Modeling with Heterogeneity

by Giuseppe Lamberti

Doctoral Dissertation

Ph.D. Tomás Aluja Banet
Advisor

UPC

Facultat de Matemàtiques
i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Barcelona, 2015

Contents

Abstract

One of the principal problems that characterizes many studies and statistical researches is represented by the nature of the considered variables: we cannot measure them directly, as is so often verified, and we need some theoretical construct to identify them. Searching for this theoretical construct, which is also known as a latent or intangible variable, means analyzing the possible relationships between the observed variables and the "concept" by proposing different theories and models. The methodology used for this purpose is the structural equation model (SEM), which allows great flexibility in modeling multiple relationships between sets composing blocks of variables.

There are essentially two methods available for estimating the structural equation model: (1) the SEM-ML (Maximum Likelihood Approach to Structural Modeling Equation), also known as LISREL (Linear Structural Relations), which is based on estimating the covariance; and (2) Partial Least Squares Path Modeling (PLS-PM), developed by Herman Wold as an alternative to LISREL for estimating latent variables. The two techniques are more complementary than conflicting, and the choice between them depends on different factors, including the purpose of analysis, the nature of the model and the research context. LISREL is suitable for confirmatory studies where models are supported by a strong theory (hard modeling), whereas PLS-PM is suitable in studies where the goal is to estimate latent variables for each individual and where the models do not have a definitive theory behind the constructs.

Over the past decade, PLS-PM has become a popular statistical tool in many fields, such as marketing, information systems, and management. The PLS-PM models do not require distributional assumptions on the latent and observed variables, and they are particularly suitable for solving problems in marketing applications, where the goal is to use the results of the PLS-PM model to implement policies oriented to consumers.

Therefore, from a methodological point of view, we will use the PLS-PM approach. The PLS-PM method is based on the use of an iterative algorithm built on NIPALS (Wold, 1966), which allows us to estimate the latent variables (intangibles) and the relationships between them (path coefficients). Recently, Michel Tenenhaus has proven that PLS-PM models are a special case of multi-table analysis and that they optimize as criterion, the sum of squared covariances among the related latent variables (Tenenhaus, 2012). Hanafi (2007) has proven the monotoneconvergence of such criterion.

Traditionally, structural equation models assume homogeneity of data: that is, all individuals belong to the same population. In many cases, this assumption is not met in studies on real data, because it is reasonable to think that different groups of people, defined by different socio-economic and psychographic characteristics, present heterogeneous behaviors.

The problem of not considering the possible existence of segments between observations is that the structural equation model may lead the analyst to obtain inadequate results. For example, if a company

wants to start a marketing campaign for its own clients and, before beginning, it decides to estimate their satisfaction through a unique PLS-PM model based on a sample of heterogeneous individuals ("young" people 18-30 years old, "adults" between 30 and 50 and "older" people between 50 and 65); it seems reasonable that the obtained estimations do not reflect the behavior of each of the three groups of individuals. In these cases, considering a single model in presence of heterogeneity would be misleading. To solve this problem, it is necessary to assume the existence of different groups in the population.

The literature distinguishes two different sources of heterogeneity: the first one is assignable, and the second one is non-assignable. Basically, we have assignable heterogeneity when we use segmentation variables to define segments. With assignable heterogeneity we can distinguish two situations: known segments and unknown segments. In the first case, individuals can be divided into segments a priori, and separate models may be estimated for each of them. The hypothesis is that an individual can be uniquely assigned to a single group that is based on one or more segmentation variables. In general, the number of groups is known a priori, and it is based on a limited number of segmentation variables: generally two, or a maximum of three. In the second case, if some segmentation variables are also available and we can divide the individuals into groups, we don't know which partitions must be considered (unknown segments). In this case, we need to find some criterion that allows us to identify which are the best partitions. When the heterogeneity is not assignable beforehand, even when assuming the existence of different segments in the data, we simply use the modeling variables to define the segments, which involves applying a technique to obtain the class partition.

In the majority of survey studies, it is normal to "not know the causes" of heterogeneity and to collect a relatively large set of segmentation variables: sex, age group, education level, religion and social status, just to name a few. In these cases it would be interesting to have a tool that automatically identifies which variables define the heterogeneous groups present in the data.

In 2009 Gaston Sánchez proposed PATHMOX as a solution to this problem. PATHMOX is a new segmentation approach for detecting heterogeneity and is based on the construction of a binary tree of PLS-PM models. The goal of PATHMOX is to analyze each partition of the tree and identify the segmentation variable that maximizes the difference between the respective PLS-PM models in each child node. In other words, it identifies the two customer segments that present different behaviors regarding relationships between latent variables. This methodology provides an important advantage over previous approaches to heterogeneity, as it allows us to detect the existence of different models for different subsets of a data-set without defining a priori segments: the segments are revealed as branches of the segmentation tree. However, some improvements to the technique can be considered.

The purpose of this thesis is to extend PATHMOX in the following ways:

1. *Extension of the PATHMOX approach for detecting which constructs differentiate segments.* The PATHMOX approach allows us to detect the existence of different models in a data-set without identifying segmentation variables beforehand. However, the F-test used in PATHMOX as split criterion is a global criterion: it allows us to assess whether all the coefficients for two compared models are equal or not; but it does not indicate which particular equation and which coefficients are responsible for the difference. To identify the significant distinct equation and the responsible coefficients of the split, we will introduce the F-block test and the F-coefficient test.

2. *Extension of the PATHMOX approach for dealing with the factor invariance problem.* In the context of PLS, the algorithm works by fixing the topology of the structural model (i.e. the model of the causal relationship between the latent variables), and the goal is to detect segments that have different path coefficients; however, it does not put any restriction on the measurement model (i.e. the models

that link the observed variables with their own constructs). Thus, anytime that a significant difference is found and two child nodes are defined, the relationship among latent variables are the same in both ”child” models, but the estimation of each latent variable is recalculated. This consideration introduces the problem of invariance: if the estimations of the latent variables are recalculated (i.e. they could be different) in each terminal node of the tree, we cannot be completely certain that we are correctly comparing the distinct behaviors of two individuals who belong to two different terminal nodes. To solve this problem,we will propose an invariance test based on the χ^2 distribution, where the goal of the test is to verify if the measurement models of each terminal node can be considered equal or not.

3. *Extension of the PATHMOX approach for overcoming the parametric hypothesis of F-test.* There is one criticism of the PATHMOX approach when it is applied in the context of partial least squares path modeling; and that is it utilizes a parametric test based on the hypothesis that the residuals have a normal distribution for comparing two structural models. PLS-PM is generally used to model data that come from the survey analysis. These data are characterized by an asymmetric distribution: when an individual expresses an opinion, for example on a specific service, the opinion will likely be characterized by social bias. This situation produces skewness in the distribution of data. As we well know, it does not matter when we apply the PLS methodology, because one goal of this method is to have no assumptions about the distribution of data. However, it imposes a limitation when we compare PLS models across PATHMOX, because we cannot guarantee the normal distribution of our data, which is a necessary condition for applying the F-test as split criterion. To overcome this limitation, we will extend the test to compare two least absolute deviation robust regressions (LAD) (Koenker and Bassett, 1982) in the context of the PLS path modeling.

4. *Generalization of the PATHMOX algorithm for use in any type of modeling methodology*. The PATHMOX algorithm has been proposed for analyzing heterogeneity in the context of partial least squares path modeling. However, this algorithm can be applied to many other kinds of methodologies according to the appropriate split criterion. To generalize PATHMOX, we will consider three distinct scenarios: regression analysis (OLS, LAD, GLM regression), principal component analysis (PCA) and partial least squares path-modeling (PLS-PM).

5. Implement the methodology using the R software as a specific programming tool.

Chapter 1

Heterogeneity Overview

In many statistical applications, the most common practice is to impose homogeneity restrictions and analyze data as if they were obtained from a single population. However, this hypothesis is often unrealistic because it is unlikely that all individuals in the sample have the same set of parameter values. Consider, for example, marketing and consumer behavior research. Potential sources of heterogeneity can be due to brand awareness, product class knowledge, product usage rate, customer preferences, desire for specificity and benefits. In survey research studies, heterogeneity can be expected among different subgroups defined by gender, groups of age, ethnicity, and marital status. In educational research, assuming homogeneity among a sample of students with varying instructional backgrounds is unrealistic. In natural sciences, a sample may consist explicitly of groups such as experimental and control groups. Test results on a medical test may reflect two types of patients in the sample: those with a disease and those who are healthy. In this chapter we consider the problem of heterogeneity. In the first section, we discuss two different ways of approaching heterogeneity: we will call them heterogeneity *in population* and *in models*. We continue defining the segmentation variables (section (1.2)), analyzing the different sources of heterogeneity (section (1.3)) and the different ways of dealing with heterogeneity (i.e. how two introduce it in our model) (section (1.4)). We conclude (section (1.5)) by presenting a background of the techniques used in classical approaches, which treat heterogeneity in models.

1.1 Approaches to the Heterogeneity Problem

Classically, heterogeneity refers to mixtures of populations forming differentiated clusters, a situation that we may refer to as *population heterogeneity*. A different case is *model heterogeneity*, where the objective is to find segments by following different models. Although at first glance the distinction may seem blurred, it is not: the fundamental point is that cluster analysis does not take into account the hypothesized structural relationships among variables, whereas modeling segmentation does consider the assumed structural relations. Furthermore, there is a different perspective in the two approaches: population heterogeneity starts by using observations to define different groups, whereas model heterogeneity starts with models that define segments. Finally, population heterogeneity also has a predictive purpose: we want to be able to identify an assignment rule that can be generalized for new observations. This differs from model heterogeneity, in which we investigate merely the presence of different models in our data.

1.2 The Nature of a Heterogeneity Variable

We assume that heterogeneity is assignable to third variables, such as socio-demographic variables (e.g., age, gender, family size, occupation and education), geographic variables (e.g., country, state, neighborhood, size of metropolitan area and climate), psychographic variables (e.g., lifestyle, values or personality) and behavioral variables (e.g., usage rate, seeking profits, readiness to buy and user status). From these, we have collected a set of observed covariates that are potential sources of heterogeneity; or, at least, this set of covariates can explain heterogeneity. We call this set segmentation variables in order to distinguish them from the modeling variables.

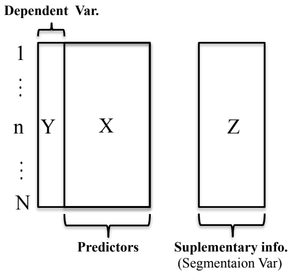


Figure 1.1: The available data structure: *Y* dependent variable, *X* predictors, *Z* supplementary information (i.e.. segmentation variables)

1.3 Sources of Heterogeneity

The source of heterogeneity may be *assignable* or *non-assignable*. Basically, we have assignable heterogeneity when we use segmentation variables to define segments. In the assignable heterogeneity we can distinguish two situations: *known segments* and *unknown segments*. In the first case (see case 1 Table (1.2)), individuals can be divided into segments a priori, and for each of them separate models may be estimated. The hypothesis is that an individual can be uniquely assigned to a single group, based on one or more segmentation variables. The objective is to compare the different groups (corresponding to the different models) that have been identified. In general, the number of groups is known a priori for most applications, and it is based on a limited number of segmentation variables: two or a maximum of three. One example could be a division into segments based on gender (men and women), and social status (low, medium, high). The total number of groups will be six (male with low, medium or high social status, and female with low, medium or high social status). Table (7.1) shows all obtained groups.

Variables	**Social Status**		
Gender	Male-low	Male-medium	Male-hight
	Female-low	Female-medium	Female-Hight

Table 1.1: Obtained groups by segmentation variables gender and social status

In the second case (see case 2 in Table (1.2)), if some segmentation variables are also available and we can divide the individuals into groups, we don't know which partitions must be considered (*unknown segments*). In this case, we need to find some criterion that allows us to identify which are the best partitions. When the heterogeneity is non-assignable beforehand (see case 3 in Table (1.2)), even if we assume the existence of different segments in the data, we just use the modeling variables to define the segments. This involves applying a technique to obtain the class partition (Lubke and Muthén, 2005).

	Assignable	**Non assignable**
Known	Case 1	■
Unknown	Case 2	Case 3

Table 1.2: Classification of heterogeneity by assignable - no assignable, known - unknown

1.4 Modeling Heterogeneity

There are several ways of dealing with assignable heterogeneity: introducing moderating variables in the model, estimating the models into the potential heterogeneous groups, or by a global test comparing the homogeneous model versus the heterogeneity we want to test.

1. **Moderating variables** Classically, heterogeneity is tested by including the segmentation variables as terms in the model (called moderating variables). These represent the heterogeneity we want to test:

$$y = f\Big(\beta_0 + \beta_1' x + \beta_{0U} U + \beta_{1U}' U x\Big) + \varepsilon \tag{1.1}$$

   where x represents the vector of explanatory variables of the model, and U the moderating (or interaction) variable we want to test. Heterogeneity is assessed by means of a statistical test on the coefficients of the interaction terms ($\beta_{0U} = 0$ and $\beta_{1U}' = 0$). In practice, this approach has limited possibilities of exploration; it implies that the source(s) of heterogeneity must be known a priori, with a small number (normally one) of moderating variables (segments are defined by the levels of the moderating variable).

2. **Comparison of coefficients (multi-group approach)** Another way of dealing with heterogeneity is to fit the model in all groups, and then to assess the similarity of the obtained coefficients:

$$y_A = f\Big(\beta_{0A} + \beta_{1A}' x_A\Big) + \varepsilon_A \tag{1.2}$$

$$y_B = f\Big(\beta_{0B} + \beta_{1B}' x_B\Big) + \varepsilon_B \tag{1.3}$$

   where A and B represent the two segments whose homogeneity we want to test. Here, heterogeneity is assessed variable by variable and it forms the model by establishing whether or not the difference in coefficients is significant ($\beta_{0A} = \beta_{0B}$ and $\beta_{1A}' = \beta_{1B}'$). As before, we have limited possibilities of exploration: the source(s) of heterogeneity must be known a priori and we can only deal with a limited number of segments (normally two).

3. **Global comparison of models** We test the identity of two models. One represents the homogeneity assumption and the other represents the heterogeneity we want to test:

$$H_0 : \begin{bmatrix} y_A \\ y_B \end{bmatrix} = f\left[\beta_0 \quad + \begin{pmatrix} x_A \\ x_B \end{pmatrix} + \quad \beta_1\right] + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix} \tag{1.4}$$

$$H_1 : \begin{bmatrix} y_A \\ y_B \end{bmatrix} = f\left[\begin{pmatrix} 1 & x_A & 0 & \\ & 0 & 1 & x_B \end{pmatrix} + \begin{pmatrix} \beta_{0A} \\ \beta_{1A} \\ \beta_{0B} \\ \beta_{1B} \end{pmatrix}\right] + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix} \tag{1.5}$$

   where the null hypothesis represents the homogeneity situation, and the alternative hypothesis represents the heterogeneity that we want to test. Heterogeneity is then assessed by looking at the global statistic of the model comparison (i.e. during regression, it means comparing the sum of squares of residuals for both models: $SSR_0 = SSR_1$). With this approach, we do not need to know beforehand the sources causing heterogeneity. We can deal with a large number of segments and we can adapt the alternative hypothesis to test the heterogeneity that fits our a priori knowledge. In other words, this is a very flexible approach for exploring heterogeneity; but on the other hand, we may run the risk of overfitting.

To explain these three different approaches, let as consider an example. We are interested in fitting the *salary* (y) by *age* (x_1) and *experience* (x_2) of the workers in a big company. For the sake of simplicity, we utilize a multiple regression, but we could generalize this example to other kinds of models. The model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{1.6}$$

Analyzing the data, we suspect that the men earn more than the women (factor of heterogeneity: *sex*) How can we verify if female employees earn less than male employees?

The moderating variable approach introduces heterogeneity as an interaction between predictor variables and a factor. The model can be rewritten as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 F + \beta_4 F x_1 + \beta_5 F x_2 + \varepsilon \tag{1.7}$$

Once the factor is introduced, we can verify if it is significant through a simple test hypothesis (testing the nullity of interaction coefficients):

$$\begin{aligned} &H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ &H_1 : \overline{H_0} \end{aligned} \tag{1.8}$$

The multi-group approach considers one model for each factor level, then afterwards it compares each coefficient in the obtained models. Considering that the factor *sex* presents two levels (*male- M* and *female - F*), we obtain:

$$\begin{aligned} y^M &= \beta_0^M + \beta_1^M x_1^M + \beta_2^M x_2^M + \varepsilon^M \\ y^F &= \beta_0^F + \beta_1^F x_1^F + \beta_2^F x_2^F + \varepsilon^F \end{aligned} \tag{1.9}$$

The test hypothesis (testing the equality of corresponding coefficient) in this case will be:

$$\begin{aligned} &H_0 : \beta_0^M = \beta_0^F \qquad \beta_1^M = \beta_1^F \qquad \beta_2^M = \beta_2^F \\ &H_1 : \overline{H_0} \end{aligned} \tag{1.10}$$

The last possibility is to consider two models: one global, as if data were obtained from a single population; and another that contains the interaction with the factor. In this case we can write the two models as:

$$\begin{aligned} &H_0 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \varepsilon \\ &H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 F + \beta_4 F x_1 + \beta_5 F x_2 + \varepsilon \end{aligned} \tag{1.11}$$

So the test hypothesis (testing of equality of both models) will be:

$$\begin{aligned} &H_0 : SSR_0 = SSR_1 \\ &H_1 : \overline{H_0} \end{aligned} \tag{1.12}$$

If we consider the subgroups obtained by partitioning the data according to factor levels, the equation (1.7) can be rewritten as:

$$\begin{aligned} y^M &= \beta_0 + x_1^M \beta_1^M + x_2^M \beta_2^M + \varepsilon^M \\ y^F &= \beta_0 + x_1^F \beta_1^F + x_2^F \beta_2^F + \varepsilon^F \end{aligned} \tag{1.13}$$

Thus, we can observe that the moderating variable and the multi group approaches can be formulated in the same way. If we now consider the equation (1.11), we can note that the alternative hypothesis has the same expression as the equation (1.7), whereas the null hypothesis is nothing more than the model without considering the factor.

