The Use of Partial Least Squares Path Modeling in Causal Inference for Archival Financial Accounting Research

Chin Fei Goh*, Mohamad Bilal Alib, Amran Rasliab

*Faculty of Management, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
bFaculty of Education, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: cfgoh2@live.utm.my

1.0 INTRODUCTION

Multivariate regression has been commonly used in financial accounting research especially in identifying theoretical prediction of variables [1]. Researchers in the arena of financial accounting research often have to resolve econometric problems such as non-normality of data, endogeneity, measurement and others to ensure the validity of causal inference in empirical studies [2-5].

There is also a steadily increasing trend in using the structural equation modeling to perform causal inference in the archival financial accounting research [6-9]. However, multivariate regression is still the predominant statistical method. In this regard, Lee et al., argue that the inadequate knowledge of Partial Least Squares-Structural Equation Modeling (PLS-SEM) among financial accounting researchers could explain why they have been reluctant to use PLS-SEM in empirical studies [1]. In a similar vein, the authors perceive that the lack of PLS-SEM in financial accounting causal inferences is surprising because the algorithms of both multivariate regression and PLS-SEM are optimized to maximize the explained variance of dependent variable [1; 10].

The authors also recognize that the uniqueness of archival data that are not obtained through entirely random experiment in which they are often regarded as observed variables [3; 11]. A good example is that financial statements published by publicly listed companies are generally verified by qualified auditors and endorsed by board of directors. Thus, researchers are unlikely able to enhance or alter the archival data during data preprocessing.

The goal of this article is to illustrate how PLS-SEM can be used for archival financial accounting research. This article also answers the call from Lee et al., to investigate potential benefits of using PLS-SEM in archival financial accounting research [1]. This article will first compares multivariate regression and structural equation modeling. It followed by describing the difference between covariance-based SEM (CB-SEM) and PLS-SEM in order to present an overview of such methods to readers. The article will then reviews common statistical issues in archival financial accounting research. Subsequently, the guidelines of using PLS-SEM based on data characteristic and research design are presented. The final section provides a conclusion.

2.0 MULTIVARIATE REGRESSION AND SEM

Central to the discipline of statistical analysis is the concept of psychometric and econometric analyses [12-13]. The psychometric analysis, e.g., factor and confirmatory factor analysis, has been used to identify indicators (or causes) of unobserved variables. By contrast, econometric analysis is used to estimate the causal relationship between predictor and dependent variables in theoretical models. Psychometric and econometric analyses can well explain the major differences between multivariate regression and structural equation modeling (see Table 1).

In general, multivariate regression is understood to be first generation whereas the structural equation modeling is the second generation of statistical methods [13]. The multivariate regression is essentially an econometric analysis to estimate the causal relationship between predictor and dependent variables. By contrast, structural equation modeling incorporates both econometric and psychometric analyses in the statistical estimation [13]. For this reason, structural equation modeling is best suited to measure unobserved variables (or theoretical constructs) that using survey-based data and permits measurement error in statistical...
estimations. In other words, structural equation modeling offers a distinctive advantage because it verifies both the validity of measures (or constructs) in psychometric analysis together with estimated relations in the econometric analysis. As a result, the researchers tend to utilize the structural equation modeling to estimate the statistical significance of unobserved variables (constructs) in the empirical studies.

Table 1 Comparison between multivariate regression and structural equation modeling

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Multivariate Regression</th>
<th>Structural Equation Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>The statistical tests of theoretical propositions are based on an empirical model.</td>
<td>Allow the flexibility of interplay between theory and data.</td>
</tr>
<tr>
<td>Model Complexity</td>
<td>The analysis is limited to simple models with one dependent variable.</td>
<td>An interrelated regression models can be estimated simultaneously.</td>
</tr>
<tr>
<td>Measures</td>
<td>The variables are directly observable. Assume no measurement error in observed variables.</td>
<td>The variables can be observed variables and unobserved variables. Unobserved variables can be estimated via multiple observed variables. Allow measurement error in modeling.</td>
</tr>
</tbody>
</table>

Sources: Derived from Fornell [14]

Another key difference between multivariate regression and structural equation modeling in research design is the complexity of models [14-15]. Multivariate regression is an estimation method that limited to simple model structure, i.e., model with single dependent variable. Structural equation modeling, on the contrary, can incorporate multiple dependent variables in an interrelated regression models simultaneously.

