Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1
© 2023 Little Lion Scientific

LATENT MODELING FOR PREDICTING MULTIDIMENSIONAL DATA

YASMINA AL MAROUNI[1], YOUSSEF BENTALEB[2]
[1]Engineering Sciences Laboratory, ENSA Kenitra, Ibn Tofail, Morocco
[2]Engineering Sciences Laboratory, ENSA Kenitra, Ibn Tofail, Morocco
E-mail : [1]yasmina.almarouni@uit.ac.ma, [2]youssef.bentaleb@uit.ac.ma

ABSTRACT

The purpose of the current paper is to find and adapt a statistical method that tackles two main issues. First, prediction in a multidimensional context whether for quantitative or categorical data. Second, modeling a complex cause-effect relations. In particular, the use of structural modeling of latent and manifest variables to derive the regression equation. This leads us to the discussion of the two main methods within Structural Equation Modeling (SEM): Partial Least Squares Path Modeling (PLS-PM) and Linear Structural Relations (LISREL). Upon a thorough comparison of the two methods, it was determined that the best approach is PLS-PM. Nevertheless, it is essential to acknowledge that this method has its limitations. To address these shortcomings, the authors have proposed an adaptation of the PLS-PM. The paper concludes with the practical application of the developed method. This enabled us, using a small sample of non-quantitative variables, to model the phenomenon of cybercrime among children via latent and manifest variables and to predict whether a child might become a victim of cybercrime.

Keywords: *SEM, PLS, LISREL, PLS-PM, Path modeling, Categorical data, Missing values, Small sample size, Latent variables.*

1. INTRODUCTION

In today's world, the amount of data we generate and collect is continuously growing. This vast pool of data holds a wealth of knowledge and information that can be used in many ways. About 23 years ago, statisticians started to focus on how to handle and make sense of high-dimensional data [1]. But as the amount and complexity of data increase, we encounter new challenges in managing and exploring it. The explosion of data not only increases the number of variables we need to consider in our analyses but also raises issues like multi-collinearity, where variables are highly interrelated, and the likelihood of encountering various types of data.

To address these challenges and effectively analyze and understand data in the context of big data, we need specific statistical methods. These methods should be able to reduce the complexity of data by boiling it down to its most important elements and handle different types of data. Consequently, the challenge is to find the most appropriate statistical method that allows to analyze causality relationships between observed and unobserved variables and derive regression equation while overcoming those challenges:

1. Large number of variables.
2. Presence of qualitative variables.
3. Presence of data with missing values.

In this context, Structural Equation Models (SEM) presents a viable option for modeling the relationships between observed and unobserved variables. SEM can model latent variables and analyze relationships between them. It has been used in various fields such as science, business, and education, proving its versatility in handling complex data structures. SEM allows for the estimation of complex cause-effect relationships within path models, making it an essential tool in our analysis [2].

This estimation can be approached from two main perspectives: the maximum likelihood approach, often associated with LISREL, and the partial least squares path modeling (PLS-PM) approach, which is also referred to as partial least squares structural equation modeling (PLS-SEM) [3]–[5].

Further on, we will explore both methods and adapt the one most pertinent to our current study. This is crucial as we have encountered various challenges, including limitations related to sample size, variable types, and other pertinent issues.

To validate our approach, we applied the adapted method in a use case to identify the cause-effect relationship between the behavior of children on the internet and their likelihood to become cybercrime victims. The study intends to determine the regression equation that will predict whether a child may be a victim of cybercrime using missed and/or incomplete data from a survey conducted with 490 children from different cities in Morocco.

The rest of the article is structured as follows: we will start by presenting structural equation modeling and its applicability in modeling complex phenomena. Following that, we will conduct a comparative analysis of the two structural equation modeling (SEM) methods, PLS-PM and LISREL. We will then describe the chosen method, outline the challenges encountered, and propose necessary adaptations. The theoretical part will be supported by an application before the concluding section.

2. STRUCTURAL EQUATION MODELING

Structural equation modeling (SEM) refers to a diverse set of mathematical models, computational algorithms, and statistical methods that match a network of concepts to data [2].

Structural equation modelling is also a comprehensive statistical approach to test hypotheses exploring the relationships between observed and latent variables. It is a methodology to represent, estimate and test a theoretical network of (mainly) linear relationships between variables [6].

SEM models include several statistical methodologies that permit to estimate the causal relationships. It is defined according to a theoretical model, linking two or more Latent Variables (LV), each measured through several Manifest Variables (MV). In the following, the causality scheme will be presented using the drawing conventions (Figure 1).

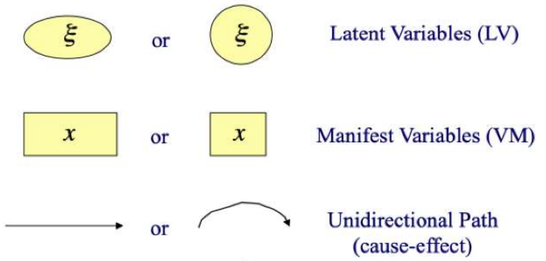


Figure 1: Drawing Convention.