1.5 Heterogeneity Techniques Background

The following overview focuses only on heterogeneity in the model in order to investigate the principal techniques proposed in the literature for dealing with heterogeneity. This classification is made in accordance with three criteria:

- Sources of heterogeneity: assignable and non-assignable a priori.
- Different ways in which heterogeneity can be included: moderating variables, comparing coefficients, global comparison of models.
- Types of variables of interest: observed variables, latent variables.

It is important to remark that this is not an exhaustive overview of all the techniques available for analyzing heterogeneity, but just a background of the classical approaches. A summary table of the classification techniques (see Table (1.3)) is proposed in the last paragraph of the chapter.

1.5.1 Modeling with Assignable Heterogeneity

In the context of observed variables, assignable heterogeneity is analyzed in most of cases by the moderating variables approach. In general, multiple linear regression (MLR) is used to model data, and the interaction terms are tested by T-test. One alternative is based on the global comparison of models approach and is called the F-test, which was proposed independently by Chow (1960) and Lebart (1985). When the variables are non-normal distributed (GLM context), the heterogeneity is also generally analyzed by moderating variables, in this case. The test hypothesis corresponding to the T-test is Wald's test. The *AIC* (Akaike, 1974), *BIC* (Schwarz, 1978) criteria and the likelihood ratio test are other possibilities, and they are related with the global comparison of models approach (McCullagh and Nelder, 1989). When the variables of interest are latent, the classical approach for detecting assignable heterogeneity is the comparison of coefficients approach. Here, the differences can be introduced by considering whether the analysis is confirmative (structural equation modeling (SEM)) or explorative (partial least squares path modeling (PLS-PM)). In the first case, the classical technique is the the multi-group common factor analysis (MG-CFA) (Joresk*ö*g, 1971; Sorbom, 1974). This technique generalizes the CFA situations where the same factor structure is now specified for each level of the categorical variables (i.e.. groups). In Henseler (2007) it is possible to find a classification of PLS-PM multi-group analysis. The principal techniques are: the re-sampling parametric approach (Chin, 2000), the re-sampling non-parametric approach (Chin, 2003), the moderation testing approach (Chin *et al.*, 2003; Henseler and Fassot, 2005; Tenenhaus *et al.*, 2006), and Henseler's approach (Henseler, 2007). An additional method is the possibilistic PLS path modeling (Palumbo and Romano, 2008).

1.5.2 Modeling Assignable Heterogeneity when the Number of Segments is Unknown

A distinct situation is when data come from surveys or researches that contain more information (i.e.. observed heterogeneity) than the one that is used for the models establishment. For instance, in many marketing studies, such as those on customer satisfaction, it is usual to collect socio-demographic variables and psycho-demographic variables like age, gender, social-status, or consumer habits, all of which take no part in the model but can be extremely useful for segmentation purposes. In these cases, even if heterogeneity is observed, we don't know which are the segmentation variables that produced the best groups. In order to incorporate the available external variables, also known as segmentation variables, the PATHMOX algorithm was proposed in 2009 by Gastón Sánchez y Tomas Aluja. In the context of structural equation modeling (SEM), an additional method was recently proposed and is known as the SEM-TREE algorithm (Brandmaier *et al.*, 2013).

1.5.3 Modeling with Non-assignable Heterogeneity

When we are modeling observed variables with non-assignable heterogeneity, almost all the techniques in the literature focus on the possibility of separating the population in subgroups (clustering). Since these techniques do not concern heterogeneity in the models, we will not include them in our classification. When latent variables are considered in the context of SEM, factor mixture models are used (Lubke and Mutheń, 2005). This technique assumes that the heterogeneity is represented by a latent class variable. In PLS, the simplest approach for unobserved heterogeneity consists of a sequential two-step procedure that combines cluster analysis in the first step with a multi-group analysis in the second step: firstly, groups are formed by performing some clustering-based procedure that takes into consideration the cause-effect structure of the models; then, a multi-group analysis is performed among the separate models for each cluster. These approaches include the finite mixture partial least squares (FIMIX-PLS) approach (Hahn *et al.*, 2002; Ringle *et al.*, 2005), and the PLS genetic algorithm segmentation (PLS-GAS) approach (Ringle and Schlittgen, 2007). An interesting approach is the response-based procedure for detecting unit segments (REBUS) approach (Trinchera *et al.*, 2007; Esposito Vinzi *et al.*, 2007, 2008). This technique works by first computing the so-called residuals of the model, and then applying a hierarchical clustering to these residuals. The computation of residuals is intended to take into account the structural part of the analyzed PLS path model. The final step consists of analyzing the local models defined during the cluster procedure.

1.5.4 Summary of Heterogeneity Classification Techniques

The following table (Table (1.3)) summarizes the classification of the techniques according to: the available sources of heterogeneity (assignable – known segments, assignable – unknown segments, non-assignable), the three different ways in which the heterogeneity can be treated (moderating variables, comparison of coefficients, global comparison of models) and the two possible types of variables of interest (latent variables (LV), manifest variables (MV)).

	Heterogeneity classification techniques					
	Assignable				**No Assignable**	
	Known segments		**Unknown segments**			
	MV	**LV**	**MV**	**LV**	**MV**	**LV**
Moderating variables	T-test Wald test					
Comparison of coefficients		MG-CFA Re-sampling parametric Re-sampling no parametric Moderation testing Henselers Possibilistic PLS-PM				Factor Mixture Models REBUS-PLS FIMIX-PLS PLS-GAS
Global Comparison of models	F-test AIC, BIC likelihood ratio test			PATHMOX SEM-TREE		

Table 1.3: Classification of techniques to model heterogeneity according to the sources of heterogeneity, the three different ways in which the heterogeneity can be treated, and the types of variables of interest

Chapter 2

Overview of Methods

Heterogeneity represents an important aspect in almost all statistical methodologies. We have differentiated several ways to approach this problem. In this chapter, we provide some references to the techniques used for analyzing heterogeneity and, when possible, we present a way to compare models through a global comparison approach. We address several modeling approaches that are in accordance with the nature of variables: we consider the regression models (section (2.1)) as models with observed variables; in this context, we introduce linear multiple regression (LRM), quantile regression (LAD) as robust regression, and a generalized linear model (GLM). Principal component analysis (PCA) (section (2.2)) and partial least squares path modeling (PLS-PM) (section (2.3)) are treated as models that fit latent variables. It is important to remark that this not an exhaustive description of techniques, but merely an overview to help the reader understand the context in which we consider the heterogeneity.

2.1 Regression Models

Wage of an employee as a function of her work experience. Price of a house as a function of its number of bedrooms. These are just an example of situations in which we can use regression models (RM) to analyze data. Regression models are one of the most powerful statistical tools for investigating relationships between variables. More specifically, RM help one understand how the value of the dependent variable changes when any one of the independent variables is varied.

The developments leading to the overview of RM extend over more than a century, and they are characterized by vast scientific production. The basics of regression models can be identified in the works of Legendre and Gauss (early nineteenth century - multiple linear regression) and Fisher (1920s → 1935: analysis of variance (ANOVA) and design of experiments (1920), likelihood function (1922), exponential family (1934)).

From 1934 up through 1970, we find many publications on different kinds of RM based on the exponential family. In particular, these studies adapt the regression models to different distributions of dependent variables. In the literature, we can find probit analysis (Bliss, 1935); logit for proportions (Berkson, 1944; Dyke and Patterson, 1952); log linear models for counts (Birch, 1963); and regression models for survival data (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Glasser, 1967). However, it is in 1972 when Nelder and Wedderburn went a step further in unifying the theory of RM by publishing their article on generalized linear models (GLM). They showed how many of the most common linear regression models, listed above, were in fact members of one family and could be treated in the same way: the maximum likelihood estimates for all of these models could be obtained using the same algorithm, iterated weighted least squares. The year 1973 represents another important step forward. The problem of outliers is discussed, and robust regressions are introduced as a possible solution. In this year Huber introduced M-estimation for regression. Least absolute deviation (LAD) - introduced as a criterion in the 19th century - was formalized as an alternative to robust regression (Koenker and Bassett, 1978). In the 1980s, several alternatives to M-estimation were proposed as attempts to overcome the lack of resistance.

Regression models involve the following elements:

- The unknown parameters (also known as coefficients), denoted by β.
- The independent variables or predictors, X.
- The dependent variable, y.
- The link function f, which relates the dependent variable with the predictors and the coefficients. In general, the link function is known; its form depends on the distribution of the dependent variable.
- The ε term expresses the random fluctuation of the dependent variable following a specified probability distribution.

A regression model can be formalize as:

$$y = f(X,\beta) + \varepsilon \tag{2.1}$$

The coefficient β expresses the relationship between dependent variable and predictors. This parameter is unknown and needs to be fitted. The two most common approaches for the estimation of β are the *method of maximum likelihood* (ML) and the *method of least squares* (OLS). Usually, the method of maximum likelihood is used for the estimation of the parameters in generalized linear models, whereas the least squares method is more common in linear regression. OLS and ML give the same results when $f(y)$ is assumed to be multivariate normally distribuated.

2.1.1 Linear Regression Model (LR)

The linear regression model assumes a linear relationship (in parameters) between a dependent variable, y, and a set of explanatory variables, X. We can formalize the model as:

$$y = f(X,\beta) + \varepsilon \quad \Rightarrow \quad y = X\beta + \varepsilon \tag{2.2}$$

where y is a $(n \times 1)$ column vector, X is a $(n \times p)$ matrix of independent variables, β is the vector of parameters and ε is a $(n \times 1)$ column vector of error terms.

The linear regression assumes the following hypothesis:

1. $\{X,y\}$ i.i.d. (independent and identically distributed)
2. The error terms has zero mean: $E(\varepsilon) = 0 \quad \Rightarrow \quad E(y) = X\beta$.
3. The variance of the error is constant and is equal to: $Var(\varepsilon) = E(\varepsilon'\varepsilon) = \sigma^2 I$.
4. The error is independent from the predictors: $Cov(\varepsilon X) = 0$.
5. The error is Normally distributed: $\varepsilon \sim N(0,\sigma^2 I)$.
6. Explanatory variables are fixed (i.e. they are not random variables).

Usually, the estimation of the β coefficients is obtained by ordinary least squares (OLS). To apply this method, we have to find the vector β, such that it minimizes the sum of the square of the difference between the residuals of the observed y and the fitted model $\hat{y}$. Noticing that $\varepsilon = (y - X\beta)$, we have to find the vector β, such that:

$$S(\beta) = E(\varepsilon'\varepsilon) = E((y - X\beta)'(y - X\beta)) = \min \tag{2.3}$$

The resulting OLS estimator of β is:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2.4}$$

2.1.1.1 Testing the Equality of Two Linear Regression Models

The equality of two regression models can be tested by using the hypothesis test introduced by Lebart *et al.* (1979), which is an F-type based statistic. The test is similar to that introduced by Chow (1960), and which was discussed in Ghilagaber (2004) and Moreno *et al.* (2005), for testing the equality between sets of coefficients in two linear regressions. This test is the only one that follows a global comparison approach that compares two models: one that is common to all data (and thus represents a homogeneous situation); and another that includes the interaction between the factor and the predictive variables (which corresponds to a heterogeneous situation). This approach is based on two lemmas: **Lemma 1** and **Lemma 2**, which were introduced by Lebart (see Lebart *et al.*, 1979 pg. 201:202, 208:209). Thus, it is convenient to introduce them before beginning the description of the test.

2.1.1.2 Lemma 1 and Lemma 2

Let us consider a generic linear model in matrix notation:

$$y = XB + \varepsilon \tag{2.5}$$

The residuals are assumed to be normally distributed with zero mean and finite variance, that is, $E(\varepsilon) = 0$ and $V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n$ (with σ^2 unknown).

Lemma 1

Let ε be a normally distributed vector in R^n with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 I_n$

1. Let Q be an $(n \times n)$ symmetric and idempotent matrix. Then, the quadratic form $\varepsilon' Q\varepsilon/\sigma^2$ follows a χ^2 distribution with v degrees of freedom (where v is the rank of Q).

2. Let L be a matrix such that $LQ = 0$. Then, the vectors $L\varepsilon$ and $Q\varepsilon$ follow an independent normal distribution. In particular, the vector $L\varepsilon$ and the variable $\varepsilon' Q_0 \varepsilon/\varepsilon^2$ are independent.

Proof. If the matrix Q is symmetric, there exists Λ and H, so that $Q = H'\Lambda H$, where H is the matrix of eigenvectors, orthonormal of Q ($H'H = HH' = I$), and Λ is the diagonal matrix of the eigenvalues. The eigenvalues are in this case all 0 and 1, due to Q being an idempotent matrix. If we consider $v = H\varepsilon$, then the linear transformation of ε, v follows a normal distribution with $E(\varepsilon) = HE(\varepsilon) = 0$ and $V(\varepsilon) = E(H\varepsilon\varepsilon' H') = \sigma I_n$. Consequently, the reduced component v/σ follows a standardized and independent normal distribution. The quadratic form $\varepsilon' Q\varepsilon/\sigma^2$ can be written:

$$\varepsilon' Q\varepsilon/\sigma^2 = \varepsilon' H'\Lambda H\varepsilon/\sigma^2 = v'\Lambda v/\sigma^2 = \sum_i v_i^2/\sigma^2 \tag{2.6}$$

The last sum includes all terms of the eigenvalues equal to 1(i.e. $tr(Q)$), since the other values are zero. Thus, since χ^2 with p degrees of freedom is defined as the sum of squares of p standardized normal variables, the first point of **Lemma 1** is verified.

Let us consider now the second point of **Lemma 1**. The vectors $L\varepsilon$ and $Q\varepsilon$ are two linear transformations of ε that follow a normal distribution. The covariance COV($L\varepsilon$, $Q\varepsilon$) is null and they are independent

$$COV(L\varepsilon, Q\varepsilon) = E\{(L\varepsilon)(Q\varepsilon)'\} = \sigma^2 LE(\varepsilon\varepsilon')Q = \sigma^2 LQ = 0 \tag{2.7}$$

Since the covariance is null, the second point of **Lemma 1** is verified. Then $\varepsilon' Q\varepsilon = (Q\varepsilon)'(Q\varepsilon)$ is independent of $L\varepsilon$.

□

Lemma 2

Let ε be a vector with normal distribution in R^n with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 I_n$. Given two matrices, X and X_0, where X_0 is defined as $X_0 = XA$ (for any matrix A), we can define the following matrices Q and Q_0 as:

$$Q = I_n - X(X'X)^{-1}X' \tag{2.8}$$

$$Q_0 = I_n - X_0(X_0'X_0)^{-1}X_0' \tag{2.9}$$

It can be shown that the quotient:

$$F = \frac{(\varepsilon' Q_0\varepsilon - \varepsilon' Q\varepsilon) \Big/ (v_0 - v)}{\varepsilon' Q_0 \varepsilon \Big/ v} \tag{2.10}$$

follows an F distribution with $(v_0 - v)$ and v degrees of freedom, where:

- v is the rank of Q
- v_0 is the rank of Q_0

Proof. A geometric representation (see Figure (2.1)) allows to proof **Lemma 2**. Let us consider R_{X_0} to be a sub-space generated from the columns of X_0, the distance of y and R_X to be the longitude of $\varepsilon = Q\varepsilon$. That means:

$$\sum \varepsilon_i^2 = \varepsilon'\varepsilon = \varepsilon' Q\varepsilon \tag{2.11}$$

On the other hand, the relationship $X_0 = XA$ means that the space R_X is generated by the columns of X. The orthogonal projections are represented in Figure (2.1), where R_X is the plane and R_{X_0} is at the right of the plane.