Multivariate regression and structural equation modeling differ not only in model complexity, but also in the way these methods were used based on research objectives [14-15]. Multivariate regression is generally used to predict the theoretical relationship in an empirical model. By contrast, structural equation modeling offers flexibility to researchers in theory or data testing in addition to predicting theoretical relationship. In particular, the researchers can identify how data is fitted with various models based on model fit indices in covariance-based SEM.

3.0 COVARIANCE-BASED AND VARIANCE-BASED SEM

Structural equation modeling can be further categorized into covariance-based SEM and variance-based SEM. Variance-based SEM is also known as Partial Least Squares Path Modeling or Partial Least Squares-Structural Equation Modeling (PLS-SEM). Table 2 presents the summary of differences between both types of SEM methods. The selection of the SEM methods are based on the advantages of the SEM methods in research design [10].

First, the selection of SEM methods is based on research objectives, i.e., to perform theory testing or predict theoretical relations [10; 16]. In general, CB-SEM analysis focuses on theory testing whereas PLS-SEM is optimized to perform predictive-causal investigation in empirical studies. The CB-SEM assumes a true model in estimation and thus it is regarded as confirmatory in the research domain. This can be explained by the fact that the algorithm of CB-SEM is designed to maximize the covariance of items in the structural paths with those from actual data. On the contrary, the algorithm of PLS-SEM will first estimates the constructs scores and then estimates the statistical significance of path coefficients in structural model. Specifically, the algorithm of PLS-SEM is geared toward exploratory study in a limited information context. Nonetheless, PLS-SEM can be used as a confirmatory analysis to create new measures or paths in an incremental study. The different algorithms in both SEM methods lead to the evaluation of CB-SEM is based on goodness-of-fit whereas PLS-SEM concentrates on the predictive power.

Second, sample size and data characteristics could also influence the choice of SEM methods [10]. A large sample size and normality of data are precondition to perform CB-SEM. By contrast, PLS-SEM is a non-parametric method that suitable for smaller sample size and (or) non-normally distributed data in analysis [17]. Nonetheless, the estimates derived from both methods tend to be consistent in the settings of large sample size, greater number of indicators and normality of data [10; 18-19].

Finally, PLS-SEM is the appropriate method to perform causal inference with formative constructs, but not for CB-SEM [10]. The rationale is that formative construct is invalid with the assumption of CB-SEM in which all latent variables are reflective constructs. Such argument is verified through a study that demonstrates using formative construct in CB-SEM tend to be biased [20]. This concern mainly attribute to the fact that zero correlations may exist between the items and occasionally the formative constructs may encounter the identification problems.
THE ISSUES OF USING MULTIVARIATE REGRESSION IN FINANCIAL ACCOUNTING RESEARCH

4.1 Data Characteristics

It is commonly known that some observed variables in archival based financial accounting empirical studies often to be characterized by skewed distributions [21-23]. To illustrate, the authors analyze five commonly used observed variables to identify whether these variables are normally distributed using a sample of manufacturing firms listed in industrial product index on Bursa Malaysia in 2006. These variables are Tobin's Q, capital expenditure scaled by total sales, debt-to-total assets ratio, total assets and total sales. The analysis of skewness and kurtosis are presented in Table 3. Clearly, the results show that all of the variables failed to meet parametric assumption because statistical z value for skewness (or kurtosis) outside the range of ±2.58 [24]. These results, however, are unsurprisingly because prior empirical studies often detect such variables that not normally distributed [21-23].

It is worth to mention that prior studies have used logarithm transformation to correct the skewed distributions of variables in order to meet parametric assumptions of multivariate regression [21-23]. The log-transformed variables, however, may lead to some problems in the analyses. First, the estimates of log-transformed variable in cross-sectional data may be inconsistent and biased with the presence of heteroscedasticity in ordinary least squares [25]. Second, using log-transformed variables in panel data regression may lead to biased estimates when some observations are zero [26]. Third, scaling issues (measurement error) may occur and lead to biased estimates when interacting log-transformed and the non-log-transformed variables in estimating moderating effects [27].