To estimate model parameters in the modeling of structural equations, there are two approaches: The maximum likelihood approach - LISREL, based on covariance, developed by Jöreskog (1970) through the software LISREL (LInear Structural RELationships) [7] and the PLS approach (PLS-PM), also known as Partial Least Squares Structural Equation Modeling (PLS-SEM), based on component estimation that maximizes the amount of variance explained, it was proposed by Wold (1975, 1982, 1985) and in 1989, the PLS 1.8 software was developed by Lohmöller [3]–[5], [8]–[11].

2.1 LISREL approach

LISREL (Linear Structural Relations) is a structural model, estimates the system of structural equations using maximum likelihood. Each manifest variable is written according to its latent variable. The first coefficient is set to 1 and the others are estimated by maximum likelihood assuming that the manifest variables follow a multinormal law [12].

The LISREL model has two parts:

- A system of structural equations formalizing the hypotheses relating to the variables considered in the study and allowing to test these hypotheses.

- A measurement model formalizing the relationships between latent and observed variables.

The LISREL model integrates and generalizes the models used in factor analysis, multiple regression analysis and causal track analysis. It is used by confirmatory factor analysis and by the analysis of systems of relations some of which are oriented logically or chronologically and can be considered causal [13].

Nevertheless, to use the LISREL method, the following prerequisites must be ensured [13]:

- Statistical models must be linear.
- The model is valid only under the following conditions:
    - Independence of observations (multilevel possible).
    - Multivariate normality of data.
    - Unidimensionality of variable blocks.
- The LISREL method requires researchers to think in terms of models and hypotheses.

2.2 PLS approach

PLS method derives its name from the use of least square regression techniques to estimate the models.

Path models are composed of two elements [4]:

• The structural model, also called the inner model in PLS-SEM, which describes the relations between the latent variables.
• The measurement model, also called external model which describe the relations between the latent and manifest variables.

The aim of PLS approach is to maximize the explained variance of the dependent latent variable.

Parameter estimation is iterative, meaning that latent variables are estimated successively by the external model (via manifest variables) and then by the internal model (via the other latent variables to which it is linked) until convergence [14].

Specifically, PLS approach has four steps:

1. Estimation of the value of the latent variables based on the scores of the manifest variables and the weights of the external model (from step 4 or arbitrarily set to initialize the iteration).

2. Estimating the structural links between latent variables (internal model).

3. Estimation of latent variables of the internal model, that is, by the values

of latent variables calculated in step 1 and the links calculated in step 2.

4. Estimating the weights of the external model using the values of the latent

variables from step 3 and returning to the first step of the process.

2. COMPARATIVE STUDY

Various comparisons have been made between LISREL and PLS-SEM (or PLS-PM). In this chapter, we compare the two methods and determine which is most appropriate for the context of the study.

For starter, the maximum likelihood approach is a covariance-based structural equation modeling developed by Jöreskog (1970) across the LISREL (Linear Structural RELationships) software [7].

While the PLS approach is a component-based estimation approach that maximizes the amount of variance explained. The PLS approach was proposed by Wold (1975, 1982, 1985) and the PLS 1.8 software was developed by Lohmöller in 1989 [8]–[11].

As for the characteristics of the two methods, LISREL is characterized by the possible presence of problems of identification and not convergence of the algorithm, latent variables are not estimated at the individual level and based on multi-normality assumptions [15].

Whereas, the PLS approach is characterized by a limited number of probabilistic assumptions, data are modelled directly by a single or multiple regression sequence, and latent variables are estimated at the individual level [15]. PLS allows to study blocks of observed variables on the same individuals in the framework of the modelling of structural relationships of latent variables [15].

The following is a summary of the results of the comparison made by Esposito Vinzi (2003) [16] based on the work done by Jöreskog and Wold (1982) [12] and Chin (2000) [17].

• Objective: PLS is oriented towards the realization of prevision and prediction, while LISREL is oriented towards parameter estimation.

• Methodology: PLS is based on variance whereas LISREL is based on covariance.

• Latent variables (LV): For PLS, latent variable is a linear combination of its manifest variables. While for LISREL, latent variable is a linear combination of all manifest variables.

• Relations between (LV) and associated manifest variables (MV): PLS covers the reflective and formative types, while LISREL covers only the reflective type.

• Missing data processing: For PLS the method used is NIPLAS (Nonlinear

estimation by Iterative Partial Least Square) and for LISREL it is maximum likelihood.

• Optimality: For PLS the optimality lies in the accuracy of the prediction while for LISREL it lies in the accuracy of the parameters.