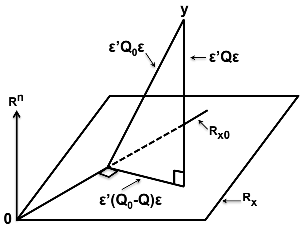


Figure 2.1: Geometric representation of the orthogonal projection of y on the sub-space R_{X_0}

From the **Lemma 1**, $\varepsilon' Q\varepsilon/\sigma^2$ follows χ^2 with Q degrees of freedom. Seeking the law of the numerator:

$$\varepsilon' Q_0\varepsilon - \varepsilon' Q\varepsilon = \varepsilon'(Q_0 - Q)\varepsilon \tag{2.12}$$

Figure (2.1) (theorem of the three perpendiculars) suggests that $Q_0 - Q$ is an orthogonal projection. Thus, $Q - Q_0$ is symmetric and the idempotent propriety remains to be verified:

$$(Q_0 - Q)^2 = Q_0 - Q_0Q - QQ_0 + Q \tag{2.13}$$

Direct calculation shows that:

$$QQ_0 = \{I_n - X(X'X)^{-1}X'\}\{I_n - X_0(X_0'X_0)^{-1}X_0'\} = Q \tag{2.14}$$

Therefore $QQ_0 = Q_0$ and $(Q_0 - Q)^2 = Q_0 - Q$, the matrix is idempotent and its rank is equal to the trace:

$$rank(Q_0 - Q) = tr(Q_0 - Q) = tr(Q_0) - tr(Q) = rank(Q_0) - rank(Q) \tag{2.15}$$

We know that a Fisher distribution is defined as the ratio between two χ^2 independent variables divided by their degrees of freedom and that, due to **Lemma 1**, the numerator of equation (2.22) follows a χ^2 distribution with $(tr(Q_0) - tr(Q))$ degree of freedom; we only have to verify if the numerator and denominator of equation (2.22) are independent. To do this, let us prove that their covariance is equal to zero:

$$COV\{(Q_0 - Q)\varepsilon, Q\varepsilon\} = E\{(Q_0 - Q)\varepsilon(Q\varepsilon)'\} = (Q_0 - Q)E(\varepsilon\varepsilon')Q = \sigma^2(Q_0 - Q)Q = 0 \tag{2.16}$$

Since the covariance is null, **Lemma 2** is verified.

□

2.1.1.3 Formulation of the Test Hypothesis

Let us consider a generic linear model in matrix notation:

$$y = XB + \varepsilon \tag{2.17}$$

The residuals are assumed to be normally distributed with zero mean and finite variance as we stated in (2.2); that is, $E(\varepsilon) = 0$ and $V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n$ (with σ^2 unknown). Let us split the n observations into two groups, A and B of n$_A$ and n_B observations. We want to test if the models are identical (i.e. the equality of all coefficients) for the two sub-populations A and B. Under the alternative hypothesis, H_1, the vectors of p coefficients of the two models (B_A and B_B) are supposedly distinct, that is:

$$y_A = X_A B_A + \varepsilon_A \quad y_B = X_B B_B + \varepsilon_B \tag{2.18}$$

Under the null hypothesis, the coefficients are equal in both segments: $B_A = B_B = B$. We can define a null hypothesis and an alternative hypothesis as:

$$H_0 : \underset{[\mathrm{n},1]}{\begin{bmatrix} y^A \\ y^B \end{bmatrix}} = \underset{[\mathrm{n},p]}{\begin{bmatrix} X^A \\ X^B \end{bmatrix}} \underset{[p,1]}{\begin{bmatrix} \beta \end{bmatrix}} + \underset{[\mathrm{n},1]}{\begin{bmatrix} \varepsilon^A \\ \varepsilon^B \end{bmatrix}} \tag{2.19}$$

$$H_1 : \underset{[\mathrm{n},1]}{\begin{bmatrix} y^A \\ y^B \end{bmatrix}} = \underset{[\mathrm{n},2p]}{\begin{bmatrix} X_1^A & 0 \\ 0 & X_1^B \end{bmatrix}} \underset{[2p,1]}{\begin{bmatrix} \beta^A \\ \beta^B \end{bmatrix}} + \underset{[\mathrm{n},1]}{\begin{bmatrix} \varepsilon^A \\ \varepsilon^B \end{bmatrix}} \tag{2.20}$$

Note that in the null hypothesis we consider that the vector of coefficients β is the same for both segments (A and B); while in the alternative hypothesis, we fit different coefficients (β_A and β_B) for the two models. The matrices X_0, X and A are required by the Lemma 2 of Lebart *et al.* (1979), such that $X = X_0 A$. To obtain the F-statistic:

$$\underset{\substack{[n,p] \\ X_0}}{\begin{bmatrix} X_1^A \\ X_1^B \end{bmatrix}} \quad \underset{\substack{[\mathrm{n},2p] \\ \mathrm{X}}}{\begin{bmatrix} X_1^A & 0 \\ 0 & X_1^B \end{bmatrix}} \quad \underset{\substack{[2p,p] \\ \mathrm{A}}}{\begin{bmatrix} I_p \\ I_p \end{bmatrix}} \tag{2.21}$$

where p is the number of explicative variables for y, and I_p is the identity matrix of order p.

Thus, we can apply Lemmas 1 and 2 of Lebart *et al.* (1979) and test the H_0 hypothesis by computing the following F-statistic with (p) y $(n-2p)$ degrees of freedom.

$$F = \frac{(SSH_0 - SSH_1) \Big/ (n-2p)}{SSH_1 \Big/ p} \tag{2.22}$$

2.1.2 Quantile Regression (LAD)

The linear regression is well-known to be highly sensitive to outlier observations, and as a consequence many robust estimators have been proposed as alternatives. One of the earliest proposals was the least-sum of absolute deviations (LAD) regression (Koenker and Bassett,1978). The LAD estimators are robust to the presence of outliers, and demonstrate to be more efficiency than OLS in cases where the term error does not have a normal distribution. The estimates are found by minimizing the sum of the absolute values of the residuals:

$$\min \sum_{i=1}^{n} |\varepsilon_i| = \min \sum_{i=1}^{n} |y_i - \hat{y}_i| = \min \sum_{i=1}^{n} |y_i - X\beta| \tag{2.23}$$

We would like to use differential calculus to find the LAD estimate of β, but the derivative of the LAD function with respect to β does not exist at the minimum value of β. However, the resulting minimization problem can be solved very efficiently by linear programming methods (Koenker and Bassett, 1978; see also Koenker and d'Orey, 1994; Koenker, 2005).

LAD can be seen as a case of the more general quantile regression. In this case, the objective function to be minimized can be written as:

$$\min \sum_{i=1}^{n} \rho_\alpha(\varepsilon_i) = \min \sum_{i=1}^{n} \rho_\alpha(y_i - X\beta) \tag{2.24}$$

where $\rho_\alpha(\cdot)$ is the tilted absolute value function that yields the α-th sample quantile as its solution. It is defined as:

$$\rho_\alpha(\varepsilon_i) = \begin{cases} \alpha\varepsilon_i & \text{if} \quad \varepsilon_i \geq 0 \\ (1-\alpha)\varepsilon_i & \text{if} \quad \varepsilon_i < 0 \end{cases} \tag{2.25}$$

For a more detailed presentation, excellent references are Birkes and Dodge (1993) and Koenker and Hallok (2001).

2.1.2.1 Testing the Equality of Two LAD Regression Models

Koenker and Bassett introduced a test for testing the equality of two LAD regression models in 1982. This approach can be considered an approximation of the likelihood ratio test when the number of observations is sufficiently large (see also Dielman and Pfaffenberger, 1990). The presentation of the test follows the description of Birkes and Dodge (1993).

Consider a generic linear model in matrix notation:

$$y = XB + \varepsilon \tag{2.26}$$

Let us split the n observations into two groups, A and B, of n$_A$ and n_B observations. As before, we want to test if the models are identical (i.e. the equality of all coefficients) for the two sub-populations A and B. The test hypothesis may be formulated in the same way as the hypothesis test that compares two linear models (see equations (3.46) and (3.47)). When the number of observations is sufficiently large, the test statistic is

$$F_{LAD} = \frac{SAR_{H_0} - SAR_{H_1}}{(p-q)\hat{\tau}} \tag{2.27}$$

where:
- SAR_{H_0} represents the sum of absolute values of the residuals under the null hypothesis.
- SAR_{H_1} represents the sum of absolute values of the residuals under the alternative hypothesis
- $\hat{\tau}$ is the measure of the size of the random error
- p is the number of variables of the model in the alternative hypothesis
- q is the number of variables of the model in the null hypothesis

Thus, when n is large, the F_{LAD} follows an F-distribution with 1 and $p-q$ degrees of freedom.

2.1.3 Generalized Linear Model (GLM)

Linear models are one of the most used approaches in statistical applications. However, there are many situations in which we cannot assume normal distribution of data. In many cases, we find it is more realistic to fit the observed variable as a discrete variable. For instance, when we analyze count processes in which the dependent variable can assume only positive values, (i.e. number of deaths for a specific disease) it is more suitable to fit the data considering a Poisson distribution than a Normal distribution. The generalized linear model (GLM) allows removal of the normal hypothesis from the dependent variable and to fit the data by adopting discrete distributions. In a GLM, each outcome of the dependent variable y is assumed to be generated from a particular distribution of the *exponential family*, a large range of probability distributions that include Normal, Binomial and Poisson, among others. For a more detailed presentation and discussion, excellent references are Dobson (1990) and McCullagh and Nelder (1989).

The mean, μ, of the distribution depends on the independent variable X and is obtained through the relation:

$$E\Big(y\Big) = \mu = g^{-1}\Big(X\beta\Big) \tag{2.28}$$

where $E(y)$ is the expected value of y; $X\beta$ is the linear predictor (a linear combination of unknown parameters); and g is the link function.

The GLM model consists of three elements:

1. *A probability distribution from the exponential family*

   $y = (y_1, \cdots, y_n)$ are independent random variables with distribution from the exponential family. The density function of an exponential family has the general form:

   $$f\Big(y, \theta, \phi\Big) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{2.29}$$

   where θ is known as a canonical parameter and ϕ is a fixed (known) scale (dispersion) parameter. Note that $a(\cdot)$ and $b(\cdot)$ are some specific functions that distinguish one member of the exponential family from the others.

2. *A linear predictor*

   The linear predictor is the quantity that incorporates the information about the independent variables into the model. It is related to the expected value of the data (thus, "predictor") through the link function. η is expressed as a linear combination of unknown parameters β. η can be expressed as $\eta = X\beta$.

3. *A link function*

   The link function provides the relationship between the linear predictor and the mean of the distribution function. If we indicate with μ the mean value of random variable y, it is related with the linear predictor by the function $g(\cdot)$:

   $$g\Big(\mu\Big) = \eta \Leftrightarrow \mu = h\Big(\eta\Big) \tag{2.30}$$

   where $h(\cdot) = g^{-1}(\cdot)$

The unknown parameters, β, are typically estimated with the maximum likelihood method. Following the general form of the exponential family (see equation 2.29), we can define the Maximum likelihood as:

$$
\begin{aligned}
log\Big(\theta,\phi,y\Big) &= log \prod_{i=1,\cdots,n} f_{Y_i}\Big(y,\theta,\phi\Big) \\
&= \sum \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\} \\
&= \sum \frac{y\theta - b(\theta)}{a(\phi)} + \sum c\Big(y,\phi\Big)
\end{aligned}
\tag{2.31}
$$

The maximum of the ML function is obtained by differentiating the log-likelihood function with respect to each element θ_j of θ and solving the simultaneous equations:

$$
\frac{\partial \mathscr{L}(\theta)}{\partial \theta_j} = 0 \quad j = 1,\cdots,p \tag{2.32}
$$

In general, the equation (2.32) cannot be solved directly and an iterative procedure is used. The most common methods are the Newton-Raphson and iteratively reweighed least squares (see Nelder and Wedderburn,1972); McCullagh and Nelder,1989).

2.1.3.1 Testing the Equality of Two General Linear Models

The comparison between two general linear models under the global comparison approach can be performed in three ways: the likelihood ratio test, the AIC (Akaike, 1974) and BIC (Schwarz, 1980) criteria. The first one is a test hypothesis based on the χ^2 statistic; the second and the third can be considered an indexes that measure the goodness of fit of a model: the lower the indeces are, the better the goodness of fit. Thus, the comparison is made by choosing the model with the lowest index value.

The likelihood ratio test

The likelihood ratio test is a criterion used for comparing two nested models. The form of the test is suggested by its name:

$$
LRT = -2log_e\left(\frac{\mathscr{L}(\hat{\theta}_s)}{\mathscr{L}(\hat{\theta}_g)}\right) \tag{2.33}
$$

the ratio of two likelihood functions. The simpler model (s) has fewer parameters than the general (g) model. Asymptotically, the test statistic is distributed as an χ^2 distribution, with degrees of freedom equal to the difference in the number of parameters between the two models. LRT can be presented as the difference in the log-likelihoods (recall that $log(A/B) = logA - logB$), and this is often handy, since they can be expressed in terms of deviance. Then:

$$
LRT = -2\Big(log_e(\mathscr{L}_s) - log_e(\mathscr{L}_g)\Big) = -2log_e(\mathscr{L}_s) + 2log_e(\mathscr{L}_g) = D_s - D_g \tag{2.34}
$$

where D is defined as the scaled deviance (Wedderburn, 1972) of the model.

Thus, the LRT can be computed as the difference in the scaled deviance of the two models. Let us assume that we want to compare two models: M_1, (that represent the model of null hypothesis) with p_1 parameters, and M_2 (that represent the model of alternative hypothesis), with p_2 parameters, where $M_1 \subset M_2$ and $p_2 > p_1$. We then compute MLs for both models and the respective deviances (D):

$$
D_{M1} = -2log_e(\mathscr{L}(M_1)) \quad D_{M2} = -2log_e(\mathscr{L}(M_2)) \tag{2.35}
$$

The LRT will be:

$$LRT = D_{M1} - D_{M2} \tag{2.36}$$

Then, under the null hypothesis that M_2 is the true model, the difference between the deviances follows an approximate χ^2_v. Where $v = p_2 - p_1$ is the difference of the parameters of the model M_1 and M_2.

The AIC and BIC Criteria

AIC and *BIC* are two criteria also related to the likelihood function. They are defined as:

$$AIC = -2\mathscr{L}(\hat{\beta}, y) + 2p \tag{2.37}$$

$$BIC = -2\mathscr{L}(\hat{\beta}, y) + p\log n \tag{2.38}$$

where p is the number of parameters and n is the number of observations. The AIC (Akaike, 1974) is the Kullback-Leibler divergence between the true model and the best choice inside a family. The BIC (Schwarz, 1980) criterion is the Bayesian choice between models: the most probable a posteriori model is the one with minimal BIC. Given a set of candidate models for the data, the preferred model is the one with the minimum index value. Hence, these indices not only reward goodness of fit, but also include a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting (increasing the number of parameters in the model almost always improves the goodness of the fit).

2.2 Principal Component Analysis (PCA)

The principal component analysis is a multidimensional analysis technique that seeks to express a set of variables using a reduced number of uncorrelated variables called components. PCA was originally introduced by Pearson (1901) and independently by Hotelling (1933). Rao's (1964) paper is remarkable for the large number of new ideas concerning uses, interpretations and extensions of PCA. Gower (1966) discussed links between PCA and various other statistical techniques, and also provided a number of important geometric insights. Finally, Jeffers (1967) gave impetus to the actual practical side of the subject by discussing two case studies in which the use of PCA goes beyond that of a simple dimension-reducing tool. An extensive review of the topic can be found in most textbooks on multivariate techniques, such as in Escofier and Pagés (1998), Lebart, Morineau and Piron (2000), Timm (2002), Saporta (2006) and Tenenhaus (2006). For a more detailed presentation and discussion, excellent references are Jackson (1991), Aluja and Morineau (1999), and Jolliffe (2002).

Let X be a matrix of n-observations, p-variables and a-rank. The goal is to reduce the number of variables of interest $(x_1, x_2, \cdots, x_p,)$ into a smaller set of components $(c_1, c_2 \cdots c_a)$, obtained as linear combination of the X variables, with each being c_a of maximal variance. The linear relationships between the components and the variables are:

$$\begin{aligned}
c_1 &= u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p \\
c_2 &= u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p \\
c_3 &= u_{13}x_1 + u_{23}x_2 + \cdots + u_{p3}x_p \\
&\vdots \\
c_a &= u_{1a}x_1 + u_{2a}x_2 + \cdots + u_{pa}x_p
\end{aligned} \tag{2.39}$$

In matrix notation we have:

$$C = XU \tag{2.40}$$

where:

- C is a matrix $(n \times a)$ of components
- U is a matrix $(a \times p)$ of vectors indicating the direction of the components

The new variables are derived in decreasing order of importance in the sense that c_1 accounts as much as possible for the variation in the original data amongst all linear combinations of $(x_1, x_2, \cdots, x_p)$. Then, c_2 is chosen to account as much as possible for the remaining variation, subject to being uncorrelated with c_1, and so on. The new variables, $(c_1, c_2 \cdots c_a)$ defined by this process are the principal components. Figure (2.2) shows a graphical representation of a PCA problem:

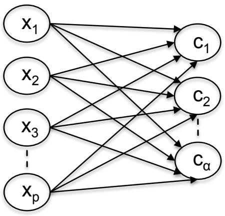


Figure 2.2: Representation of a PCA problem with a path diagram

2.2.1 Derivation of the Principal Components

Let us consider a matrix X with n observations and p variables. Given this matrix, we can define a variable space R^p with as many dimensions as variables. Each variable represents one co-ordinate axis. Each observation (each row) of the X-matrix is placed in the p-dimensional variable space. Consequently, the rows of the matrix X form a swarm of points in this R^p space. Let us apply a mean-centering to our data. This procedure involves the subtraction of the variables' average from the data, and it corresponds to the re-positioning of the coordinate system, such that, the average is now the origin of the axes in R^p. Let us denote the new transformation of R^p as $\bar{R^p}$. After mean-centering, the data set is ready for the computation of the first principal component c_1. This component is the line in $\bar{R^p}$ space that, passing through the average point, best approximates data in the least squares sense (i.e it is the line that maximizes the variation in the swarm of points).