Finally, log-transformed variables require careful interpretation because of its effect is multiplicative in nature [28].

Table 3 Analysis of skewness and kurtosis statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market-to-book ratio</td>
<td>10.809</td>
<td>133.798</td>
</tr>
<tr>
<td>Capital expenditure</td>
<td>3.598</td>
<td>17.994</td>
</tr>
<tr>
<td>Firm leverage</td>
<td>4.808</td>
<td>40.232</td>
</tr>
<tr>
<td>Total assets</td>
<td>13.451</td>
<td>184.202</td>
</tr>
<tr>
<td>Total sales</td>
<td>5.972</td>
<td>42.079</td>
</tr>
</tbody>
</table>

Sources: Researchers' own construction

4.2 Endogeneity

In financial accounting research, endogeneity often becomes one of criteria to assess the robustness of results in econometric analyses [2]. Such problem occurs when the predictor variables are not theoretically exogenous in research settings thus leading to spurious or inconsistent estimates. Stated differently, researchers must not interpret that a significant relationship in ordinary least squares analyses is a causal relationship with the presence of endogeneity. In financial accounting research, instrumental variables methods are widely used to mitigate the endogeneity problems in econometric analyses [2].

Table 2 The difference between the covariance-based SEM and variance-based SEM

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Covariance-based SEM</th>
<th>PLS-SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>The algorithm attempts to generate estimates for the latent constructs in the structural paths and the corresponding measurement loadings by maximizing the covariance of any connected two items in the structural paths so that similar to the covariance obtained from actual sample data.</td>
<td>The algorithm involves two important processes. First, the algorithm attempts to generate estimated score of latent constructs based on the connected items. Second, the algorithm generates PLS estimates based on the immediate blocks of a particular construct in the structural path.</td>
</tr>
<tr>
<td>Implication</td>
<td>Focus on covariance of all items in the proposed model based on the goodness-of-fit and chi-square statistic.</td>
<td>Focus on the maximization of variances of dependent variables. PLS-SEM is a predictive-oriented approach.</td>
</tr>
<tr>
<td>Distributional Assumptions</td>
<td>CB-SEM is a parametric-approach which assumes there are identical distributions in observations, and these observations are independent.</td>
<td>No distributional assumption is made. PLS-SEM is essentially a non-parametric approach.</td>
</tr>
<tr>
<td>Confirmatory/Exploratory Studies</td>
<td>CB-SEM utilizes full information, i.e., maximum likelihood, under the assumption of a “true” model. Thus, The CB-SEM focuses on confirmatory analysis.</td>
<td>PLS-SEM can be used in an exploratory study, which is a limited information approach (i.e., the theoretical knowledge is relatively limited). PLS-SEM can be used in confirmatory purpose when the research objective is to perform an incremental study to create new structural paths or new measures by relying on a theory-driven baseline model.</td>
</tr>
<tr>
<td>Sample size</td>
<td>Relatively large sample size required in analysis. The required sample size is based on the Cohen statistical power analysis.</td>
<td>The sample size requirement can be based on the OLS regression rule, which is 20 cases per dependent variable. As a general rule of thumb the minimum sample size is 100 and 200.</td>
</tr>
<tr>
<td>Construct</td>
<td>CB-SEM is limited to the use of reflective construct.</td>
<td>Reflective and formative construct can be used.</td>
</tr>
</tbody>
</table>

Sources: Derived from Chin [10]
4.3 Construct

Although archival data are often regarded as observed variables, but archival data also can be used to measure unobserved variables (constructs) that defined by researchers. For example archived data have been used to measure unobserved variables, e.g., financial disclosure index and corporate governance related indices, in financial accounting research [29-30]. Specifically, researchers have to use a summed scale and index to measure unobserved variables from archival data [31-32]. Normally, researchers will follow prior self-constructed or regulatory-constructed index in estimating unobserved variables prior to perform multivariate regression. Nevertheless, such approach may be less reliable because multivariate regression assume there is no measurement error [14]. Furthermore, some unobserved variables, e.g., decisional process, are commonly analyzed together with archival data may lead to biases in measurement when using multivariate regression [33].