• Model complexity: Large (ex: 100 VL, 1000 VM) for PLS and reduced or moderate (<100 VM) for LISREL.

• Min sample size: 30-100 cases for PLS, 200-800 cases for LISREL.

Here we are going to shackle by the conclusion of another comparison. The authors [18] concluded that the principal motivation in using the PLS method consists of exploration and prediction, it is more an exploratory approach than a confirmatory one. In addition, PLS modelling avoids problems arising from the small sample size and can estimate very complex models either reflective or formative via many latent and manifest variables. Not to mention the fact that PLS modelling has less rigorous assumptions about the distribution of variables and errors.

After comparing the two methods, PLS regression stands out, especially when dealing with a large number of predictors, significant sample sizes, and multi-collinearity issues. Based on all its features, the PLS-PM is the most appropriate method to model the phenomenon of cybercrime for children, through its predictive capacity. However, it's important to note that PLS was initially designed for quantitative variables, presenting a challenge when we apply it to big data that includes various data types. Hence, its implementation in our use case has been challenging, so what are these challenges and how do we overcome them?

3. THE SELECTED APPROACH - PLS-PM

According to the result of the discussion in the third session, the PLS-PM is the most suitable method for the study. In this chapter, the authors detailed the PLS-PM model and equations.

2.2.1 Model equation

Let J be the number of groups of variables, the data consists of J groups of $X_j = \{x_{j1}, .., x_{jkj}\}$ observed on n individuals.

The x_{jh} variables are called manifest variables and are assumed to be centered. X_j is the matrix of observed data from the j^{ieme} group.

Each group of manifest variables represents the observable expression of a reduced centered latent variable ξ_j to which it is related by the following equation:

$$x_{jh} = \pi_{jh}\xi_j + \varepsilon_{jh} \qquad (1)$$

With the following:

• ε_{jh}: Random term, assumed to mean zero and not correlated to the latent variable ξ_j.

• The signs ξ_j of π_{jh} are to be specified.

The structural relations between latent variables are described as follows:

$$\xi_j = \Sigma_i \beta_{ij} \xi_i + \nu_j \qquad (2)$$

With:

• ν_{jh}: Random term, assumed to mean zero and not correlated to the latent variable ξ_i (of the equation (2)).

3.2.2. Model presentation

The PLS model consists of two sub-models, the internal model that describes the relations among the latent variables and the external model (or measurement model) that describes the relations

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

among the manifest variables and the corresponding latent variable, presented in the figure [14] bellow:

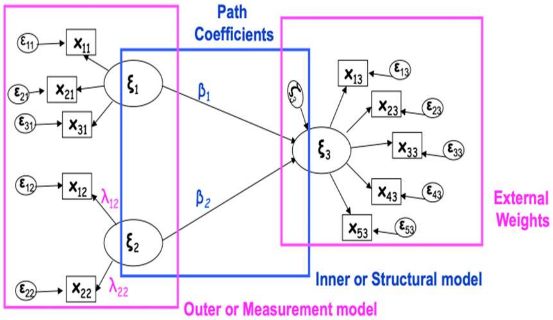


Figure 2: Path model with latent variables.

The βi are used to identify the sign of the correlation between latent variables:

- βi = 1, if the correlation between ξi and ξj is positive.

- βi = -1, if the correlation between ξi and ξj is negative.

- βi = 0, if ξi and ξj are not related to each other: The coefficient βi is then structurally nil.

The external model is divided into two types:

- Formative type: The latent variable ξj is the reflection of the manifest variables of the block Xj.

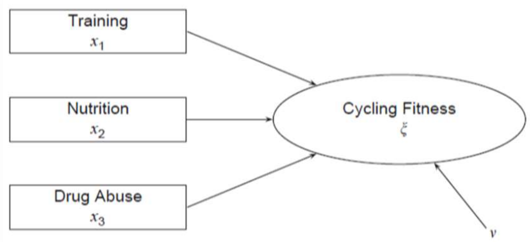


Figure 3: Example of formative type.

- Reflexive or effect indicators : Manifest variables reflect their latent variable, for example:

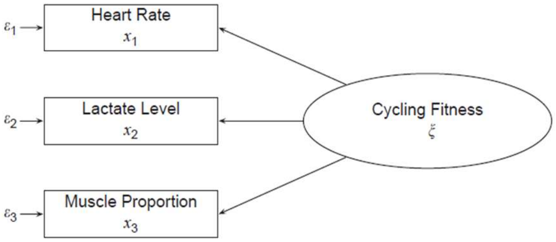


Figure 4: Example of reflective measurement model.

4. PLS-PM CHALLENGES AND THE PROPOSED ADAPTATION

Notwithstanding their advantages, in the context of this study, PLS-PM method has faced challenges that can create impediments like the nature of the data.