Algebraic Derivation

Let us consider the linear combination $C = XU$. The first principal component, c_1, will be the linear combination $c_1 = Xu_1$ which maximizes the variance of the swarm of points, where u_1 is a vector of p-coefficients. The condition of c_1 with maximum variance implies looking for u_1, which maximizes:

$$var(Xu_1) = E(Xu_1)^2 = E(u_1'X'Xu_1) = u_1'Su_1 = \text{max} \tag{2.41}$$

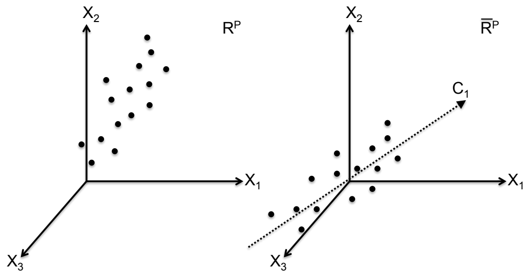


Figure 2.3: Geometrical derivation of the first principal component c_1. The R^p space represents the space defined by the original variables. The $\bar{R^p}$ space represents the space defined by the original variables with centered observations

where S is the variance-covariance matrix. It is clear that the maximum will not be achieved for finite u_1, so a normalization constraint must be imposed. The goal is to maximize $u_1'Su_1$ subject to $u_1'u_1 = 1$. Using Lagrange multipliers, we have to maximize:

$$u_1'Su_1 - \lambda(u_1'u_1) = \max \tag{2.42}$$

Differentiation with respect to u_1 gives:

$$Su_1 - \lambda u_1 = 0 \quad \text{or} \quad Su_1 = \lambda u_1 \tag{2.43}$$

Maximization implies that λ must be as large as possible; that is, u_1 is the eigenvector corresponding to the largest eigenvalue, λ_1, of S.

The second principal component, c_2, is the linear combination $c_2 = Xu_2$, which is uncorrelated with c_1. It has maximum variance, so that the k-th component $c_k = Xu_k$ is found to have maximum variance, subject to being uncorrelated with the previous components $c_1, c_2, \cdots, c_{k-1}$. Up to $a \leq p$ *PCs* could be found, but it is hoped that most of the variation in X will be accounted for by m *PCs* $(m << a)$.

2.3 Partial Least Squares Path Modeling (PLS-PM)

PLS path modeling is a statistical method that has been developed for analyzing latent variables in the context of structural models. The objective of PLS is to obtain scores of latent variables for predictive purposes without using the model for explaining the covariation of all the indicators. According to Chin (1998), parameter estimates are obtained based on the ability to minimize the residual variances of all dependent variables (both latent and observed). In Wold's (1975 a) paper, the main principles of partial least squares for principal component analysis (Wold, 1966) were extended to situations with more than one block of variables. Other presentations of PLS Path Modeling given by Wold appeared in the same year (Wold,1975 b, c). Wold (1980) provides a discussion on the theory and the application of partial least squares for path models in econometrics. The specific stages of the algorithm are well described in Wold (1982) and in Wold (1985). An important reference was provided by Lohm*ö*ller in 1989, when he published the book Latent Variable Path Modeling with Partial Least Squares. The book contains his research results and presents a detailed and wide-ranging account of the capabilities of PLS-PM. In his book, statistical, modeling, algorithmic, and programming aspects of the PLS methodology are treated in great depth. He also extended the basic PLS algorithm in various directions to show the scope of problems that can be handled with PLS. Extensive reviews on the PLS approach to structural equation models with further developments are given in Chin (1998) and in Tenenhaus *et al.* (2005). They describe the PLS-PM model's characteristics as an algorithm that allows fitting the parameters, which they base on the *Handbook of Partial Least Squares* (Esposito Vinzi *et al.*, 2010).

2.3.1 Latent Constructs

PLS-PM is a method that allows us to analyze complex systems of relationship between variables. In this context, we can differentiate *manifest variables* (MV) – which are defined also as observed variables or indicators – and *latent variables* (LV), which are also known as constructs or factors. The latent variables can also be distinguished as *endogenous* (LV that can be at the same time dependent and predictors of other latent variables) and *exogenous* (LV that can be only predictors of other latent variables). The MV and LV can be related in two distinct ways: *reflective* and *formative*. In the first case, manifest variables are considered to be caused by the latent variables. In the second, the latent construct is supposed to be formed by its indicators (Diamantopoulos and Winklhofer, 2001). The main difference between reflective and formative patterns is related to the cause-effect relationships between indicators and their constructs. To better understand this concept, let us consider, as an example, the performance of a firm; we can create a reflective scale that measures a manager's views on how well the firm is performing, or we can create a set of metrics for firm performance that measure various elements, such as profitability, return on equity, market share, etc.

2.3.2 Data Structure

PLS path modeling aims to estimate the relationships among $Q(q = 1,\cdots,Q)$ blocks of manifest variables. Each block q is an expression of one unobservable construct. In each block q, we assume P variables $P(p = 1,\cdots,P)$ observed on N units $(n = 1,\cdots,N)$. The resulting data (x_{npq}) are collected in a column partitioned data table X:

$$X = \Big[X_1 \cdots, X_q \cdots, X_Q \Big] \tag{2.44}$$

where X_q is the generic q-th block made of P_q variables.

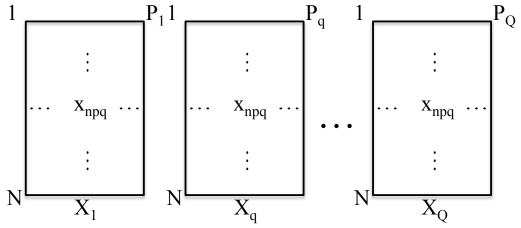


Figure 2.4: Data structure in PLS-PM

2.3.3 Model Specification

The PLS model can be divided into two parts: the *structural model* (inner model), which relates the endogenous latent variables with other latent variables, and *measurement model*, also known as outer model, that analyzes the relationships between manifest and latent variables.

Structural Model

The structural model describes the cause-effect relationships between the latent variables. Associations between LV can be represented by a recursive system of linear equations. The LV can play the role of response variable (endogenous latent variable) or explanatory variable (exogenous latent variable). For simplicity we will not adopt a different notation for these two types of latent variables. The expression of the structural model is:

$$\xi_j = \beta_{0j} + \sum_{q:\xi_q \to \xi_j} \beta_{qj}\xi_q + \zeta_j \tag{2.45}$$

where ξ_j with $(j = 1, \cdots, J)$ is the generic endogenous latent variable, the parameter β_{qj} is defined as a path coefficient (representing the link between the j-th and q-th latent variable) and ζ_q is the inner residual term. Assuming the condition of the specification of the predictor, which assures desirable estimation properties in classical ordinary least squares (OLS) model, we obtain:

$$E(\xi_j) = \beta_{0j} + \sum_{q:\xi_q \to \xi_j} \beta_{qj}\xi_q \tag{2.46}$$

The predictor specification condition involves:

$$E(\zeta_q) = E(\xi_j \zeta_q) = 0 \tag{2.47}$$

which means that the residuals have zero mean and are uncorrelated with the LVs.

Measurement Model

The measurement model formulation depends on the direction of the relationships between the latent variables and the corresponding manifest variables (Fornell and Bookstein, 1982). As a matter of fact, different types of measurement models (see Figure(2.5)) are available: the *reflective model* , *the formative model* and the *MIMIC model* (a mixture of the two previous models).

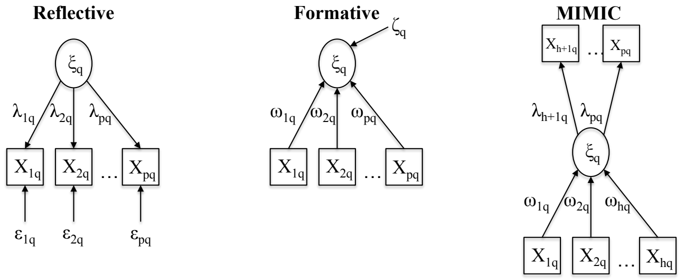


Figure 2.5: Path diagram of the three possible relationships between latent variables and manifest variables. In order from left to right: reflective, formative, and MIMIC

In a *reflective model*, the block of manifest variables related to a latent variable, is assumed to measure a unique underlying concept. Each MV reflects (i.e. is an effect of) the corresponding LV. In more formal terms, in a reflective model each MV is related to the corresponding LV by a simple regression model:

$$x_{pq} = \lambda_{p0} + \lambda_{pq}\xi_q + \varepsilon_{pq} \tag{2.48}$$

where λ_{p0} is the intercept, λ_{pq} is the loading associated to the p-th manifest variable in the q-th block and the error term ε_{pq} represents the imprecision in the measurement process. Standardized loadings are often preferred for interpretation purposes, due to the fact they represent correlations between each manifest variable and the corresponding latent variable. An assumption (predictor specification) behind this model is that the error ε_{pq} has zero mean and is uncorrelated with the latent variable of the same block, that is: $E(x_{pq}|\xi_q) = \lambda_{p0} + \lambda_{pq}\xi_q$ and $E(\varepsilon_{pq}\xi_q) = 0$.

In the *formative model*, each MV of each sub-block of MV represents a different dimension of the underlying concept. The latent variable is defined as a linear combination of the corresponding manifest variables. Thus the measurement model could be expressed as:

$$\xi_q = \sum_{p=1}^{P_q} \omega_{pq} x_{pq} + \delta_q \tag{2.49}$$

where ω_{pq} with $(p = 1, \cdots, P_q)$, is the coefficient linking each MV of the q block to the corresponding LV, and the error term δ_q represents the fraction of the corresponding latent variable not accounted for the block of manifest variables. Also, in this case we assume the predictor specification, which in this case will be: $E(\xi_q|x_{pq}) = \sum_{p=1}^{P_q} \omega_{pq} x_{pq}$ and $E(\varepsilon_{pq}\xi_q) = 0$.

The *MIMIC* can be considered a mix of reflective and formative ways. In this case there are two linear equations:

$$x_{hq} = \lambda_{h0} + \lambda_{hq}\xi_q + \varepsilon_{hq} \quad \text{for} \quad h = 1 \quad \text{to} \quad p_1 \tag{2.50}$$

$$\xi_q = \sum_{h=p_1+1}^{P_q} \omega_{hq} x_{hq} + \delta_q \tag{2.51}$$

the p_1 first MVs follow a reflective way and $(p - p_1)$ the last ones a formative way. The predictor specification hypotheses are both assumed by the reflective and formative ways.

2.3.4 Algorithm Estimation

In PLS path modeling an iterative procedure permits estimation of the outer weights $\hat{\omega}_{pq}$ and the latent variable scores $\hat{\xi}_q$. The estimation procedure is named partial, since it solves blocks one at a time by means of alternating single and multiple linear regressions. The path coefficients β_{pq} are estimated afterwards by means of a regular regression between the estimated latent variable scores in accordance with the specified network of structural relations.

The procedure can be summarized in five main steps:

Imput: $X = \begin{bmatrix} X_1 \cdots, Xq \cdots, XQ \end{bmatrix}$ i.e. Q blocks of centered manifest variables

1. **step** *Outer estimation*: The latent variables are defined as linear combination of their manifest variables

2. **step** *Inner estimation*: The latent variables are defined as linear combination of the adjacent latent variables

3. **step** *Updating Outer Weights*: the outer weights are updated according to the internal and external estimation

4. **step** *Convergence*: outer estimation and inner estimation are compared, and they must converge. If not, the algorithm restarts with **step 1**

5. **step** *Coefficients estimation*: the path and loading coefficients are estimated according to the inner and outer models.

Output: $\left[\hat{\omega}_{pq}, \hat{\xi}_q, \hat{\beta}_{pq}\right]$

Step 1 Outer Estimation

In the first step, the latent variables are defined as linear combinations of their manifest variables. In mode A, this estimation is done trying to obtain a set of weights able to estimate a latent variable accounting for as much variance as possible for the indicators and the construct. In the case of mode B estimation we can't assure high cummunalities for these blocks. The weights are scaled to give $\hat{\xi}_q$ unit variance. This standardization is done to avoid scale ambiguity of the LV. The standardized latent variables are defined as Y_q. We can calculate the Y_q as:

$$\hat{\xi}_q = Y_q = \pm\left(\sum_{p=1}^{P_q} \omega_{pq} x_{pq}\right) \tag{2.52}$$

where the ± sign shows the sign ambiguity. This ambiguity is solved by choosing the sign that makes Y_q positively correlated to the majority of x_{pq}:

$$sign\left(\sum_{p=1}^{P_q} sign\left[cor\left(x_{pq}Y_q\right)\right]\right) \tag{2.53}$$

The standardized LV is finally expressed as:

$$Y_q = \sum_{p=1}^{P_q} \omega_{pq} x_{pq} \tag{2.54}$$

The algorithm begins with an initial outside approximation of the LV by using arbitrary weights, which are scaled to obtain unit variance. Following Chin's suggestion (1999), we can set initial weights with equal value to perform a first approximation of the latent variable as a simple sum of its indicators. This option is based on a scenario in which the researcher, having no additional information, would obtain the best first approximation of the LV as a summation of the MV.

Step 2 Inner Estimation

In this step, the latent variables in the inner model are defined as linear combinations of the adjacent latent variables. Only the adjacent latent variables are considered means that, estimating of each LV, we consider only those have direct relationships with it. We indicate them with the index q'. The internal estimation Z_q, the name of the standardized ξ_q, is defined by:

$$Z_q = \sum_{\substack{q' : \beta_{qq'} \neq 0 \\ \beta_{q'q} \neq 0}} \left(\varepsilon_{qq'} Y_q\right) \tag{2.55}$$

where $\varepsilon_{qq'}$ are the inner weights which are assumed to be scaled so that the variable in parentheses is standardized. There are three options for calculating the inner weights:

1. The Centroid Scheme

$$\varepsilon = \begin{cases} sign\left(cor\left(Y_q, Y_{q'}\right)\right) & \xi_q \xi_{q'} \quad adjacents \\ 0 & otherwise \end{cases} \tag{2.56}$$

   This scheme considers only the sign direction of the correlations between am LV and its adjacent (neighboring) LVs. It does not consider the direction or the strength of the paths in the structural model. Some problems may be present when a correlation is close to zero, causing sign changes during the iterations from $+1$ to -1.

2. The Factor Scheme

$$\varepsilon = \begin{cases} cor\left(Y_q, Y_{q'}\right) & \xi_q \xi_{q'} \quad adjacents \\ 0 & otherwise \end{cases} \tag{2.57}$$

   The factor scheme uses the correlation coefficient as the inner weight. It considers the strength of the paths in the structural model.

3. The Path Scheme.

   $\varepsilon_{qq'}$ is calculated as a coefficient of the multiple regression of Y_q on Y_q' if ξ_q is predictor of ξ_q'. The path weighting scheme has the advantage of taking into account both the strength and the direction of the paths in the structural model. However, this scheme presents some problems when the LV correlation matrix is singular.

Step 3 Updating Outer Weights

We can conceive the inside approximation as a stage in which the information contained in the inner relations is incorporated into the estimation process of the latent variables. Once the inside approximation is done, the internal estimates Z_q must be considered with regard to their indicators. This is done by updating the outer weights ω_{pq}. There are basically two ways of calculating the outer weights: *mode A*, and *mode B*. Each mode corresponds to a different way of relating the MVs with the LVs in the theoretical model. *Mode A* is used when the indicators are related to their latent variables through reflective way. Instead, *mode B* is preferred when indicators are associated with their latent variables in a formative way. However, there is also a third option called *mode C*, which is rarely used in practice and it is supposed to be used when the indicators of an LV are connected by MIMIC way.

1. Mode A

   In the reflective way, each weight ω_{pq} is the regression coefficient of Z_q in the simple regression of x_{pq} on Z_q, i.e. the simple regression $x_{pq} = \omega_{pq} Z_q$ where:

$$\omega_{pq} = \left(Z_q' Z_q\right)^{-1} Z_q' x_{pq} = COV\left(x_{pq}, Z_q\right) \tag{2.58}$$

2. Mode B

   In the formative way, Z_q is regressed on the block of indicators related to the latent construct X_q, and the vector ω_q of weights ω_{pq} is the regression coefficient in the multiple regression:

$$\omega_{pq} = \left(X_q'X_q\right)^{-1} X_q'Z_q \tag{2.59}$$

3. Mode C

   Mode C is implemented in Lohm$\ddot{o}$ller's version, and it is a special case of Mode B. The MIMIC way is a kind of mix between reflective and formative ways, so the path coefficients for the h MVs related in a reflective way are estimated by a simple linear regression: $x_{hq} = p_{hq}Y_q$ for $h = 1$ to p_1; and the path coefficients for the for $h = p_1 + 1$ to P. MVs related in a formative way are estimated by a multiple linear regression:

$$Y_q = \sum_{h=p_1+1}^{P_q} \omega_{hq}x_{hq} \tag{2.60}$$

Step 4 Check for Convergence

In every iteration step, say $S = 1, 2, 3, \cdots$, convergence is checked by comparing the outer weights of step S against the outer weights of step $S-1$. For example, Wold (1982) proposed $|\omega^{\hat{S}-1}{}_{jk} - \hat{\omega}^{S}{}_{jk}| < 10^{-5}$ as a convergence criterion. Although convergence of the PLSPM algorithm is still an open issue, some advances in understanding this aspect have been achieved in the last decade. On the one hand, the algorithm always converges with two blocks. On the other hand, for three or more blocks there is no guarantee that it will always converge (although in practice it usually does). Interestingly, Hanafi (2007) has proved that convergence is achieved when using Mode B and what he calls "Wold's procedure" (as opposed to Lohm$\ddot{o}$ller's procedure), provided the data blocks in Mode B are of full rank.

