilitates theory-building by creating new constructs, including unobservable variables and structural paths in theoretical models. This feature is beneficial for incremental studies that lead to theory testing. Unlike multivariate regression, PLS-SEM enables confirmatory factor analysis on measurements with unobserved variables. Multivariate regression assumes the absence of measurement error when considering such measures on unobserved variables, leading to biased estimates. Thus, PLS-SEM outperforms multivariate regression in this aspect. PLS-SEM not only examines the combined effects of observed variables but also

simultaneously conducts confirmatory factor analysis to ensure that the observed variables exhibit the same attribute. In contrast, multivariate regression is limited to one dependent variable, while PLS-SEM is a more suitable technique for testing moderation and mediation in complex models. PLS-SEM can estimate multiple dependent variables (endogenous constructs) more effectively in complex models compared to multivariate regression. Furthermore, PLS-SEM offers a more straightforward approach to mediation analysis, as it can estimate mediation effects in a complete model, whereas multivariate regression requires multiple steps to estimate both moderating and mediating effects. Some PLS-SEM software can generate direct, indirect, and total effects in models in a single analysis, facilitating moderation effects analysis even with multiple dependent variables in a complex model. PLS-SEM can be used as an empirical method for examining moderating effects in various ways. For instance, it models interaction terms for non-normally distributed variables. In cases where data fail to meet normality assumptions, logarithm transformation can be applied in PLS-SEM before modeling interaction terms. This approach helps mitigate scaling issues, measurement errors, and biased estimates. Moreover, PLS-SEM allows for multigroup analyses, which involve examining systematic differences between parameters for different groups. Overall, PLS-SEM presents valuable advantages for analyzing empirical data, particularly in the context of archival data, providing insights into complex relationships and accommodating non-linear and non-normally distributed variables.

Standard of reporting archival data in PLS-SEM

The present study focuses on accessing and utilizing published or archival data sets and understanding secondary data through the PLS-SEM technique, driven by the growing interest in big data and its mining.

One important implication for discussion is the need to go beyond reporting model fit indices and results and also explore sample size and model validation. It is crucial to report model fit indices in PLS-SEM, and key indicators include $\chi 2$, CFI, RMSEA, TLI, GFI, NFI, SRMR, AIC, and BIC (Table 2). The first and second implications emphasize the importance of considering the choice and usage of fit indices in line with the study's objectives. Researchers should not overlook GFI and NFI as essential indicators of model performance and fit, despite their tendency to be neglected. However, it is important to address the different properties and sensitivity of these fit indices to various factors, such as data distribution, missing data, model size, and sample size, as cautioned by Barrett (2007). The third implication stems from the first two and suggests that theoretical support should correspond to most fit indices. Variable selection is a crucial aspect of SEM analysis, and researchers need to justify whether the selected variables effectively represent the phenomenon under investigation. In PLS-SEM analysis using archival data, indices like CFI, RMSEA, and SRMR help detect model misspecification and assess relative fit, while AIC and BIC (as shown in Table 2) are primarily used for model selection, explaining the model's quality in terms of type, structure, and hypotheses.

Considering these implications, reporting results involves two key steps: reporting estimates and the modeling process. Users must thoroughly describe the results of hypothesis tests and include fundamental indices such as p-values, R square, and standard errors as overall fit indices, which indicate the validity and reliability of each path. These indices also provide evidence in cases of poor overall fit. Following the reporting standards outlined by the American Psychological Association (APA) is generally required, encompassing five key steps: model specification, data preparation, PLS-SEM estimation, model evaluation and modification, and reporting of findings. In reporting the model specification process, researchers should provide information on theoretical plausibility, positive or negative direct effects of variables, data sampling method, sample size, and model type. The data preparation process should include an assessment of normality, analysis of missing data and methods used for handling it, as well as transformations. Estimating SEM requires reporting the input matrix, estimation method, software brand and version, and procedures for scaling latent variables. Model evaluation and modification involve reporting fit indices, cut-off points or values, and any necessary model modifications mentioned in Table 2. Reporting findings should encompass latent variables, including factor loadings, standard errors, p-values, R square, standardized and unstandardized structure coefficients, and graphical representations of the model. Sample size estimation and reporting play a crucial role, with recommendations from various authors suggesting different approaches. Traditionally, sample size varies based on fit indices, model size, and variable distribution. Factors such as the amount of missing data, variable reliability, and strength of path parameters are also important considerations. While the general recommendation is often 100-200, others propose five cases per parameter in the model (Tabachnick & Fidell, 2001). However, specific fit indices or power analysis of the model should guide sample size determination, and caution should be exercised in applying general rules. Monte Carlo simulation and equations proposed by Kim (2005) can aid in calculating sample size based on model fit indices and statistical power. Model validation, although less common, serves as evidence for the hypothetical model and involves testing the model on two or more random datasets with a large sample size. This validation process

ensures similarity of parameters across different datasets from the same population, especially when the model is developed based on various datasets.