Indeed, it's important to note that PLS-PM was originally designed to handle quantitative variables exclusively. Additionally, the relatively small sample size, approximately 100 in this case, poses further challenges. Furthermore, the inconsistency of PLS path coefficient estimates in scenarios involving reflective measurements can introduce adverse consequences for hypothesis testing.

The fundamental question that needs to be addressed is whether the standard PLS-PM approach is suitable for our specific context or whether we must adapt the PLS-PM to tackle these challenges effectively.

We will now proceed to address each of these challenges separately and conclude this section with a summary.

4.1 How to deal with inconsistency

As mentioned previously, the external model is composed by reflective and/or formative measurements. In the case of reflective measurement, the inconsistency of PLS path coefficient estimates can have adverse consequences for hypothesis testing. To remedy this, an extension of PLS called consistent PLS (PLSc) was introduced [19].

Consistent PLS (PLSc) [19], first mathematically developed by Dijkstra, "aims to 'adjust' the attenuated regression path estimates to the correct value using the reliability of the constructs" [20]. Indeed, PLSc provides a correction for estimates when PLS is applied to reflective constructs, so, the path coefficients, interconstruct correlations, and indicator loadings become consistent [19].

The four steps of consistent PLS [19] are:

1. The application of traditional PLS to provide latent variable scores and to estimate latent variable correlations and weights.

2. The determination of the new reliability coefficient for each reflective construct.

3. The correction of the original latent variable correlations by using the new reliability coefficient for attenuation, and thus obtains consistent correlations of latent variables.

4. The estimation of consistent path coefficients in a least squares manner based on the consistent correlations of latent variables.

4.2 How to deal with small sample

In this section, the objective is to find whether the authors could use PLS-SEM with a small sample size or not, subsequently identify the conditions required to use it if exists.

To begin with, according to two previous studies, the PLS-SEM worked well with small samples [21], [22].

In addition, a more recent simulation study [23] stated that PLS-SEM is a good choice when the sample is small. Moreover, the consistent PLS (PLSc) approach provided corrected model estimates while maintaining the ability to process complex models when the sample size is limited [19], [24].

Second, some researchers believed in the 0 times rule. This rule indicates that the sample size should be equal to the larger of 10 times the maximum number of arrowheads pointing at a latent variable anywhere in the PLS path model [4], [17].

Based on the above, PLS-SEM can be used when the sample is small [21]–[23], notably when using consistent PLS [19], [24], and it is preferable to follow the 10 times rule [4], [25].

4.3 How to deal with categorical predictors with missing values

Historically, the PLS method has been designed to address only quantitative variables. In order to process categorical variables, the most common method was to replace each categorical variable by its indicator matrix [15], in which case, the categories will be intercepted as variables. Other adaptations have been proposed, so that a comparative study of methods for dealing with non-quantitative variables has been carried out [26].

Based on the results of the cited comparison [26], and in order to treat categorical data with missing values by PLS regression, the authors [27] chose to adapt PLS for categorical predictors (PLS-CAP) [28].

The PLS-CAP has also been extended by the PLS-PM to obtain appropriate quantification of reflective categorical indicators in the PLS-PM [29]. Indeed, the PLS-CAP algorithm can naturally be extended to PLS-PM because PLS regression is a particular PLS path model with two reflective blocks[29], [30].

In the same context, here we will adapt the PLS path model using PLS1 for categorical predictors with missing values (PLS1-CAP-MV) [27] to address categorical variables with missing values.

In fact, PLS1-CAP-MV is based on PLS-CAP [28] algorithm and PLS1-NIPLAS [15] to handle categorical variables with missing values in the context of a single response variable.

Using the same quantification method as PLS1-CAP-MV, we will first replace the initial matrix with a matrix containing only quantitative variables before running the PLS path model algorithm.

4.4 How to transform categorical variables into quantitative variables

The quantification method used is based on Hayashi’s first quantification method [31] and used in (PLS1-CAP-MV). To describe this method, we will use the succeeding notations:

- $X = \{x_1,..,x_k\}$: The matrix of k variables observed on n individuals.
- ${x^{ql}}_{li}$: The value of the observation i of the vector ${x^{ql}}_{l}$.
- ${x^{qqs}}_{l}$: The vector of the quantified values without rows containing missing values.
- ${x^{qq}}_{li}$ ou Xqqli: The value of the observation i of the vector ${x^{qq}}_{l}$.
- ${u^{(e)}}_{li}$: Equal to true if the value of individual i of the vector u_l exists.
- ${x^{(e)}}_{ji}$: Equal to true if the value of individual i of the x_{ji} exists.

The algorithm starts by initializing u_l with the response variable and follows the next steps.