Step 5 Coefficient Estimation

This stage of the algorithm consists of calculating of the path and loading coefficient estimates, $\hat{\beta}_{qq'}$ and $\hat{\lambda}_{qq'}$, according to the inner and outer models. For the structural model the path coefficients are estimated by ordinary least squares in the multiple regression of Y_q on the Y_q''s related to it,

$$Y_q = \sum_{q'} \beta_{qq'}Y_{q'} \tag{2.61}$$

$$\beta_{qq'} = \left(Y_q'Y_q\right)^{-1} Y_q'Y_{q'} \tag{2.62}$$

For the measurement model, the loading coefficients are estimated depending on their corresponding way. In the reflective way, the loading coefficients are the regression coefficients of the simple linear regression of x_{pq} on Y_q:

$$x_{pq} = \lambda_{pq} Y_q \tag{2.63}$$

$$\lambda_{pq} = \left(Y_q' Y_q\right)^{-1} Y_q' x_{pq} \tag{2.64}$$

In the formative way, the weight coefficients ω_{pq}'s coincide with the outer weights obtained in the first stage. This is because we perform the multiple linear regression of Y_q on the x_{pq}:

$$Y_q = \sum_{p=1}^{P_q} \omega_{pq} x_{pq} \tag{2.65}$$

$$\omega_{pq} = \left(X_q' X_q\right)^{-1} X_q' Y_q \tag{2.66}$$

2.3.5 PLS - Validation

PLS path modeling lacks a well identified global optimization criterion, so there is no global fitting function to assess the goodness of the model. Furthermore, it is a variance-based model strongly oriented to prediction. Thus, model validation focuses mainly on the model's predictive capability. According to the PLS-PM structure, each part of the model needs to be validated: the measurement model, the structural model and the overall model. A detailed description of PLS validation can be found in Tenenhaus *et al.* (2005) and in the *Handbook of Partial Least Squares* (Esposito Vinzi *et al.*, 2010).

2.3.6 Tenenhaus' Contribution of PLS as a Special Case of the Regularized Generalized Canonical Correlation Analysis (RGCCA)

In PLS path modeling, we consider that each block X_j is the expression of an unobserved latent variable (LV), and that structural relations (i.e. multiple regression equations) exist between the latent variables. We consider that two blocks X_j and X_k are connected if the associated latent variables are related: LV(X_j) explains LV(X_k) or vice versa. Let us define the design matrix $C = \{c_{jk}\} : c_{jk} = 1$ for when blocks X_j and X_k are connected in the structural equation model; otherwise it is 0. We know that the PLS algorithm depends on two modes (A and B) for the latent variables' outer estimation, and at least three schemes (centroid, factorial and structural) for the latent variables' inner estimation. When selecting either Mode B or A and either the centroid or factorial schemes, the inner and outer latent variable estimations are identical to the inner and outer components of RGCCA (Tenenhaus, 2011).

RGCCA and likewise the GCCA represent a framework for modeling linear relationships between several blocks of variables observed for the same set of individuals. Considering a network of the connections between these blocks, the objective of RGCCA is to find linear combinations of block variables, called components, such that the block components explain their own block well, and/or the block components that are assumed to be connected are highly correlated. The principal difference regarding the GCCA is represented by the introduction of a τ_j shrinkage constant (Ledoit and Wolf, 2004) for the estimation of the $Var(X_j a_j)$. This constant allows us to overcome the limitation on the number of observations (we know that to estimate the $Var(X_j a_j)$ correctly, we need, at least, $n_j > p_j$). In RGCCA, the objective can be formalized as:

$$\begin{aligned} &\max \sum_{j,k=1, j\neq k}^{J} c_{jk} g(COV(X_j a_j, X_k a_k)) \\ &\text{subject to the constraints} \quad \tau_j \parallel a_j \parallel^2 + (1-\tau_j) Var(X_j a_j) = 1, \quad j = 1, \ldots, J \end{aligned} \tag{2.67}$$

where X_j and X_k represent two generic blocks; a_j and a_k are the outer weights that allow generating the two generic latent variables; and τ_j is a shrinkage constant that assumes values between 0 and 1.

The equation (2.67) has no analytical solution. Tenenhaus proposed the PLS algorithm in 2011 to solve this problem. If we consider a generic block X_j, the initial step of the algorithm defines the weights a_j in an arbitrary way. The outer component (i.e. outer PLS estimation) that summarizes the block information is calculated as: $y_j = X_j a_j$, in accordance with $\tau_j \parallel a_j \parallel^2 + (1-\tau_j)Var(X_j a_j) = 1$. The inner component (i.e. inner PLS estimation), which takes into account the relation between the blocks, is obtained as: $z_j = \sum_{k \neq j} e_{jk} y_k$, where the inner weights e_{jk} are defined as:
- e_{jk} = $c_{jk} sign(cor(y_k y_j))$ $\rightarrow$ Centroid scheme
- e_{jk} = $c_{jk} cor(y_k y_j)$ $\rightarrow$ Factorial scheme

The outer weights a_j is updated as:

$$a_j = \frac{[\tau_j I + (1-\tau_j)\frac{1}{n}X_j' X_j]^{-1} X_j' z_j}{\sqrt{z_j X_j [\tau_j I + (1-\tau_j)\frac{1}{n}X_j' X_j]^{-1} X_j' z_j}} \tag{2.68}$$

These steps are iterated until the convergence of the criterion.

When $\tau_j = 0$, Mode B and the centroid or factorial schemes are selected for the inner estimation, and when $\tau_j = 1$ (however, with Mode A selected for weight estimation in this case), the inner and outer components of RGCCA (Tenenhaus, 2011) are equal to the inner and outer latent variable estimation obtained by the PLS algorithm of Wold (1982).

Chapter 3

PATHMOX algorithm

In this chapter, we present the PATHMOX algorithm. We start by introducing the heterogeneity problem (section (3.1)) as one of the principal reasons for considering the existence of different sub-groups in a dataset and, consequently, of the need to apply different models to each one of them. We describe the PATHMOX (section (3.8.1) and (3.8.1)) algorithm as a possible solution to the heterogeneity problem. Here, a brief introduction on segmentation trees is proposed. We continue by analyzing the F-test split criterion (section (3.4)), utilized by the algorithm in the context of PLS-PM, and we propose three different extensions: in the first one (section (3.5)), we discuss the possibility of achieving which constructs differentiate segments; in the second one (section (3.6)), we propose a non-parametric approach to overcoming the parametric hypothesis of the F-test; in the third one (section (3.7)), we analyze a possible solution for overcoming the invariance problem. We conclude the chapter by discussing how to generalize the PATHMOX approach to other methodologies (section (3.9)). To this end, we consider two different frameworks: (1) multiple regression (section (3.10)), which takes into account ordinary least squares, least absolute deviation and generalized linear regression; and (2) principal component analysis (section (3.11)).

3.1 The Heterogeneity Problem

A common practice, in data collection is to store some additional information as socio-demographic variables, for example: sex, social status, or age. For instance, in survey analysis, it is usual to consider some generic information about the respondents that goes beyond the variables of interest, such as geographic, demographic, psychographic or behavioral factors. These variables, known as segmentation variables, represent an important source of heterogeneity that must be considered through analysis. In general, if we have some previous knowledge regarding which of these factors produces significant differences in the results of the analysis, or if we do not have any previous knowledge but only one or two segmentation variables are collected, we can analyze this source of heterogeneity by the splitting data in accordance with the segmentation variables. We can then consider a different model for each obtained group. However, the segmentation variables in almost all cases are more than two, and we don't have any previous knowledge about the factors. Therefore, we cannot identify the groups a priori.

In order to avoid losing all this information, an algorithm solution can be considered. We propose a solution based on the segmentation binary tree approach (where only binary partitions are considered). As we will see, the algorithm meets three important objectives: to identify subgroups of population by partitioning data; to compare the obtained groups; to identify the most significant differences. The last two goals are achieved by introducing a specific "split" criterion, which is strictly related to the nature of the analyzed variables.

3.2 Introduction to the Segmentation Tree

A segmentation tree has greatly increased in popularity during recent years. It is a statistical system that mines data to predict or classify future observations based on a set of decision rules. It is characterized by a special form called "tree structure" because its graph has the appearance of a tree (although the tree is generally displayed upside down). The first approaches were proposed by Sonquist and Morgan (1964) and Sonquist, Baker and Morgan (1971) with their methodology called Automatic Interaction Detection (AID). A different scheme was developed by Kass with the CHAID algorithm (Kass,1980). A new impulse was given to segmentation tasks by the work of Breiman (Breiman *et al.*, 1984) with the CART (Classification And Regression Tree) algorithm, together with the work of Ross Quinlan and his ID3 algorithm (Quinlan, 1986). Successors to the ID3 algorithm have been developed with the C4.5 algorithm (Quinlan, 1993) and the C5 algorithm (Quinlan, 1998). Decision trees have three types of nodes: root, internal, and leaf nodes. The root node is that which contains the entire set of elements in the sample; the root node has no incoming branches, and it may have zero or two (or more) outgoing branches. Internal nodes are those that have one incoming branch and two (or more) outgoing branches. Leaf or terminal nodes have no outgoing branches. The branches are the lines connecting the nodes. Every node that is preceded by a higher level node is called "child" node. A node that gives origin to two (or more) child nodes is called a "parent" node. Terminal nodes are nodes without children.

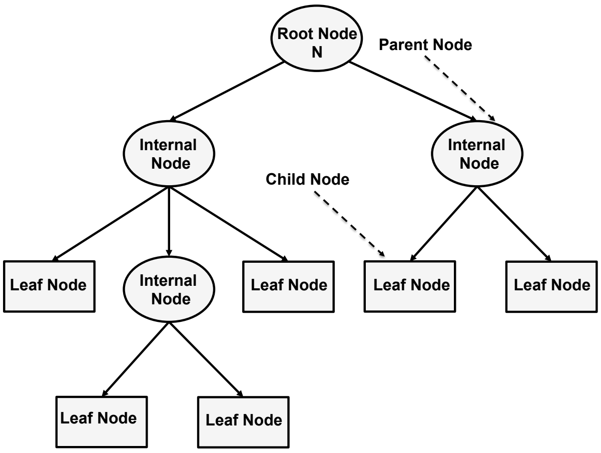


Figure 3.1: Graphical representation of a segmentation tree structure with its different types of nodes

According to Apté and Weiss (1997), the basic steps of the tree building procedure can be summarized as follows. The initial state of a decision tree is the root node to which all the examples from the training set are assigned. If it is the case that all examples belong to the same class, then no further decisions need to be made to partition the examples, and the solution is complete. If the examples at this node belong to two or more classes, then a test is made at the node, which will result in a split. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained.

3.2.1 Binary Segmentation Tree

When the nodes of the tree only can be partitioned into two, we have a binary segmentation tree. According to Breiman *et al.* (1984), there are some steps to follow when you are building a binary segmentation tree. The steps include: establish the set of admissible divisions for each node; adopt a criterion for the goodness of the split, that is, the way in which the nodes are divided from parent nodes to child nodes (*split criterion*); fix a rule to arrest the growth of the tree (*stopping rule*); and, once a node is recognized as a terminal, set a rule that allows determining how to classify cases that are contained in it. The process begins by establishing the set of admissible splits for each node, a procedure that is related to the nature of the segmentation variables. Following Sánchez (2009), the number and type of possible binary splits will vary depending on the scale (e.g., binary, nominal, ordinal, and continuous) of the segmentation variables. According to this classification, the number of the binary partitions can be summarized as:

Type of Variable	Binary splits
Binary	1
Nominal	$2^{k-1}-1$
Ordinal	$k-1$
Continuous	$k-1$

Table 3.1: Number of binary splits according to the type of variable

where K represents the number of variables' levels.

The process will continue until when all observations are divided into child nodes. The criterion by which the nodes are divided from parent nodes to child nodes will be fulfilled when a high degree of homogeneity in each terminal node is found. The stopping rule is optional, but it is useful for avoiding over-fitting problems. One possibility is to let the tree grow freely and, subsequently, to remove the less significant branches (post pruning-process). Another strategy is to identify a stop criterion, like setting the minimum number of observations that each node must contain, the number maximum of splits, or the minimum value of the splitting criterion (pre-pruning process). The final step is to assign each terminal node to a class. In this case you can follow a rule known as "plurality": the terminal node is classified according to the most represented group.

3.3 The PATHMOX Algorithm

As we have said in section (3.1), we want the algorithm to achieve three objectives: identify subgroups of population by partitioning data, compare the obtained groups by verifying the presence of significant differences and, if found, identify the most significant. To perform these three goals, the procedure adapts the principles of binary segmentation processes to produce a segmentation tree with different models in each of the obtained nodes. Unlike a classic decision tree, our algorithm has no prediction purpose (there are no predefined classes to be predicted); but it has an identification goal to detect different models present in the data. To this end, it identifies the set of splits (based on the segmentation variables) and does so with superior discriminating capacity – in the sense that it separates the models as much as possible. Here, a split criterion must be found to decide whether two models calibrated from two distinct segments (successors of a node) can be considered different. In the context of partial least squares path modeling, this algorithm was proposed for the first time by Gastón Sánchez in 2009 and given the name PATHMOX. It represented a new point of view in observing heterogeneity in PLS-PM models. However, it is important to note that PATHMOX can be generalized to many other kinds of methodologies and statistical fields according to the appropriate split criterion. Following the same approach, the SEM tree procedure (Brandmaier *et al.*, 2013) has been recently proposed as a different recursive partitioning technique in the framework of structural model equations. Differences between PATHMOX and SEM Trees primarily reflect differences in the application orientation: the former is applied in the context of partial least squares path modeling, the latter is utilized for the LISREL approach and, consequently, in the underlying estimation techniques. These are least squares for PATHMOX and maximum likelihood for SEM Trees. For a comparison of PLS vs LISREL, see Jöreskog and Wold (1982).

3.3.1 Algorithm Description

The procedure starts by establishing a set of admissible partitions for each node of the tree. This aspect is directly related to the nature of segmentation variables (see Table (3.1)). Once all possible subgroups have been considered, the second step is to identify the best partition (*split criterion*). The best split is obtained by applying a criterion (i.e test) that allows comparison of all sub-models calculated for each partition.

The criterion must be able to identify a degree of difference between the two compared sub-models. In general, the degree of difference is associated with a p-value: the lower the p-value, the more different the partitions are. The optimal partition is chosen by comparing the different p-values of each test and selecting the lowest. The choice of the split criterion is related to both the nature of the analyzed data and the kind of methodology utilized to approach the data. A detailed description of criteria is proposed in the next sections. This step is repeated iteratively, thus obtaining a ranking of the best segmentation variables and the different nodes of the tree. The last phase is to identify a criterion for discriminating whether the segment is an intermediate node or a terminal one (*stopping rule*). Here, a pre-pruning rule is adopted. One node is considered terminal when:

- The segmentation variables for that parent node have been exhausted and no more splits are feasible.
- The selected best split has an associated p-value exceeding the significance threshold.
- The number of elements in one node is less than or equal to some predefined minimum number of elements.

The algorithm has three parameters that have to be defined by the analyst:

1. The significance p-value threshold.
2. The minimum number of elements for a node that avoids fragmentation of small size nodes.
3. The grow tree depth level that allows avoiding an intractable number of segments.

We can summarize the procedure in the following Table:

Algorithm 1 PATHMOX Algorithm

```
Step 1. Start with the global model at the root node

Step 2. Establish a set of admissible partitions for each segmentation variable in each node of the tree

Step 3. Detect the best partition by:

        3.1. Comparing all binary partitions in all segmentation variables

        3.2. Applying the test, calculating for each comparison a p-value

        3.3. Sorting the p-values in a descending order

        3.4. Choosing as the best partition the one associated to the lowest p-value

Step 4. If (stop criteria[1] = false) than

        repeat by step 3

        1. Posible stop criteria:

        a. The number of individuals in the group falls below a fixed level

        b. The p-values test are not significant

        c. Maximum number of tree depth attained
```

3.4 The Split Criterion and Partial Least Squares Path Modeling PLS-PM

Gastón Sanchéz proposed the F-global test in 2009 as a criterion for comparing two different PLS path models by extending the test for comparing two linear regressions introduced by Lebart *et al.* (1979). This test focuses on the relationships between the path coefficients of the structural model, and it is based on the consideration that comparing two structural models can be put in terms of comparing two regression models. Structural models are in fact nothing more than a set of regressions among latent variables, one regression for each endogenous variable. This paragraph follows the proof described in his work (see Sanchéz 2009, pp.145:148).