4.4 Moderation and Mediation

Moderation and mediation analyses are considered to be one of main themes to describe the causal relationship in financial accounting research. A moderator can be interpreted as a variable that strengthens (or weakens) the cause-effect relations, on the one hand; a mediator is a third variable that exerts an intermediary process to the causal-effect relations, on the other [34]. In multivariate regression, products of sums approach (i.e., interaction terms) is used for estimating the moderation effect whereas separate multi-step process (e.g., Barron and Kenny) was used to analyze the mediating effect [1]. Since multivariate regression is limited to single dependent variable, the moderating and mediating effects are not examined simultaneously a complex model with multiple dependent variables.

5.0 WHEN TO USE PLS-SEM IN ARCHIVAL-BASED FINANCIAL ACCOUNTING RESEARCH

Undoubtedly, multivariate regression is the de facto statistical method in financial accounting research [1]. Meanwhile, PLS-SEM is a under-utilized statistical method although the estimation technique of multivariate regression (e.g., ordinary least squares) and PLS-SEM method is similar, i.e., maximize the prediction of original raw scores. Both multivariate regression and PLS-SEM are predictive-oriented methods. For this reason, the authors argue that PLS-SEM could be a substitute to multivariate regression in predicting the causal relationship in archival financial accounting research. Specifically, researchers could justify the choice of PLS-SEM to estimate cause-effect relations based on data characteristics and research considerations. Table 4 provides an overview concerning using multivariate regression and PLS-SEM in archival financial accounting research.

First and foremost, PLS-SEM is an appropriate method when empirical data cannot meet the parametric assumption of multivariate regression, i.e., normality of data [10; 17]. Since empirical archival data essentially are documented historical data, researchers are unlikely to enhance of data. For example, financial data that derived from annual reports of publicly listed companies are generally verified by qualified auditors and bound to fulfill the laws of regulatory bodies. As a result, the researchers cannot alter the data even though the archival data could not meet parametric assumption. As discussed previously, it is a common practice for researchers to perform logarithm transformation to correct the skewed distributions of variables in archival-based financial accounting research and thus may lead to biased estimates and interpretation problems in multivariate regression. In this regard, PLS-SEM can be used because it is a non-parametric method because both PLS-SEM and multivariate regression are predictive-oriented methods. A study has also shown that the estimation of PLS-SEM is robust using skewed data [19]. In addition, the authors argue that PLS-SEM could yield higher statistical power to estimate non-linear relationship of predictor variables in empirical studies. The rationale is that multivariate regression require normally distributed data that are unlikely to existed in variables of non-linear relationship. By contrast, PLS-SEM allows the modeling for non-linear terms in heterogenous data [35].

Prior empirical studies suggest that some predictor variables (e.g., corporate governance) are difficult to be observed directly [8; 36-37]. Therefore, multivariate regression, which assumes variables can be measured directly, is not appropriate and suffers with measurement error [15]. For example, Johnson and Greening [37] and Azim [8] suggest that investigating the combined effects of unobserved corporate governance mechanisms renders a new research direction. In this regard, PLS-SEM could be used to fulfill the research considerations. First, the combined effect of observed variables can be examined provided that those observed variables were measuring the same attribute [17]. Second, single-item construct can be used alongside with multi-item constructs in PLS-SEM, but not for CB-SEM [17]. Single-item construct is also known as concrete construct that represented singular attribute [38]. For this reason, PLS-SEM is appropriate because some archival data particularly financial data can be conceptualized as singular and concrete in terms of attribute [11]. This argument can be detected in all empirical studies using multivariate regression that assume no measurement error in estimation.

In addition, PLS-SEM can be used for theory building in research, i.e., to create new constructs (unobservable variables) or structural path in theoretical models [10]. This usually occurs when the researchers intend to first perform incremental study prior to theory testing. Stated differently, it could be that researchers intend to migrate from multivariate regression based theoretical models to structural equation modeling in theory testing. In this research context, PLS-SEM algorithm can confine the new constructs and measures to immediate block of constructs in the structural paths [10]. On the contrary, CB-SEM, which is full information approach, may not appropriate to test new constructs or structural paths in theoretical models because it is most likely result in poor goodness-of-fit indices. One explanation is that the goodness-of-fit indices in CB-SEM relies on the how close the covariance of two items in structural paths to the covariance that obtained from data [39]. Thus, the goodness-of-fit indices can be worsen because of a unreliable single measure [40]. The goodness-of-fit indices are also affected by sample size, normality of data and estimation procedures [41].