Step 1: Calculation of quantified values

The objective of this step [27] is to replace all qualitative variables with missing values by quantitative variables with the same missing values.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

For each $l \in (1: L)$:

$$x^{qq(e)}_{l} = G^{(e)}_{l}(G^{(e)}_{l}{}' \, G^{(e)}_{l})^{-1} \, G^{(e)}_{l}{}' \, u_{1l}{}^{Xqqli\,(e)} \quad (3)$$

With:

- $x^{qq}{}_{l}{}^{(e)}$: Vector of $x^{ql}{}_{li}$ if $x^{ql}{}_{li}$ exists.

- $G^{(e)}{}_{l}$: Dummy matrix of $x^{qq(e)}{}_{l}$.

$u_{1l}{}^{Xqqli\,(e)}$: Vector containing the u_{1li} values of the u_1 vector for observations where $x^{ql}{}_{li}$ exists.

The resulting quantified variable is:
$x^{qq}{}_{l}$ = { for each observation i: if i exists $x^{qq}{}_{li}$ else NA }.

Step 2: Deduce the new quantified matrix

After the calculation of the quantified variables, comes the step of the matrix deduction which will be used as input to the classical method PLS1-PM. The following equation will use the notations below:

- X_{qt}: The matrix of quantitative variables.

- X_{qq}: The matrix of quantified variables.

- X_0: The juxtaposition of X_{qt} and X_{qq}.

$$X_0 = [X_{qt}|X_{qq}] \quad (4)$$

Note that X_0 contains the same missing values as the original matrix, making the choice of how to manage missing values more flexible [27].

4.5 Summary of the adaptation

To deal with all encountered problems, first we began by the quantification of categorical data, second, applied a combination of PLS-PM and consistent PLS. Third, the path model should respect the maximum pointing arrowheads to the latent variable [4], [25].

The new method will follow these steps :

1. Create a new matrix initialized by the quantitative variables of the initial matrix.
2. For each qualitative variable: Replace each qualitative value with a numerical value while retaining the missing values.
3. Concatenate the result quantified variables with new matrix.
4. Use the new matrix as input of consistent PLS-PM algorithm.

5. APPLICATION

The objective of this chapter is to have a modeling of the phenomenon of cybercrime for children via latent and manifest variables to identify the cause effect relations between the behavior of children on the Internet and the fact that they become victims of cybercrime. It's also aimed to predict whether a child could become victim of cybercrime. To do this, we applied the proposed solution mentioned before as follow.

First, we started with the quantification step using a python algorithm using the quantification method found in paper [27]. Furthermore, we moved on to the identification of latent and manifest variables, followed by modeling the studied phenomenon while respecting the 10 times rule [4], [25] to overcome the potential small sample size problems. Finally, we applied the consistent PLS-PM algorithm using SAS.

However, within the context of this study, this method has encountered challenges that can create barriers as well as the barriers that can be caused by the nature of the data.

5.2. Data description

The study sample came from a study conducted by Moroccan Centre polytechnic research and innovation (CMRPI). It's coming from a questionnaire about cybercrime answered by 490 students (10-12 years old) from different Moroccan cities. The study sample size is 100.

The study sample was derived from research conducted by the Moroccan Center for Polytechnic Research and Innovation (CMRPI). It originates from a questionnaire focused on cybercrime, which was completed by 490 students aged between 10 and 12 years old, hailing from various cities in Morocco. However, for the purposes of our study, the sample size is reduced to 100.

The study covers the following quantitative and categorical variables:

- Manifest variables: Chat_with_stranger (CWStr), Computer_available (CA), Facebook_access (FA), Instagram_access (InsA), Internet_search_lang (IntSL), Internet_use_avg_hrs (IntUAH), Internet_use_freq (IntUF), Request_cam (RC), Request_pers_info (RPI), Smartphone_available (SmAv), Snapchat_access (SnA), Tablet_available (TblA), Twitter_access (TwA), Victim_internet_crime (VIC), Whatsapp_access (WA).

- Latent variables: Internet_use (IntU), Tools_availability (TlA), Stranger_relationship (StrR), Social_network_availability (SNA).
- The variable explained: Victim internet crime

5.3. Pattern of causality

In this study, the theoretical model was designed with the help of experts in the functional domain. The resulting causality scheme is presented in the figure (*Figure 5*) using the drawing conventions (*Figure 1*).

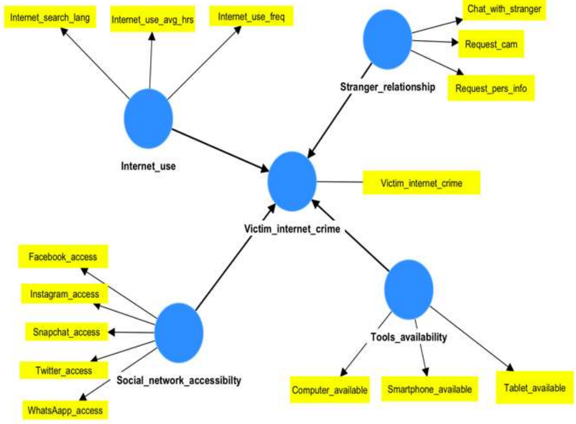


Figure 5: Causality scheme.