Models Comparison in Pathmox

The F-global test comparison is based on the global model comparison at the structural level. Each binary split defines a pair of nodes, each of which will have its associated structural model, i.e. its associated set of path coefficients. Then, we perform a global comparison test on the identity of the two models, meaning that the sets of path coefficients in the two child nodes are equal to those of the parent node. The model of the parent node corresponds to a homogeneous situation, and the model of the child nodes corresponds to a heterogeneous situation. To that end, Sanchéz (2009) adapted the identity test of two regression models by Lebart *et al.* (1979) and Chow (1960), and he used it to detect the most significant split.

Testing the Equality of Two PLS Path Models

Let us consider a path model, as in Figure (3.2), with two endogenous latent variables,η_1 and η_2. Its generalization into more complex models is straightforward, with the inconvenience of complicating the notation:

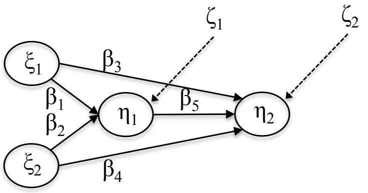


Figure 3.2: Path diagram of a PLS model with two endogenous variables

The structural equations for both endogenous constructs are:

$$\eta_1 = \beta_1\xi_1 + \beta_2\xi_2 + \zeta_1 \tag{3.1}$$

$$\eta_2 = \beta_3\xi_1 + \beta_4\xi_2 + \beta_5\eta_1 + \zeta_2 \tag{3.2}$$

The disturbance terms ζ_1 and ζ_2 are assumed to be normally distributed with zero mean and finite variance; that is, $E(\zeta_1) = E(\zeta_2) = 0$ and $Var(\zeta_1) = Var(\zeta_2) = \sigma^2\mathrm{I}$. It is also assumed that $Cov(\zeta_1, \zeta_2) = 0$.

We can define the following matrices:

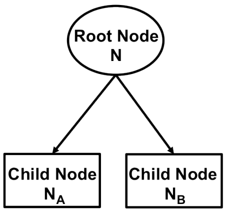
$$\begin{aligned}
X_1 &= [\xi_1, \xi_2] && \text{a column matrix with the explicative latent variables of } \eta_1 && (3.3)\\
B_1 &= [\beta_1, \beta_2] && \text{a vector of path coefficients for the regression of } \eta_1 && (3.4)\\
X_2 &= [\xi_1, \xi_2, \eta_1] && \text{a column matrix with the explicative latent variables of } \eta_2 && (3.5)\\
B_2 &= [\beta_3, \beta_4, \beta_5] && \text{a vector of path coefficients for the regression of } \eta_2 && (3.6)
\end{aligned}$$

The structural equations in matrix form are expressed as:

$$\eta_1 = X_1 B_1 + \zeta_1 \qquad (3.7)$$

$$\eta_2 = X_2 B_2 + \zeta_2 \qquad (3.8)$$

We assume that the parent node is divided in two child nodes or segments



In each segment we compute its own structural model:

$$Segment^A : \eta_1^A = X_1^A B_1^A + \zeta_1^A \quad \text{and} \quad \eta_2^A = X_2^A B_2^A + \zeta_2^A \qquad (3.9)$$

$$Segment^B : \eta_1^B = X_1^B B_1^B + \zeta_1^B \quad \text{and} \quad \eta_2^B = X_2^B B_2^B + \zeta_2^B \qquad (3.10)$$

with $\zeta_1^A\ N(0,\sigma^2 I)$ and $\zeta_2^A\ N(0,\sigma^2 I)$, $\zeta_1^B\ N(0,\sigma^2 I)$ and $\zeta_2^B\ N(0,\sigma^2 I)$.

Now, we want to investigate if the models of the two segments A and B are different or not. Thus, the test hypothesis will be: *The coefficients of segments A and B, are equal in both equations at the same time.*

We can define a null hypothesis and alternative hypothesis for each structural equation as:

$$H_{0\eta_1} : \underset{[n,1]}{\begin{bmatrix}\eta_1^A\\ \eta_1^B\end{bmatrix}} = \underset{[n,p_1]}{\begin{bmatrix}X_1^A\\ X_1^B\end{bmatrix}} \underset{[p_1,1]}{\begin{bmatrix}\beta_1\end{bmatrix}} + \underset{[n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^B\end{bmatrix}} \qquad (3.11)$$

$$H_{0\eta_2} : \underset{[n,1]}{\begin{bmatrix}\eta_2^A\\ \eta_2^B\end{bmatrix}} = \underset{[n,p_2]}{\begin{bmatrix}X_2^A\\ X_2^B\end{bmatrix}} \underset{[p_2,1]}{\begin{bmatrix}\beta_2\end{bmatrix}} + \underset{[n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^B\end{bmatrix}} \qquad (3.12)$$

$$H_{1\eta_1} : \underset{[n,1]}{\begin{bmatrix}\eta_1^A\\ \eta_1^B\end{bmatrix}} = \underset{[n,2p_1]}{\begin{bmatrix}X_1^A & 0\\ 0 & X_1^B\end{bmatrix}} \underset{[2p_1,1]}{\begin{bmatrix}\beta_1^A\\ \beta_1^B\end{bmatrix}} + \underset{[n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^B\end{bmatrix}} \qquad (3.13)$$

$$H_{1\eta_2} : \underset{[n,1]}{\begin{bmatrix}\eta_2^A\\ \eta_2^B\end{bmatrix}} = \underset{[n,2p_2]}{\begin{bmatrix}X_2^A & 0\\ 0 & X_2^B\end{bmatrix}} \underset{[2p_2,1]}{\begin{bmatrix}\beta_2^A\\ \beta_2^B\end{bmatrix}} + \underset{[n,1]}{\begin{bmatrix}\zeta_2^A\\ \zeta_2^B\end{bmatrix}} \qquad (3.14)$$

Note that in the null hypothesis we consider that the path coefficients are the same for both segments, whereas in the alternative hypothesis we fit different path coefficients for segments A and B. Each pair of hypotheses can be combined in more compact expressions:

$$H_0: \underset{[2n,1]}{\begin{bmatrix}\eta_1^A\\ \eta_2^A\\ \cdot\\ \eta_1^B\\ \eta_2^B\end{bmatrix}} = \underset{[2n,p_1+p_2]}{\begin{bmatrix}X_1^A & 0\\ 0 & X_2^A\\ \cdot\cdot & \cdot\cdot\\ X_1^B & 0\\ 0 & X_2^B\end{bmatrix}} \underset{[p_1+p_2,1]}{\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^A\\ \cdot\\ \zeta_1^B\\ \zeta_2^B\end{bmatrix}} \tag{3.15}$$

$$H_1: \underset{[2n,1]}{\begin{bmatrix}\eta_1^A\\ \eta_2^A\\ \cdot\\ \eta_1^B\\ \eta_2^B\end{bmatrix}} = \underset{[2n,2p_1+2p_2]}{\begin{bmatrix}X_1^A & 0 & 0 & 0\\ 0 & X_2^A & 0 & 0\\ \cdots & \cdots & \cdots & \\ 0 & 0 & X_1^B & 0\\ 0 & 0 & 0 & X_2^B\end{bmatrix}} \underset{[2p_1+2p_2,1]}{\begin{bmatrix}\beta_1^A\\ \beta_1^B\\ \cdot\\ \beta_2^A\\ \beta_2^B\end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^A\\ \cdot\\ \zeta_1^B\\ \zeta_2^B\end{bmatrix}} \tag{3.16}$$

The matrices X_0, X and A required in Lemma 2 of Lebart *et al.* (1979), such that $X = X_0A$. can be defined:

$$X_0 = \underset{[2n,p_1+p_2]}{\begin{bmatrix}X_1^A & 0\\ 0 & X_2^A\\ \cdot\cdot & \cdot\cdot\\ X_1^B & 0\\ 0 & X_2^B\end{bmatrix}} \quad X = \underset{[2n,2p_1+2p_2]}{\begin{bmatrix}X_1^A & 0 & 0 & 0\\ 0 & X_2^A & 0 & 0\\ \cdots & \cdots & \cdots & \\ 0 & 0 & X_1^B & 0\\ 0 & 0 & 0 & X_2^B\end{bmatrix}} \tag{3.17}$$

$$A = \underset{[2p_1+2p_2,p_1+p_2]}{\begin{bmatrix}I_{p_1} & 0\\ 0 & I_{p_2}\\ \cdot & \cdot\cdot\\ I_{p_1} & 0\\ 0 & I_{p_2}\end{bmatrix}} \tag{3.18}$$

where I_P is the identity matrix of order equal to the number of p variables.

Then, assuming that the random perturbations associated with the latent variables are uncorrelated with equal variance, we can apply Lemmas 1 and 2 of Lebart *et al.* (1979). Hence, the F-statistic measuring the discrepancy between the two models:

$$F_{Global} = \frac{(SS_{H_0} - SS_{H_1}) \Big/ (p_1 + p_2)}{SS_{H_1} \Big/ [2n - 2(p_1 + p_2)]} \tag{3.19}$$

where SS_{H_0} and SS_{H_1} stand for the corresponding sum of squares of residuals in both models. It then follows, under the null hypothesis, that we have an F distribution with $p_1 + p_2$ and $2n - 2(p_1 + p_2)$ degrees of freedom.

3.5 Extending the PATHMOX Approach for Detecting Which Constructs Differentiate Segments

As we can see, the PATHMOX approach allows us to detect the existence of different path models in a dataset without identifying segmentation variables beforehand: the different segments are reveled as branches of the segmentation tree. However, the F-test used in PATHMOX as split criterion is a global criterion: it allows assessing whether all the path coefficients for two compared structural models are equal or not, but it does not indicate which particular structural equation and which path coefficients are responsible for the difference. To identify the significantly distinct structural equation and the responsible path coefficients of the split, we introduced the F-block test and the F-coefficient test.

3.5.1 F-Block Test

To detect which structural regression (i.e. structural equation) is responsible for the global difference, we have extended the F-global test, to compare the equality of each structural equation of the model. We will call the statistic of this comparison F-block (or block-test) . Let us consider the same model shown in Figure (3.2) with two endogenous variables, η_1 and η_2:

$$\eta_1 = \beta_1\xi_1 + \beta_2\xi_2 + \zeta_1 \tag{3.20}$$

$$\eta_2 = \beta_3\xi_1 + \beta_4\xi_2 + \beta_5\eta_1 + \zeta_2 \tag{3.21}$$

Let us assume that the F-global test gives a significant p-value. We want to investigate which structural equation is responsible for the difference. The test hypothesis will be: *for every one of the structural equations, the path coefficients of segments A and B are equal, whereas the others can freely vary.*

For the sake of simplicity, we test whether the first structural equation is equal in both segments while letting equation two vary freely. In this case the null hypothesis, H_0, states that the structural equation shown in Figure (3.21) is equal for segments A and B, while the alternative hypothesis, H_1, states that all structural equations are different. The two hypotheses can be written as follows:

$$H_0: \underset{[2n,1]}{\begin{bmatrix}\eta_1^A\\ \eta_2^A\\ \eta_1^B\\ \eta_2^B\end{bmatrix}} = \underset{[2n,p_1+2p_2]}{\begin{bmatrix}X_1^A & 0 & 0\\ 0 & X_2^A & 0\\ X_1^B & 0 & 0\\ 0 & 0 & X_2^B\end{bmatrix}} \underset{[p_1+2p_2,1]}{\begin{bmatrix}\beta_1\\ \beta_2^A\\ \beta_2^B\end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^A\\ \zeta_1^B\\ \zeta_2^B\end{bmatrix}} \tag{3.22}$$

$$H_1: \underset{[2n,1]}{\begin{bmatrix}\eta_1^A\\ \eta_2^A\\ \eta_1^B\\ \eta_2^B\end{bmatrix}} = \underset{[2n,2p_1+2p_2]}{\begin{bmatrix}X_1^A & 0 & 0 & 0\\ 0 & X_2^A & 0 & 0\\ 0 & 0 & X_1^B & 0\\ 0 & 0 & 0 & X_2^B\end{bmatrix}} \underset{[2p_1+2p_2,1]}{\begin{bmatrix}\beta_1^A\\ \beta_1^B\\ \beta_2^A\\ \beta_2^B\end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^A\\ \zeta_1^B\\ \zeta_2^B\end{bmatrix}} \tag{3.23}$$

where $n = n_A + n_B$ is the number of elements in the model containing the two nodes; and p_j is the number of explicative latent variables for each j-th endogenous construct $j = 1, \dots, J$ (in this example $J = 2$).

We can define the matrices X_0. The X corresponding to both hypotheses are:

$$X_0 = \underset{[2n,p_1+2p_2]}{\begin{bmatrix}X_1^A & 0 & 0\\ 0 & X_2^A & 0\\ X_1^B & 0 & 0\\ 0 & 0 & X_2^B\end{bmatrix}} \quad X = \underset{[2n,2p_1+2p_2]}{\begin{bmatrix}X_1^A & 0 & 0 & 0\\ 0 & X_2^A & 0 & 0\\ 0 & 0 & X_1^B & 0\\ 0 & 0 & 0 & X_2^B\end{bmatrix}} \tag{3.24}$$

Then, we can see that $X_0 = XA$ defining the matrix A as:

$$A = \underset{[\,2p_1+2p_2, p_1+2p_2]}{\begin{bmatrix} I_{p_1} & 0 & 0 \\ 0 & I_{p_2} & 0 \\ I_{p_1} & 0 & 0 \\ 0 & 0 & I_{p_2} \end{bmatrix}} \tag{3.25}$$

where I_P is the identity matrix of order equal to the number of p variables.

Then, as before, we can apply Lemmas 1 and 2 of Lebart *et al.* (1979). Hence, the F-statistic measuring the discrepancy between the two models:

$$F_{Block} = \frac{(SS_{H_0} - SS_{H_1}) \Big/ p_1}{SS_{H_1} \Big/ 2(n - p_1 - p_2)} \tag{3.26}$$

where SS_{H_0} and SS_{H_1} stands for the corresponding sum of squares of residuals in both models, follows, under the null hypothesis, an F distribution with p_1 and $2(n - p_1 - p_2)$ degrees of freedom.

3.5.2 F-Coefficient Test

Let us suppose now that the difference between the first structural equation in segments one and two is significant, that is, there is a difference between segment A and segment B. We want to investigate which are the the responsible coefficients for such a difference. Let us consider the same structural model shown in Figure (3.2). For the sake of simplicity, we test the equality of coefficient β_1 in the first equation of both segments. We re-adapt the same global F-test to this situation. The test hypothesis will be: *for every path coefficient of segments A and B, we test if they are equal; whereas the others can vary freely.*
The two hypotheses are written as follows:

$$H_0 : \underset{[2n,1]}{\begin{bmatrix} \eta_1^A \\ \eta_2^A \\ \eta_1^B \\ \eta_2^B \end{bmatrix}} = \underset{[2n, 2\sum_{j=1}^{P} p_j - 1]}{\begin{bmatrix} \xi_1^A & \xi_2^A & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \xi_1^A & \xi_2^A & \eta_1^A & 0 & 0 & 0 & 0 \\ \xi_1^B & 0 & 0 & 0 & 0 & \xi_2^B & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \xi_1^B & \xi_2^B & \eta_1^B \end{bmatrix}} \underset{[2\sum_{j=1}^{P} p_j - 1, 1]}{\begin{bmatrix} \beta_1 \\ \beta_2^A \\ \beta_3^A \\ \vdots \\ \beta_5^B \end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix} \zeta_1^A \\ \zeta_2^A \\ \zeta_1^B \\ \zeta_2^B \end{bmatrix}} \tag{3.27}$$

$$H_1 : \underset{[2n,1]}{\begin{bmatrix} \eta_1^A \\ \eta_2^A \\ \eta_1^B \\ \eta_2^B \end{bmatrix}} = \underset{[2n, 2\sum_{j=1}^{P} p_j]}{\begin{bmatrix} \xi_1^A & \xi_2^A & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \xi_1^A & \xi_2^A & \eta_1^A & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \xi_1^B & \xi_2^B & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi_1^B & \xi_2^B & \eta_1^B \end{bmatrix}} \underset{[2\sum_{j=1}^{P} p_j, 1]}{\begin{bmatrix} \beta_1^A \\ \beta_2^A \\ \beta_3^A \\ \vdots \\ \beta_5^B \end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix} \zeta_1^A \\ \zeta_2^A \\ \zeta_1^B \\ \zeta_2^B \end{bmatrix}} \tag{3.28}$$

Calling X_0 the design matrix of the null hypothesis and X the design matrix of the alternative hypothesis, we have $X_0 = XA$, where:

$$A = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{[\,2\sum_{j=1}^{P} p_j, \sum_{j=1}^{P} p_j - 1]} \tag{3.29}$$

Then, as before, we can apply Lemmas 1 and 2 of Lebart *et al.* (1979). Hence, the F-statistic measuring the discrepancy between the two models:

$$F_{Coefficient} = \frac{(SS_{H_0} - SS_{H_1}) \Big/ 1}{SS_{H_1} \Big/ 2(n - \sum_{j=1}^{P} p_j)} \tag{3.30}$$

where SS_{H_0} and SS_{H_1} stand for the corresponding sums of squares of residuals in both models, which follows, under the null hypothesis, an F distribution with 1 and $2(n - \sum_{j=1}^{P} p_j)$ degrees of freedom.