Unlike multivariate regression, PLS-SEM offers confirmatory factor analysis on measurement in formulating unobserved variables. Prior studies in financial accounting research have largely depended on summed scale and index to measure unobserved variables (see previous discussion). Such measures on unobserved variables (or constructs) in multivariate regression are assumed with the absence of measurement error and thus may lead to biased estimate [42]. For this reason, PLS-SEM could be a better approach compared to multivariate regression since the former accounts for measurement error of unobserved variables in estimation. To illustrate, Li et al., built a construct in PLS-SEM analysis to measure international ownership by using three type of international ownership [36]. This approach not only investigates the combine effects of observed variables, but also perform a confirmatory factor analyses to verify those three observed variables exhibit same attribute.
PLS-SEM could also be an alternative approach to test moderation and mediation in a complex model. As discussed previously, multivariate regression is limited to only one dependent variable. In comparison, PLS-SEM is a better approach to estimate multiple dependent variables (or endogenous constructs) in a fairly complex model simultaneously. That is, the variables in the PLS-SEM analysis can become the predictor and dependent variables at the same time. With regard to mediation analysis, PLS-SEM is also a straightforward approach compared to multivariate regression because the former can estimate the mediation effect in a complete model, whereas the latter involves several steps to estimate the moderating and mediating relationships [15]. It is also worth to mention that some PLS-SEM software can generate the direct, indirect and total effects in the models directly [43]. In a similar vein, PLS-SEM allows the moderation effects could also be estimated in a complex model with multiple dependent variables.

Finally, PLS-SEM is an useful empirical method to examine the moderating effect in two ways. First, PLS-SEM can be used to model the interaction terms for non-normally distributed variables. The rationale is that when data failed to meet the normality requirement of multivariate regression, researchers may tempt to perform logarithm transformation prior to modeling interaction terms. As a result, interaction terms between log-transformed and non-log-transformed variables may lead to scaling issues (measurement error) and biased estimates [27]. Second, PLS-SEM offers multigroup analyses to examine different group of data [17; 44]. The multigroup moderation is aimed to examine if there were systematical difference between the parameters for two (or more) groups. Thus, the authors suggest that multigroup moderation may be useful in financial accounting research because prior studies have emphasized the heterogeneity of cause-effect relations between different institutional or industry contexts [45–46].

Thus far the authors have argued that PLS-SEM is beneficial in various research design contexts. However, PLS-SEM is inappropriate to be applied in a model with endogeneity concern. The PLS-SEM only supports recursive model in which there is no causal loops in analyses [17]. Owing to this limitation, the estimation of PLS-SEM will be unreliable when endogeneity appears to be main problem in econometric analysis [2]. Thus, the authors advise researchers to perform endogeneity test using multivariate regression to ensure the exogeneity of predictor variables prior to the use of PLS-SEM. By doing so, PLS-SEM could be applied to ensure the consistency and robustness of results. Alternatively, PLS-SEM could be used if researchers could justify the research settings in their empirical studies are exogenous [3].

### 6.0 CONCLUSION

The demand for inference studies is of paramount importance in financial accounting research. In this regard, under-utilized statistical methods that can address the deficiencies in popular research design should not be overlooked [47]. The authors felt that PLS-SEM represents one of alternative statistical methods in financial accounting research although multivariate regression is more well-received among researchers. Thus, the purpose of this article was to present an non-technical overview with regard to multivariate regression, covariance based SEM and PLS-SEM as well as the guidelines to use PLS-SEM in archival financial accounting empirical research.

This study is beneficial to financial accounting researchers who may lack understanding on the use of PLS-SEM in archival data-based research context. The authors have identified some of research considerations in financial accounting research which can explain why PLS-SEM is a appropriate statistical method in empirical analyses. Thus, the major contribution of this article is to provide necessary background to financial accounting researchers concerning how PLS-SEM can be applied to in archival financial accounting research based on data characteristic and research considerations.

### References