The schema above shows that the manifest variables reflect their latent variables, which means that the indicators are reflective.

5.3.1 Causality schema with indicators

The variables have been grouped into 5 Latent variables, each measured through a number of observable indicators.

The measurements were determined by PLS-PM algorithm. As a reminder, the algorithm has been launched after the quantification step (transforming categorical variables into quantitative variables) and the modeling step. Therefore, we obtained the causality schema with the scores below:

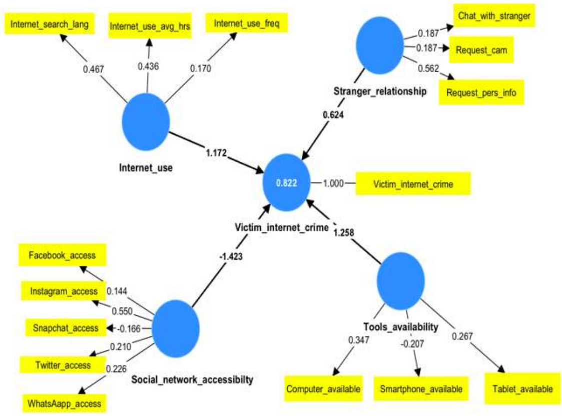


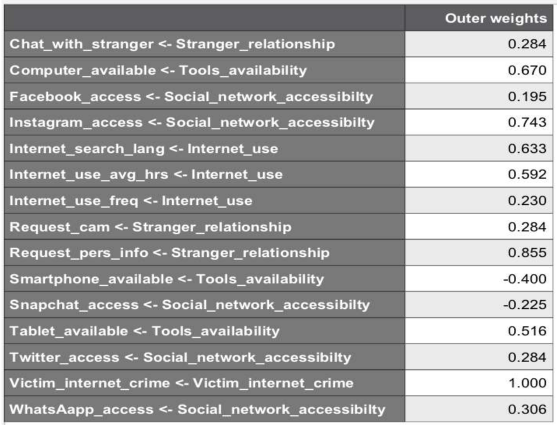
Figure 6: Causality scheme with scores scheme.

5.3.2 External model

At the end of the algorithm of the PLS approach, the final weights are obtained to link the manifest variables to the latent variables of the model.

The following table shows the outer weights of the manifest variables.

Table 1: Outer weights

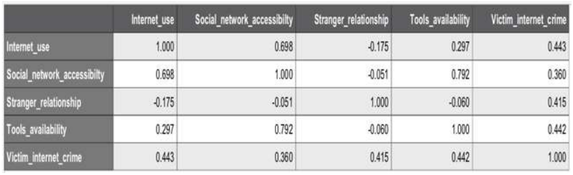


	Outer weights
Chat_with_stranger <- Stranger_relationship	0.284
Computer_available <- Tools_availability	0.670
Facebook_access <- Social_network_accessibilty	0.195
Instagram_access <- Social_network_accessibilty	0.743
Internet_search_lang <- Internet_use	0.633
Internet_use_avg_hrs <- Internet_use	0.592
Internet_use_freq <- Internet_use	0.230
Request_cam <- Stranger_relationship	0.284
Request_pers_info <- Stranger_relationship	0.855
Smartphone_available <- Tools_availability	-0.400
Snapchat_access <- Social_network_accessibilty	-0.225
Tablet_available <- Tools_availability	0.516
Twitter_access <- Social_network_accessibilty	0.284
Victim_internet_crime <- Victim_internet_crime	1.000
WhatsAapp_access <- Social_network_accessibilty	0.306

5.3.3 Internal model

Correlations between latent variables are detailed in Table 2 below:

Table 2: Latent variables – Correlations



	Internet_use	Social_network_accessibilty	Stranger_relationship	Tools_availability	Victim_internet_crime
Internet_use	1.000	0.698	-0.175	0.297	0.443
Social_network_accessibilty	0.698	1.000	-0.051	0.792	0.360
Stranger_relationship	-0.175	-0.051	1.000	-0.060	0.415
Tools_availability	0.297	0.792	-0.060	1.000	0.442
Victim_internet_crime	0.443	0.360	0.415	0.442	1.000

The results obtained show that the four latent variables have almost the same importance for “Victim internet crime”, the most important is “Internet use” (correlation coefficient = 0.43) and the least important is “Social network accessibility” (correlation coefficient = 0.36).

Therefore, to decrease the variable “Victim internet crime”, parents should act on the four latent variables starting with the variable “Internet use”, then on the “Tools availability”, “Stranger relationship”, “Social network availability”.

At the end of the algorithm of the PLS approach, the final weights are obtained to link the manifest variables to the latent variables of the model.