Addressing evidence of causal relationships and variable selections in archival data

Selecting an appropriate model and variables based on the research goal is a fundamental aspect of statistical analysis. This section aims to provide guidance to practitioners and researchers on using Partial Least Squares Structural Equation Modeling (PLS-SEM) in archival research. Understanding the purpose of PLS-SEM is crucial for researchers to effectively apply this methodology. One of the key reasons for using PLS-SEM in archival research is when the data and research objectives are tied to an underdeveloped theoretical foundation. In such cases, researchers may have limited prior knowledge of the causal relationships among constructs. Another important reason to opt for PLS-SEM is when the sample size is limited, as is often the case in non-parametric analyses. With a limited sample size, it becomes challenging to make specific assumptions about the distribution of the data or handle missing data. Nonetheless, researchers can confidently use PLS-SEM to test causal relationships even in situations with constrained sample sizes and limited theoretical support (Hair et al., 2013).

Considering these constraints, PLS-SEM becomes a viable option due to its algorithm based on maximum likelihood estimation. It is generally recommended to start with a smaller dataset during the initial stages and then apply PLS-SEM. This approach helps generate sufficient and necessary evidence to assess causality, select variables, and establish causal relationships. Monecke and Leisch (2012) recommend this selection process as it allows for collecting long-term data and updating hypotheses. Their recommendations are based on the centrality of evidence for establishing causal relationships. While various approaches are available, one crucial step is specifying the causal relationships and correlations among the constructs. Shipley (2002) emphasizes the importance of justifying these causal relationships and correlations, along with theoretical underpinnings, as their absence weakens the hypotheses' causal claims. To address this challenge, Bollen and Pearl (2013: 304) suggest (1) imposing zero coefficients and (2) imposing zero covariance on the model. According to Bollen and Pearl (2013: 304), strong causal assumptions ultimately assign specific parameter values to those relationships.

Theoretical and practical implication

The theoretical and practical implications of the above considerations for using PLS-SEM in archival research are as follows:

Addressing underdeveloped theoretical foundations: PLS-SEM offers a valuable approach when there is limited prior knowledge or an underdeveloped theoretical base. Researchers can confidently apply PLS-SEM to explore causal relationships even in the absence of strong theoretical support. This enables the investigation of research questions and the generation of empirical evidence to build theoretical foundations.

Overcoming sample size limitations: PLS-SEM is particularly suitable for situations with limited sample sizes. Non-parametric requirements and the algorithm of PLS-SEM, based on maximum likelihood, make it a robust technique for analyzing data with small sample sizes. Researchers can employ PLS-SEM to derive meaningful insights from their data, even when traditional statistical approaches may not be applicable.

Flexibility in data distribution and missing data: PLS-SEM does not impose strict assumptions about data distribution, making it well-suited for analyzing archival data that may not meet normality assumptions. Additionally, PLS-SEM provides flexibility in handling missing data, allowing researchers to address data gaps effectively. This flexibility enables researchers to utilize valuable archival data without discarding observations due to missing values.

Importance of model specification and justification: Proper model specification and theoretical justification are crucial in PLS-SEM. Researchers must clearly define causal relationships and correlations among constructs and provide theoretical support for these relationships. This helps strengthen the validity and reliability of the research findings and enhances the overall quality of the study.

Model validation using multiple datasets: Model validation plays a vital role in ensuring the robustness and generalizability of findings in archival research. Researchers can employ techniques such as testing the model on multiple random datasets to validate the model's performance. This validation process requires a large sample size to assess the consistency of parameters across different datasets.

Overall, the practical implications emphasize the usefulness of PLS-SEM in addressing the challenges associated with underdeveloped theoretical foundations, limited sample sizes, and analyzing archival data. By leveraging the flexibility and robustness of PLS-SEM, researchers can gain valuable insights from their data and advance empirical knowledge in their respective fields.