3.6 Extending the PATHMOX Approach for Overcoming the Distribution Hypothesis

One criticism of the PATHMOX approach, when applied in the context of partial least squares path modeling, is that it utilizes a parametric test based on the hypothesis that the residuals have a normal distribution when comparing two structural models. PLS-PM is generally used to model survey analysis data. In general, these data are characterized by an asymmetric distribution: when an individual expresses an opinion, for example on a specific service, the opinion will likely be skewed positive. This situation produces skewness in the distribution of data. As we well know, it does not matter when we apply PLS, because one goal of this methodology is to obviate assumptions on data distribution. However, it represents a limit when we compare PLS models across PATHMOX, because we can not guarantee the normal distribution of our data, which is the condition needed to apply the test. In 2009, Sánchez conducted a simulation study with no normal data and showed that the test clearly detects differences when data also present an asymmetric distribution. In any case, there is a theoretic limit to applying the F-test. To overcome this limit, we can extend the test to compare two robust LAD regressions (Koenker and Bassett, 1982) in the context of partial least squares path modeling.

When the number of the observations is sufficiently large and by following the proofs we gained from the F-global, F-block and F-Coefficients tests, we can rewrite the F-statistics utilizing robust regression (LAD):

$$F_{Global_{LAD}} = \frac{(SAR_{H_0} - SAR_{H_1}) \Big/ (p_1 + p_2)}{SAR_{H_1} \Big/ [2n - 2(p_1 + p_2)]} \sim F_{p_1+p_2, 2n-2(p_1+p_2)} \tag{3.31}$$

$$F_{Block_{LAD}} = \frac{(SAR_{H_0} - SAR_{H_1}) \Big/ p_1}{SAR_{H_1} \Big/ 2(n - p_1 - p_2)} \sim F_{p_1, 2(n-p_1-p_2)} \tag{3.32}$$

$$F_{Coefficient_{LAD}} = \frac{(SAR_{H_0} - SAR_{H_1}) \Big/ 1}{SAR_{H_1} \Big/ 2(n - \sum_{j=1}^{P} p_j)} \sim F_{1, 2(n-\sum_{j=1}^{P} p_j)} \tag{3.33}$$

where:
- SAR_{H_0} and SAR_{H_1} are the absolute sum of residuals under the null and alternative hypothesis;
- p_1 and p_2 are the number of the variables for the two segments;
- n is the number of observation

3.7 Extending the PATHMOX Approach for Dealing with the Factor Invariance Problem

In the context of PLS, the algorithm works by fixing the topology of the structural model (i.e. the model of the causal relationships between the latent variables), and the goal is to detect segments having different path coefficients in the structural model. However, this does not place any restrictions on the measurement models (i.e. the models that link the observed variables with their own constructs). Thus, anytime that a significant difference is found and two child nodes are defined, the relationships among latent variables are the same in both "child" models, but the estimation of each latent variable is recalculated. This consideration introduces the problem of invariance: if the estimation of the latent variables are recalculated (i.e. they could be different) in each terminal node of the tree, how can we be completely certain that we are correctly comparing the distinct behaviors of two individuals who belong to two different terminal nodes? As an illustration, let us consider the customer satisfaction of a company's clients: when PATHMOX is applied to data and a significant difference is found between men and women, for example, how can we be sure that we are correctly comparing the satisfaction of a man and a woman if the satisfaction is recalculated in each terminal model? To solve this problem, an invariance test can be considered, where the goal of the test is to verify if the measurement models of each terminal node may be assumed equal or not.

3.8 Invariance Test

To test the invariance of the measurement model across different sub PLS-PM models we follow the same approach of model comparison, first stated by Lebart *et al.* (1979), Chow (1960). We propose to perform a global comparison of models test applied to the outer model. The current model is formed by the juxtaposition of all outer models of the identified segments, each one with their corresponding specific weights defining the constructs, whereas the null model is formed assuming the same weights for every construct in all segments. Then, we can test the invariance of weights across the sub-models. Non significance of the statistic reveals that we can assume a unique set of weights for every construct in all sub-models, that is, there is factor invariance, otherwise we will accept that not only structural models differ but also measurement models in the detected segments do.

3.8.1 Testing the Equality of the Sub-model Weights

Let's take a PLS-PM model with k latent variables. Without loss of generality, to formalize the test, we can take a very simple case; let be S the number of segments in this case $(S = 2)$ we will call them segment A and segment B; let be a model with two latent constructs $(k = 2)$, ξ and η each with its corresponding associated blocks X and Y, with p_1 and p_2 indicators respectively (in general and p_k denotes the number of indicators of block k). Let n be the total number of individuals $(n = n_A + n_B)$.

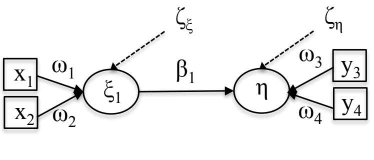


Figure 3.3: Path diagram of a PLS model with two latent constructs

We want to investigate if we can assume the existence of common weights for both sub-models. The null hypothesis means that there are common weights, for both constructs, ω_ξ, ω_η, whereas, the alternative hypothesis specifies that every segments has its own specific weights. Then, we can write the measurement model as a concatenation of all constructs in all segments and define the two hypotheses as follow:

$$H_0 : \begin{array}{cccccc} y & & X_0 & \beta_0 & & \varepsilon_0 \\ \begin{bmatrix}\xi^A\\ \eta^A\\ \xi^B\\ \eta^B\end{bmatrix} & = & \begin{bmatrix}X^A & 0\\ 0 & Y^A\\ X^B & 0\\ 0 & Y^B\end{bmatrix} & \begin{bmatrix}\omega_\xi\\ \omega_\eta\end{bmatrix} & + & \begin{bmatrix}\zeta^A_\xi\\ \zeta^A_\eta\\ \zeta^B_\xi\\ \zeta^B_\eta\end{bmatrix} \\ [\mathrm{Sn},1] & & [\mathrm{Sn},p_1+p_2] & [p_1+p_2,1] & & [\mathrm{Sn},1] \end{array} \tag{3.34}$$

$$H_1 : \begin{array}{cccccc} y & & X_1 & \beta_1 & & \varepsilon_1 \\ \begin{bmatrix}\xi^A\\ \eta^A\\ \xi^B\\ \eta^B\end{bmatrix} & = & \begin{bmatrix}X^A & 0 & 0 & 0\\ 0 & Y^A & 0 & 0\\ 0 & 0 & X^B & 0\\ 0 & 0 & 0 & Y^B\end{bmatrix} & \begin{bmatrix}\omega^A_\xi\\ \omega^A_\eta\\ \omega^B_\xi\\ \omega^B_\eta\end{bmatrix} & + & \begin{bmatrix}\zeta^A_\xi\\ \zeta^A_\eta\\ \zeta^B_\xi\\ \zeta^B_\eta\end{bmatrix} \\ [\mathrm{Sn},1] & & [\mathrm{Sn},Sp_1+Sp_2] & [Sp_1+Sp_2,1] & & [\mathrm{Sn},1] \end{array} \tag{3.35}$$

Then, assuming $y \sim N(X_\beta, \sigma^2)$, we know that Lebart *et al.* (1979):

1. The quadratic form $\varepsilon' Q \varepsilon / \sigma^2$, where Q is symmetric and idempotent matrix, follows a χ^2 distribution with v degrees of freedom (where v is the rank of Q).

2. Given two matrices X and X_0 where X_0 is defined as $X_0 = XA$ (for any matrix A), we can define the following matrices Q and Q_0 as:

$$Q = I_n - X(X'X)^{-1}X' \tag{3.36}$$

$$Q_0 = I_n - X_0(X_0'X_0)^{-1}X_0' \tag{3.37}$$

Then, it can be shown that $\varepsilon'(Q_0 - Q)\varepsilon$ follows an χ^2 distribution with $(v_0 - v)$ degrees of freedom, where:

- v is the rank of Q
- v_0 is the rank of Q_0

Thus, it is easy to see that the design matrices of precedent hypothesis can be written as $X_0 = XA$ taking:

$$A = \underset{[Sp_1+Sp_2, p_1+p_2]}{\begin{bmatrix} I_{p_1} & 0 \\ 0 & I_{p_2} \\ I_{p_1} & 0 \\ 0 & I_{p_2} \end{bmatrix}} \tag{3.38}$$

where I_{P_j} is the identity matrix of order p_j.

Assuming that every construct in every segment is normally distributed with equal variance; the difference of the residuals sum of squares of both hypotheses follows a χ^2 distribution with $S-1, p_1+p_2$ degrees of freedom; in general:

$$SS_{HO} - SS_{H1} \sim \chi^2_{(S-1)\sum_{k=1}^{S} p_k} \tag{3.39}$$

3.9 Generalization of the PATHMOX Algorithm

One of the most important aspects of the PATHMOX approach is that it can be generalized to other kinds of methodologies and statistical fields. This idea is based on the consideration that, even if it make sense to compare models by taking into account a set of segmentation variables, and even if an appropriate split criterion is utilized for the comparison, we can apply PATHMOX to explore the hypothesis of heterogeneity in our data. In this section, we take into account multiple regression and principal component analysis to show two generalization examples of the PATHMOX algorithm. We discuss different split criteria that allow us to compare models in each of these two methodologies.

3.10 Multiple Regression

In the context of multiple regression, we consider three different scenarios: the classic ordinary least squares regression (OLS), the least absolute regression (LAD) and the general linear regression (GLM). As in the case of PLS-PM, we perform a global comparison test on the identity of two models, meaning that the sets of coefficients in the two child nodes are equal to those of the parent node. The model of the parent node corresponds to the homogeneous situation, and the model of the child nodes corresponds to the heterogeneous situation.

3.10.1 Multiple Linear Regression

Comparing two linear multiple regressions can be performed using the hypothesis test introduced by Lebart *et al.* (1979, pp. 212), that is, an F-type based statistic. It represents the simple case (i.e. when we have just one equation) of the F-test being utilized to compare two structural models in PLS-PM.

Let us consider a generic linear model with two coefficients, β_1 and β_2:

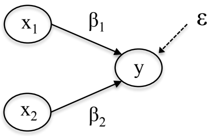


Figure 3.4: Path diagram of a regression model with two coefficients β_1 and β_2

For sake of simplicity we does not consider intercept. The model can be written as

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon \tag{3.40}$$

The disturbance term ε is assumed to be normally distributed with zero mean and finite variance, that is, $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 I$. This implies that $y \sim N(x_1\beta_1 + x_2\beta_2, \sigma^2 I)$

We can define the following matrices:

$$X = [x_1, x_2] \quad \text{a } (nxp) \text{ matrix with the explicative variables of } y \tag{3.41}$$

$$B = [\beta_1, \beta_2] \quad \text{a } (px1) \text{ vector of coefficients for the regression of } y \tag{3.42}$$

The model in a matrix form will be:

$$y = XB + \varepsilon \tag{3.43}$$

Let us suppose that the model in Figure (3.4) is associated with the parent node's model and contains the total number of n observations. Let us assume that the parent node is split into two child nodes or segments, one segment containing n_A elements and the other n_B. In each segment we compute its own model:

$$Segment^A : y^A = X^A B^A + \varepsilon^A \tag{3.44}$$

$$Segment^B : y^B = X^B B^B + \varepsilon^B \tag{3.45}$$

with ε^A $N(0, \sigma^2 I)$ and ε^B $N(0, \sigma^2 I)$.

Now we want to investigate if the models of the two segments A and B are different or not. The test hypothesis will be: *the coefficients of segments A and B are equal in both equations at the same time.*

We can define a null hypothesis and alternative hypothesis for each model as:

$$H_0 : \underset{[n,1]}{\begin{bmatrix} y^A \\ y^B \end{bmatrix}} = \underset{[n,p]}{\begin{bmatrix} X^A \\ X^B \end{bmatrix}} \underset{[p,1]}{\begin{bmatrix} \beta \end{bmatrix}} + \underset{[n,1]}{\begin{bmatrix} \varepsilon^A \\ \varepsilon^B \end{bmatrix}} \tag{3.46}$$

$$H_1 : \underset{[n,1]}{\begin{bmatrix} y^A \\ y^B \end{bmatrix}} = \underset{[n,2p]}{\begin{bmatrix} X_1^A & 0 \\ 0 & X_1^B \end{bmatrix}} \underset{[2p,1]}{\begin{bmatrix} \beta^A \\ \beta^B \end{bmatrix}} + \underset{[n,1]}{\begin{bmatrix} \varepsilon^A \\ \varepsilon^B \end{bmatrix}} \tag{3.47}$$

Note that in the null hypothesis we consider the vector of coefficients β to be the same for both segments (A and B), while in the alternative hypothesis we fit different coefficients (β^A and β^B) for the two models.

To obtain the F-statistic, the matrices X_0, X and A required by Lemma 2 of Lebart *et al.* (1979), such that $X = X_0A$, can be defined as:

$$X_0 = \underset{[n,p]}{\begin{bmatrix} X_1^A \\ X_1^B \end{bmatrix}} \quad X = \underset{[\mathrm{n},2p]}{\begin{bmatrix} X_1^A & 0 \\ 0 & X_1^B \end{bmatrix}} \tag{3.48}$$

$$A = \underset{[2p,p]}{\begin{bmatrix} I_p \\ I_p \end{bmatrix}} \tag{3.49}$$

where p is the number of explicative variables for y and I_p is the identity matrix of order p.

Thus, we can apply Lemmas 1 and 2 of Lebart *et al.* (1979) and test the H_0 hypothesis by computing the following F-statistic with (p) and $(n-2p)$ degrees of freedom.

$$F = \frac{(SS_{H_0} - SS_{H_1}) \Big/ p}{SS_{H_1} \Big/ (n-2p)} \tag{3.50}$$

3.10.1.1 The F-Coefficient Test

As in the PLS context of analyzing an OLS multiple regression and we find a significant difference between two models, it makes sense to explore which coefficients are responsible for the split. To this end, we can apply the test of equality on several coefficients, as introduced by Lebart *et al.* (1979, pp. 213). The test hypothesis will be: *for every coefficient of segments A and B, we test if it is equal, whereas the others can vary freely.*

For the sake of simplicity, we test the equality of coefficient β_1 in both segments. We can re-adapt the same F-test to this situation. The two hypotheses can be written as follows:

$$H_0: \underset{[\mathrm{n},1]}{\begin{bmatrix} y^A \\ y^B \end{bmatrix}} = \underset{[\mathrm{n},2p-1]}{\begin{bmatrix} x_1^A & x_2^A & 0 \\ x_1^B & 0 & x_2^B \end{bmatrix}} \underset{[2p-1,1]}{\begin{bmatrix} \beta_1 \\ \beta_2^A \\ \beta_2^B \end{bmatrix}} + \underset{[\mathrm{n},1]}{\begin{bmatrix} \varepsilon^A \\ \varepsilon^B \end{bmatrix}} \tag{3.51}$$

$$H_1: \underset{[\mathrm{n},1]}{\begin{bmatrix} y^A \\ y^B \end{bmatrix}} = \underset{[\mathrm{n},2p]}{\begin{bmatrix} x_1^A & x_2^A & 0 & 0 \\ 0 & 0 & x_1^B & x_2^B \end{bmatrix}} \underset{[2p,1]}{\begin{bmatrix} \beta_1^A \\ \beta_1^B \\ \beta_2^A \\ \beta_2^B \end{bmatrix}} + \underset{[\mathrm{n},1]}{\begin{bmatrix} \varepsilon^A \\ \varepsilon^B \end{bmatrix}} \tag{3.52}$$

Note that in the null hypothesis coefficient β_1 is the same for both segments, while in the alternative hypothesis we fit different coefficients for β_1^A and β_1^B. Calling X_0 the design matrix of the null hypothesis and X the the design matrix of the alternative hypothesis, we have $X_0 = XA$, where:

$$A = \underset{[\,2p,2p-1]}{\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \tag{3.53}$$

Then, as before, applying Lemmas 1 and 2 of Lebart *et al.* (1979), we can test the H_0 hypothesis by computing the following F-coefficient statistic with (1) and $(n-2p)$ degrees of freedom.

$$F = \frac{(SS_{H_0} - SS_{H_1}) \Big/ 1}{SS_{H_1} \Big/ 2(n-2p)} \tag{3.54}$$

3.10.2 Least Absolute Regression

As with OLS, LAD estimation can be applied in the context of PATHMOX in this case to obtain different robust regression models classified by a set of segmentation variables. Since the LAD estimate is nothing more than a robust linear regression, we can adapt the test for equality of two linear models, introduced by Lebart *et al.* (1979), to compare whether two LAD models are distinct or not. The problem, in this case, is that the Lebart approach is based on the hypothesis that the residuals of models are normal, and it is this supposition that allows us to consider Lemmas 1 and 2 of Lebart *et al.* (1979) and compare models by means of an F-statistic; whereas the LAD estimate does not guarantee the normal distribution of the residuals. Regardless, by following the asymptotic hypothesis that is, when the number of the observations is sufficiently large, an approximation is possible. Let us consider the same model with two coefficients, β_1 and β_2, shown in Figure (3.4). Lebart's test hypotheses are: (H_0) the coefficients of the two models are equal, and (H_1) the coefficients of the two models are distinct. Let us recall the absolute sum of residuals under the null hypothesis obtained by LAD estimation to be SAR_{H_0} and the the absolute sum of residuals under the alternative hypothesis to be SAR_{H_1}. Thus, if n is sufficiently large, we can re-adapt equations (3.50) and (3.54), and we can test the H_0 hypothesis by computing the following $\hat{F}_{LAD}$-statistics:

$$F_{Global_{LAD}} = \frac{(SAR_{H_0} - SAR_{H_1}) \Big/ p}{SAR_{H_1} \Big/ (n-2)p} \tag{3.55}$$

$$F_{Coefficient_{LAD}} = \frac{(SAR_{H_0} - SAR_{H_1}) \Big/ 1}{SAR_{H_1} \Big/ 2(n - \sum_{j=1}^{P} p_j)} \tag{3.56}$$

3.10.3 Generalized Linear Model

The comparison of two GLM regressions can be performed using the likelihood ratio test, which is an χ^2-type based statistic. Let us consider a general GLM model with two coefficients: β_1 and β_2 (see Figure (3.4)). Following a global comparison approach (hypothesis formulation: equation (3.46) and (3.47)), we can consider the *residual deviance* for the models under the null and alternative hypotheses: $Dev_{H_0} = l(H_0;y)$ and $Dev_{H_1} = l(H_1;y)$, and define the D_{LRT} likelihood ratio as:

$$D_{LRT} = -2\log\frac{Dev_{H_0}}{Dev_{H_1}} \tag{3.57}$$

Under the Null hypothesis, D_{LRT} is asymptotically distributed as χ^2 with $(N-k)$ degrees of freedom, where $N = n-p$ and $k = n-2p$.