5.3.3 Regression equation

Given the structural pattern, determining “Victim internet crime” is a complex process involving virtually all latent variables, the equation being:

VIC = 1,258 TlA − 1,423 SNA + 1,172 StrR + 0,624 StrR (5)

With:

• Internet_use (IntU), Tools_availability (TlA), Stranger_relationship (StrR), Social_network_availability (SNA) and Victim_internet_crime (VIC).

5.3.4 Model validation

In the PLS approach, there is no overall index to judge the quality of the model as a whole [32]. Three levels of model validation are defined: the quality of the external model, the quality of the internal model, the quality of each structural equation.

Our interest in this study focuses on quality of regression equation i.e., the equation of the internal model, which is determined by R^2.

R^2 is equal to **0.822** close to 1, which mean that the quality is satisfactory.

6. CONCLUSION

In this paper, we identified and adapted a statistical method that enables both prediction in a multidimensional context and latent modeling of categorical variables. To do this, we started by comparing the two SEM methods namely LISREL and PLS-PM, which enable the calculation of composite and complex indicators (latent variables).

The comparison results showed that the PLS-PM method is the most suitable due to its predictive capacity. However, this approach faces challenges when dealing with inconsistency, small simple size, categorical predictors with missing values, and categorical variables.

An adaptation of the PLS regression was proposed to address these concerns. The proposed solution combines consistent PLS, PLS- PM and the quantification step of PLS1-CAP-MV (first step) while adhering to the 10-time rule in the modelling step to avoid small sample constraints. The proposed methodology was implemented to predict the relationship between children behavior on the internet and their likelihood to become subjects to cybercrimes. Firstly, the method began with the quantification step, the resulting matrix was used as an input of the PLS method. Secondly, we modeled the schema of causality, then calculated internal and external scores using cPLS and PLS-PM. Finally, we concluded by the deduction of the regression equation. We demonstrated that the resulting regression equation can predict whether a child may be a victim of cybercrime based on its behavior on the internet.

7. FUTURE DIRECTIONS AND PERSPECTIVES

In this study, we conducted a qualitative comparison between LISREL and PLS-PM based on the literature review. In future work, we intend conduct a thorough quantitative analysis by applying both methods model big data and compare the accuracy of both models.

ACKNOWLEDGMENT

The authors would like to thank the CMRPI (Moroccan Centre polytechnic research and innovation www.cmrpi.ma) for its support.

REFERENCES:

[1] R. D. Cook and L. Forzani, “Big data and partial least-squares prediction,” *Can. J. Stat.*, vol. 46, no. 1, pp. 62–78, 2018, doi: 10.1002/cjs.11316.

[2] D. Kaplan, *Structural Equation Modeling: Foundations and Extensions*. SAGE Publications, 2008.

[3] J. Hair, M. Sarstedt, C. Ringle, and S. Gudergan, *Advanced Issues in Partial Least Squares Structural Equation Modeling*. 2017.

[4] J. Hair, G. T. M. Hult, C. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 2022. doi: 10.1007/978-3-030-80519-7.

[5] V. Esposito Vinzi, W. W. Chin, J. Henseler, and H. Wang, Eds., *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Berlin, Heidelberg: Springer, 2010. doi: 10.1007/978-3-540-32827-8.

[6] E. E. Rigdon, “Structural Equation Modeling,” in *Modern Methods for Business Research*, Psychology Press, 1998.

[7] K. G. Jöreskog and D. Sörbom, *LISREL VI, analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*, 4th ed. Mooresville, Ind.: Scientific Software, Inc., 1984.

[8] H. Wold, “Soft modelling: The Basic Design and Some Extensions,” 1982. Accessed: Dec. 23, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Soft-modelling%3A-The-Basic-Design-and-Some-Wold/c8b4365e181ca55ec5891b07b56a9f5ffeab531f

[9] J.-B. Lohmöller, *Latent Variable Path Modeling with Partial Least Squares*. Springer Science & Business Media, 2013.

[10] J. Henseler *et al.*, “Common Beliefs and Reality About PLS: Comments on Rönkkö and Evermann (2013),” *Organ. Res. Methods*, vol. 17, no. 2, pp. 182–209, Apr. 2014, doi: 10.1177/1094428114526928.

[11] E. E. Rigdon, M. Sarstedt, and C. M. Ringle, “On Comparing Results from CB-SEM and PLS-SEM: Five Perspectives and Five Recommendations,” *Mark. ZFP – J. Res. Manag.*, vol. 39, no. 3, pp. 4–16, 2017.

[12] J. K. G, “The ML and PLS techniques for modeling with latent variables : Historical and comparative aspects,” *Syst. Indirect Obs. Part I*, pp. 263–270, 1982.