Conclusion and future research

In conclusion, PLS-SEM offers a valuable methodological approach for conducting empirical research with archival data. It provides researchers with a flexible and robust technique to explore causal relationships, even in situations with limited theoretical foundations and small sample sizes. PLS-SEM's ability to handle non-normal data distributions and missing data further enhances its applicability in analyzing archival datasets. By following best practices in model specification, data preparation, and reporting, researchers can ensure the validity and reliability of their findings.

However, there are several avenues for future research in using PLS-SEM in archival research:

Methodological advancements: Further methodological developments can focus on refining the PLS-SEM technique for analyzing archival data. This could include addressing specific challenges associated with handling large and complex datasets, developing techniques for handling missing data more effectively, and improving model validation procedures.

1. Integration of PLS-SEM with other statistical techniques: Exploring the integration of PLS-SEM with other statistical techniques can enhance the analysis of archival data. This may involve combining PLS-SEM with advanced multivariate techniques or integrating it with machine learning approaches to gain deeper insights from complex datasets.
2. Comparative studies: Comparative studies that compare the results obtained from PLS-SEM with other statistical techniques commonly used in archival research can provide insights into the strengths and limitations of PLS-SEM. This can help researchers determine the most appropriate analytical approach for their specific research questions and data characteristics.

By addressing these future research directions, researchers can further enhance the application of PLS-SEM in archival research, broaden its scope of application, and contribute to the advancement of empirical knowledge in various disciplines.

References

Aguirre-Urreta, M. I. and Rönkkö, M. Statistical inference with PLSc using bootstrap confidence intervals. *MIS Quarterly,* vol. 42, no. 3, pp. 1001–1020. https://doi.org/10.25300/misq/2018/13587, 2018

Angeli, C., Howard, S. K., Ma, J., Yang, J. and Kirschner, P. A. Data mining in educational technology classroom research: can it contribute? *Computers & Education*, vol. 113, pp. 226–242, 2017.

Anderson, J. C. and Gerbing, D. W. Structural equation modeling in practice: A review and recommended two- step approach. *Psychological Bulletin*, vol. 103, 3, pp. 411–423. https://doi.org/10.1037/0033- 2909.103.3.411, 1988

Antonakis, J., Bendahan, S., Jacquart, P. and Lalive, R. On making causal claims: A review and recommendations. *The Leadership Quarterly*, vol. 21, no. 6, pp. 1086-1120. https://doi.org/10.1016/j.leaqua.2010.10.010, 2010

Barrett, P. Structural equation modeling: Adjudging model fit. *Personality and Individual Differences,* vol. 42, no. 5, 815–824. https://doi.org/10.1016/j.paid.2006.09.018, 2007

Becker, J. M., Rai, A. and Rigdon, E. E. "Predictive Validity and Formative Measurement in Structural Equation Modeling: Embracing Practical Relevance." Thirty Fourth International Conference on Information Systems, pp. 1–19. 2013

Bentler, P. M. and Yuan, K. H. Structural equation modeling with small samples: Test statistics. *Multivariate behavioral research*, 34(2), 181–197. https://doi.org/10.1207/s15327906mb340203, 1999

Bollen, K. A. A new incremental fit index for general structural equation models. *Sociological Methods & Research*, vol. 17, no 3, pp. 303–316. https://doi.org/10.1177/0049124189017003004, 1989

Bollen, K. and Pearl, J. "Eight Myths About Causality and Structural Models" In S.L. Morgan (Ed.), Handbook of Causal Analysis for Social Research, pp. 301–328, *Springer,* 2013

Buczak, A. and Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, vol 18, no. 2, pp.1153-1176. 2017

Chang, W., Franke, G. R. and Lee, N. Comparing reflective and formative measures: New insights from relevant simulations. *Journal of Business Research,* vol. 69, no. 8, pp. 3177–3185. 2016

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J. and Kirby, J. B. Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, vol. 29, no. 4, pp. 468–508. https://doi.org/10.1177/0049124101029004003, 2001

Davick, N. S. The use and misuse of structural equation modeling in management research: A review and critique. *Journal of Advances in Management Research,* vol. 35, no. 2, pp. 441–458. 2014

Dijkstra, T. K. and Henseler, J. Consistent and asymptotically normal PLS-estimators for linear structural equations (Working paper). Retrieved from http://www.rug.nl/staff/t.k.dijkstra/research, 2012.