[13] E. Jakobowicz, “Contributions aux modèles d’équations structurelles à variables latentes,” These de doctorat, Paris, CNAM, 2007. Accessed: Dec. 23, 2023. [Online]. Available: https://www.theses.fr/2007CNAM0564

[14] M.-L. Mourre, “La modélisation par équations structurelles basée sur la méthode PLS : une approche intéressante pour la recherche en marketing,” in *9ème Congrès de l’Association Française du Marketing*, La Rochelle, France, May 2013. Accessed: Dec. 23, 2023. [Online]. Available: https://hal.science/hal-03278657

[15] M. Tenenhaus, *La régression PLS: théorie et pratique*. Editions TECHNIP, 1998.

[16] V. E. Vinzi, “The PLS approach to path modeling,” *IASC-IFCS Summer Sch. Lisbon*, 2003.

[17] W. Chin, “Partial least squares for IS researchers: an overview and presentation of recent advances using the PLS approach.,” in *Proceedings of the 21st International Conference on Information Systems*, Jan. 2000, p. 742.

[18] J. Henseler, C. M. Ringle, and R. R. Sinkovics, “The use of partial least squares path modeling in international marketing,” in *New Challenges to International Marketing*, vol. 20, R. R. Sinkovics and P. N. Ghauri, Eds., in Advances in International Marketing, vol. 20. , Emerald Group Publishing Limited, 2009, pp. 277–319. doi: 10.1108/S1474-7979(2009)0000020014.

[19] T. K. Dijkstra and J. Henseler, “Consistent Partial Least Squares Path Modeling,” *MIS Q.*, vol. 39, no. 2, pp. 297–316, 2015.

[20] D. L. Goodhue, W. Lewis, and R. Thompson, “Does PLS Have Advantages for Small Sample Size or Non-Normal Data?,” *MIS Q.*, vol. 36, no. 3, pp. 981–1001, 2012, doi: 10.2307/41703490.

[21] B. S. Hui and H. Wold, “Consistency and consistency at large of Partial Least Squares estimates,” *Part II*, vol. 2, 1982.

[22] W. Chin and P. Newsted, “Structural Equation Modeling Analysis with Small Samples Using Partial Least Square,” *Stat. Strateg. Small Sample Res.*, Jan. 1999.

[23] W. Reinartz, M. Haenlein, and J. Henseler, “An empirical comparison of the efficacy of covariance-based and variance-based SEM,” *Int. J. Res. Mark.*, vol. 26, no. 4, pp. 332–344, Dec. 2009, doi: 10.1016/j.ijresmar.2009.08.001.

[24] T. K. Dijkstra and J. Henseler, “Consistent and asymptotically normal PLS estimators for linear structural equations,” *Comput. Stat. Data Anal.*, vol. 81, pp. 10–23, Jan. 2015, doi: 10.1016/j.csda.2014.07.008.

[25] D. Barclay, R. Thompson, and C. Higgins, “The Partial Least Squares (PLS) Approach to Causal Modeling: Personal Computer Use as an Illustration,” *Technol. Stud.*, vol. 2, Jan. 1995.

[26] Y. Al Marouni and Y. Bentaleb, “State of art of PLS Regression for non quantitative data and in Big Data context,” in *Proceedings of the 4th International Conference on Networking, Information Systems & Security*, in NISS ’21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 1–5. doi: 10.1145/3454127.3456615.

[27] Y. Al Marouni and Y. Bentaleb, “Treatment of Categorical Variables with Missing Values Using PLS Regression,” in *Emerging Trends in Intelligent Systems & Network Security*, M. Ben Ahmed, B. A. Abdelhakim, B. K. Ane, and D. Rosiyadi, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2023, pp. 475–485. doi: 10.1007/978-3-031-15191-0_45.

[28] G. Russolillo and C. N. Lauro, “A Proposal for Handling Categorical Predictors in PLS

Regression Framework," in *Classification and Multivariate Analysis for Complex Data Structures*, B. Fichet, D. Piccolo, R. Verde, and M. Vichi, Eds., in Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer, 2011, pp. 343–350. doi: 10.1007/978-3-642-13312-1_36.

[29] L. Trinchera, G. Russolillo, and C. N. Lauro, "USING CATEGORICAL VARIABLES IN PLS PATH MODELING TO BUILD SYSTEM OF COMPOSITE INDICATORS".

[30] M. Tenenhaus, V. E. Vinzi, Y.-M. Chatelin, and C. Lauro, "PLS path modeling," *Comput. Stat. Data Anal.*, vol. 48, no. 1, pp. 159–205, Jan. 2005, doi: 10.1016/j.csda.2004.03.005.

[31] H. C, "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematical point of view," *Stat Math*, vol. 3, pp. 121–143, 1950.

[32] V. Stan and G. Saporta, "Une comparaison expérimentale entre les approches PLS et LISREL," in *38èmes Journées de Statistique*, Clamart, France, May 2006. Accessed: Dec. 23, 2023. [Online]. Available: https://hal.science/hal-01125190