Dijkstra, T. K. and Henseler, J. Consistent partial least squares path modeling. *MIS Quarterly,* vol. 38, no. x, pp. 1-26, 2015b

Garson, G. D. Partial Least Squares: Regression & Structural Equation Models, Statistical Associates Publishing. 2016

Goodhue, D., Lewis, W. and Thompson, R. Statistical power in analyzing interaction effects: Questioning the advantage of PLS with product indicators. *Information Systems Research,* vol. 18, pp. 211–227. 2007

Hair, J. F., Risher, J. J., Sarstedt, M., and Ringle, C. M. When to use and how to report the results of PLS-SEM. *European Business Review*, vol 2, no. 3, pp. 34-56. https://doi.org/10.1108/ebr-11-2018-0203, 2019

Hair, J. F., Ringle, C. M. and Sarstedt, M. Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results, and Higher Acceptance. *Long Range Planning*, vol. 46, no. 2, pp. 1–12. 2013

Hair, J. F., Sarstedt, M., Ringle, C. M. and Gudergan, S. P. Advanced Issues in Partial Least Squares Structural Equation Modeling (2nd ed.). Thousand Oaks, CA, CA: Sage. 2017

Hampton, C. Estimating and Reporting Structural Equation Models with Behavioural Accounting Data. *Behavioural Research in Accounting*, vol. 27, no. 2, pp. 1–34. 2015

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T. M. and Calantone, R. J. Common Beliefs and Reality About PLS: Comments on Rönkkö and Evermann (2013). *Organizational Research Methods*, vol. 17, no. 2, pp. 182–209. https://doi.org/10.1177/1094428114526928, 2014

Hinson, L.A. and Utke, S. Structural Equation Modeling in Archival Capital Markets Research: An Empirical Application to Disclosure and Cost of Capital. *Financial Accounting eJournal*. 2018

Ioannidis, J. P. Why are most published research findings false? *PLoS Medicine*, vol 2, no. 8, pp. 124-138. 2016

Kante, M., Chepken, C. and Oboko, R. Partial Least Square Structural Equation Modelling' use in Information Systems: An Updated Guideline of Practices in Exploratory Settings. *Kabarak Journal of Research & Innovation*, vol. 6, pp. 67–79. 2018

Lee, L., Petter, S., Fayard, D. and Robinson, S. On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems*, vol. 12, no. 4, pp. 305-328, 2011.

Licerán-Gutiérrez, A. and Cano-Rodríguez, M. Ana Licerán-Gutiérrez & Cano-Rodríguez Using partial least squares in archival accounting research: an application to earnings quality measuring. *Spanish Journal of Finance and Accounting,* vol, 2, 143-170. https://doi.org/10.1080/02102412.2019.1608705, 2020

Lyytinen, K. Data matters in IS theory building. *Journal of the Association for Information Systems,* vol 10, no. 10, 715-720, 2009

Monecke, A. and Leisch, F. semPLS: Structural Equation Modeling Using Partial Least Squares. *Journal of Statistical Software*, vol. 48, pp. 1–32. ttps://doi.org/10.18637/jss.v048.i03 , 2012.

Ogiela, L., Ogiela, M. R. and Ko, H. Intelligent data management and security in cloud computing. *Sensors*, vol 20, no. 12, pp. 3458–3469, 2020

Reinartz, W., Haenlein, M. and Henseler, J. An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing,* vol 26, no. 4, pp. 332–344. 2009

Rigdon, E. E. Choosing PLS path modeling as the analytical method in European management research: A realist perspective. *European Management Journal.* Vol 1, pp. 1–8. 2016

Shipley, B. Start and Stop Rules for Exploratory Path Analysis. Structural Equation Modeling: A Multidisciplinary Journal, vol. 9, no. 1, pp. 554–561. 10.1207/S15328007SEM0904_5. 2002

Xu, L., Jiang, C., Wang, J., Yuan, J. and Ren, Y. Information security in big data: privacy and data mining. *IEEE Access*, vol. 2, no. 2, pp. 1149–1176. 2017

Biography

Anass Bayaga is currently a cognitive mathematics and STEM cognition professor at the Nelson Mandela University. Presently, Anass serves as an editorial board member of international journal of mathematics teaching and learning (IJMTL). Prof Bayaga currently serves as member of the Membership committee of the Mixed Methods International Research Association - MMIRA. He was also a member of the International Institute of Informatics and Systemics (IIIS). Bayaga was also a Fulbright researcher at the George Washington University, where through predictive modeling, he previously researched cognitive enhancement via mobile computing and applications in STEM, which is also presently his research group's niche/focus. His research and teaching interests are Mathematics and computational cognition (Neuro-mathematics (STEM), STEM cognitive enhancement via Human-computer interaction, and Predictive/mathematical modelling.