Detecting the Responsible Coefficient

When we find a significant difference between two models, it also makes sense in GLM to explore which coefficients are responsible for the split. To this end, we can formulate the null and alternative hypotheses as in equations (3.51) and (3.52); again we can consider the *residual deviance* for both models: $Dev_{H_0} = l(H_0;y)$ and $Dev_{H_1} = l(H_1;y)$, and define the D_{LRT} likelihood ratio (see equation (3.57)). Under the null hypothesis, D_{LRT} is asymptotically distributed as χ^2 with $(N-k)$ degrees of freedom, where $N = n-p-1$ and $k = n-2p$.

3.11 Principal Component Analysis

The basic idea of principal component analysis (PCA) is to express a set of variables using a reduced number of variables called components. However, the possibility to express the variables across the components is based on the hypothesis that we can represent all data with a single pattern in the relation of variables. Thus, we could define the obtained components as an average representation of the variables of interest. However, it is reasonable in many cases to suppose the presence of some heterogeneity in the data: the observations behave differently according to a certain factor (segmentation variable). In this case, we cannot accept the existence of a single pattern in the relation of variables, and distinct distributions must be considered. In this context, it makes sense to apply PATHMOX to the PCA, with the objective of discovering different patterns of correlation between the variables of interest, patterns that define the underlying components. To this end, we propose an ad-hoc criterion based on the correlation of the interest variables; we call it *average correlation criterion*. The criterion is presented by considering a generic PCA model (see figure (3.5)) based on three principal components.

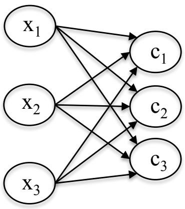


Figure 3.5: Path diagram of a PCA model with three components c_1 y c_2 and c_3

We can define the components as:

$$\begin{aligned} c_1 &= x_1u_{11} + x_2u_{12} + x_3u_{13} \\ c_2 &= x_1u_{21} + x_2u_{22} + x_3u_{23} \\ c_3 &= x_1u_{31} + x_2u_{32} + x_3u_{33} \end{aligned} \tag{3.58}$$

in a matrix form:

$$C = XU \tag{3.59}$$

3.11.1 Average Correlation Criterion

The average correlation criterion follows a comparison of coefficients approach. Thus, during the split process, we calculate two local models (one for segment A and another for the segment B) for each binary partition of all segmentation variables. We then test the equality of the two models by comparing the two representations of the variables in R^n. The degree of similarity is measured in terms of average correlation between the original variables, which is represented by the principal components: the lower the correlation is, the more different the two sub-models are. To guarantee that the degree of similarity does not depend simply on the direction of the components, a procrustean rotation is done by one representation on the other (see Figure (3.6).

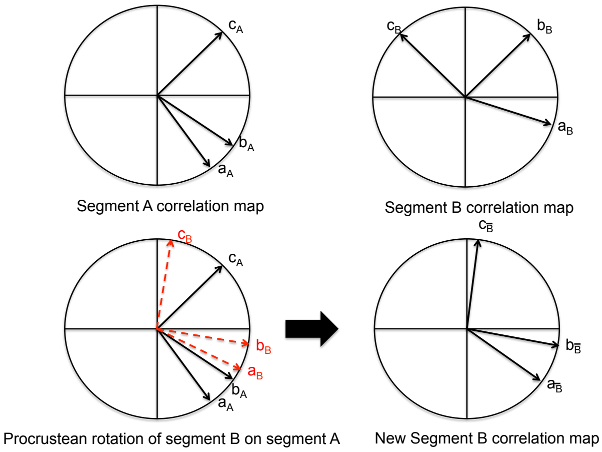

c_A
b_A
a_A
Segment A correlation map
c_B
b_B
a_B
Segment B correlation map
c_B
c_A
b_B
a_B
b_A
a_A
Procrustean rotation of segment B on segment A
$c_{\bar{B}}$
$b_{\bar{B}}$
$a_{\bar{B}}$
New Segment B correlation map

Figure 3.6: Graphical representation of the segment B procrustean rotation. Starting with the correlation maps at the top of the figure, we can observe how different the two representations seem (segments A and B). However, once we perform a procrustean rotation, the two representations present very similar patterns. We can conclude that, if we want to measure the different degrees of similarity between two representations, it is necessary to realize a procrustean rotation in order to avoid biased results.

Let us consider the model shown in Figure (3.5). For the sake of simplicity, we consider just the first two components. Let us assume that the model is associated with the parent node and contains the total number of n observations. Let us consider that the parent node is split into two child nodes or segments: one segment containing n_A elements and the other n_B. In each segment, we compute its own PCA model and we calculate the variable coordinates of the components:

$$\begin{aligned} Segment^A : C^A = X^A u^A \quad &\Rightarrow \quad \phi^A = u^A \sqrt{\lambda^A} \\ Segment^B : C^B = X^B u^B \quad &\Rightarrow \quad \phi^B = u^B \sqrt{\lambda^B} \end{aligned} \tag{3.60}$$

where:
- C^A and C^B are the principal components of segments A and B;
- ϕ^A and ϕ^B are the coordinates of the variables on the first two components;
- u^A, u^B and λ^A, λ^B are the eigenvector and the eigenvalue evaluated in the two sub-models;

Let us apply the procrustean rotation (Thurstone, 1947; Kaiser, 1958; Cattell, 1978) on the representation of segment B [1], ϕ^B obtaining $\hat{\phi}^B$ to guarantee that the degree of similitude does not depend simply on the direction of the components. We can calculate the *average correlation* as:

$$AC = \sum_{i=1}^{p} \frac{\phi_i^A \cdot \hat{\phi}_i^B}{p} \tag{3.61}$$

where:
- ϕ^A are the coordinates of the variables of the sub-model A on the first two components;

[1] To apply the procrustean rotation on the coordinates of the model of the segment B is just a reference; it can be applied indifferently on both segments

- $\hat{\phi}^B$ are the procrustean coordinates of the variables of the sub-model B on the first two components;
- p is the total number of the original variables;

As we can see, the average correlation is nothing more than the scalar product between the coordinates ϕ of the two sub-models A and B for the components, divided by the total number of variables (in this case three). This measure represents the degree of similarity between the two representations of the principal components of the original variables: the lower the correlation is, the more different the two sub-models are[2].

[2] We have tried by means a permutation test to find a reference distribution of the AC coefficient. However we haven't included it in the solution due to its computational cost. We have found an heuristic rules about the magnitude of the average correlation, which allows us to solve this problem.

Chapter 4

Simulations

In the following chapter we present a Monte Carlo simulation analysis in order to evaluate the sensitivity of the split criteria used in PATHMOX. Our simulations are based mainly on the works of Chin and Newsted (1999), Cassel *et al.* (1999, 2000), Westlund *et al.* (2001), Chin *et al.* (2003), Goodhue *et al.* (2006), and Sánchez (2009). We begin this chapter with the description of the principal objectives of the study (section (4.1)), we continue by reporting on the data generation process (section (4.2)) and the experimental factors (section (4.3)), and we conclude with the analysis of the obtained results (section (4.4), (4.5) and (4.6)).

4.1 Motivation

We have realized our simulation study with three principal objectives:

- assess the coherence of the F-block and the F-coefficient statistics in respect to the F-global;
- compare the least squares statistics with the LAD approach;
- evaluate the performance of the invariance measurement test.

The main objective behind the simulation study is to evaluate the coherence of the F-block and the F-coefficient statistics in respect to the F-global when two path models are compared. We are interested in verifying if the same difference is identified by the other two statistics when a significant difference is found by the F-global. This represents a necessary step in incorporating the F-block and the F-coefficient into PATHMOX's split process. The second objective of the study consists of comparing the three F statistics with those obtained by the LAD approximation; in this case we are interested in evaluating the degree of sensitivity in order to verify if the two approaches identify the same differences. The principal aim is establish under which conditions the LAD approximation is as suitable as the classical F statistics are. The goal of the third analysis is to investigate the performance of the invariance measurement test. Here, we want to determine if the test is specific for each terminal node of the tree, and if not, what are the different conditions under which this test is suitable for understanding if the measurement model can be considered the same for all of them. Before proceeding with the simulation studies, we check the adequacy of the simulation procedure and the corresponding PLS-PM estimation.

4.2 Simulated Models

The first step in our simulation study is the definition of the PLS model. We have considered two different models: we will call them Model 1 and Model 2; the first one is all reflective, whereas the second one contains a formative block.

Model 1

In Model 1, the data were generated according to the structural model in Figure (4.1) by following a two-step procedure (Reinartz *et al.*, 2005): first, we generated the latent variable data by following the relationship specified in the structural model;

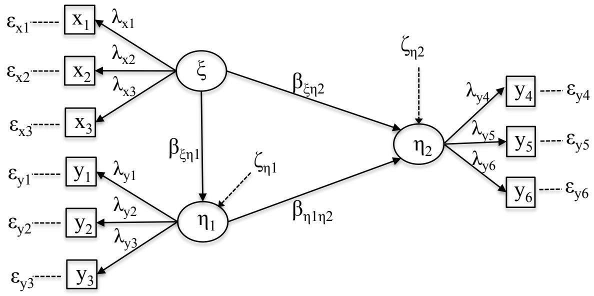


Figure 4.1: Path diagram of PLS model 1

The PLS model consists of one exogenous (ξ) and two endogenous (η_1 and η_2) latent variables. The inner structure is defined as:

$$\begin{aligned} \eta_1 &= \beta_{\xi\eta_1}\xi + \zeta_{\eta_1} \\ \eta_2 &= \beta_{\xi\eta_2}\xi + \beta_{\eta_1\eta_2}\eta_1 + \zeta_{\eta_2} \end{aligned} \tag{4.1}$$

where β are regression coefficients and ζ are the error terms associated with the endogenous latent variables. The manifest variables are denoted by x for ξ, and by y for η. The measurement models for ξ, η_1 and η_2 are reflective and defined as:

$$\begin{array}{lll} x_1 = \lambda_{x_1}\xi_1 + \varepsilon_{x_1} & y_1 = \lambda_{y_1}\xi_2 + \varepsilon_{y_1} & y_4 = \lambda_{y_4}\eta + \varepsilon_{y_4} \\ x_2 = \lambda_{x_2}\xi_1 + \varepsilon_{x_2} & y_2 = \lambda_{y_2}\xi_2 + \varepsilon_{y_2} & y_5 = \lambda_{y_5}\eta + \varepsilon_{y_5} \\ x_3 = \lambda_{x_3}\xi_1 + \varepsilon_{x_3} & y_2 = \lambda_{y_3}\xi_2 + \varepsilon_{y_3} & y_6 = \lambda_{y_6}\eta + \varepsilon_{y_6} \end{array} \tag{4.2}$$

The λ terms are coefficients, and the ε terms are random errors. We set the λ values for the three constructs equal to 0.85, 0.80 and 0.75, respectively, for $i = 1, \cdots, 3$. In the based model, the noises ε and ζ were realized from a Normal distribution, $N \sim (\mu = 0, \sigma^2 = 0.05)$. We have used a Beta distribution, $B \sim (6,3)$, for the exogenous latent variable, ξ, to reproduce the asymmetry that characterizes the way respondents answered the survey studies.

Model 2

Model 2 (see Figure (4.2)) differs from Model 1 in the way that we have related the manifest variables with the exogenous latent variable; in this case the measurement model of ξ is in fact formative and defined as:

$$\xi = \pi_{\xi 1}x_1 + \pi_{\xi 2}x_2 + \pi_{\xi 3}x_3 + \delta_\xi \tag{4.3}$$

The π terms are coefficients, and the δ_ξ term is the random error, introduced in the model due to the hypothesis that some manifest variables explaining the construct are not considered. The x values have been generated from a multivariate normal distribution with a mean vector, $\mu = (0,0,0)$, and a covariance matrix:

$$\Sigma = \begin{pmatrix} 1 & 0.01 & 0.1 \\ 0.01 & 1 & 0.01 \\ 0.1 & 0.01 & 1 \end{pmatrix} \tag{4.4}$$

For the base model, all the π terms were set to be $0.2, 0.3, 0.4$. The noises δ were realization from the Normal distribution $N \sim (\mu = 0, \sigma^2 = 0.1)$. We have used the same process adopted for the model A, to generate the endogenous latent variables and the corresponding manifest variables.

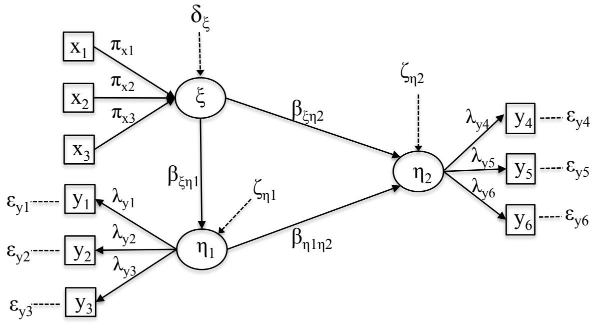


Figure 4.2: Path diagram of the PLS model 2

4.3 Data Generation Process

In generating the segments, an important condition is that the difference introduced in the two simulated segments must be controlled; that is, we have to ensure that, if we want a 0.5 coefficient of the model, we will obtain it as an average when simulating the data. To do this, let us consider the first inner equation of the model in Figure (4.1):

$$\eta_1 = \beta_{\xi\eta_1} + \zeta_\xi \tag{4.5}$$

The variance and the standard deviation of η_1 can be easily expressed as:

$$var(\eta_1) = \beta^2_{\xi\eta_1} + var(\zeta_\xi) \tag{4.6}$$

As we know, the latent variables in a PLS-PM algorithm are rescaled at each iteration. That corresponds in our model to fitting η under the restriction that $var(\eta) = 1$. This implies:

$$\begin{aligned} var(\eta_1) &= \beta^2_{\xi\eta_1} + var(\zeta_{\eta_1}) = 1 \\ var(\zeta_{\eta_1}) &= 1 - \beta^2_{\xi\eta_1} \end{aligned} \tag{4.7}$$

Applying the same logic to the second inner equation of the model in Figure (4.1), we obtain:

$$\begin{aligned} var(\eta_2) &= \beta^2_{\xi\eta_2} + \beta^2_{\eta_1\eta_2} + 2\beta_{\xi\eta_2}\beta_{\eta_1\eta_2} + var(\zeta_\xi) = 1 \\ var(\zeta_{\eta_2}) &= 1 - \beta^2_{\xi\eta_2} - \beta^2_{\eta_1\eta_2} - 2\beta_{\xi\eta_2}\beta_{\eta_1\eta_2} \end{aligned} \tag{4.8}$$

Thus, to guarantee that we can control the value of the β coefficients, we have to estimate the standard deviation of the endogenous error as:

$$\begin{aligned} SD(\zeta_{\eta_1}) &= \sqrt{1 - \beta^2_{\xi\eta_1}} \\ SD(\zeta_{\eta_2}) &= \sqrt{1 - \beta^2_{\xi\eta_2} - \beta^2_{\eta_1\eta_2} - 2\beta_{\xi\eta_2}\beta_{\eta_1\eta_2}} \end{aligned} \tag{4.9}$$

4.4 Adequacy of the PLS-PM Estimation

To show that these restrictions allow control of the path coefficients, we have run a simulation on Models 1 and 2. We have taken into account three different factors:

1. Sample size: {*100,400,1000*}
2. The error's standard deviation of the manifest variables: {*0.05, 0.2, 1*}
3. The difference in the path coefficients for both models: *high*, *medium*, and *small*

Checking the simulated inner model

First, we want to check the goodness of the simulated data, so we present in each plot the theoretical path coefficient with the corresponding density function of the least squares path coefficient, which was obtained from the data results in 500 simulations. Moreover, for each simulated data-set, we proceed to estimate the corresponding PLS-PM model. We also graphically represent the density function of the PLS-PM estimated path coefficient. This allows us to see how accurate the PLS-PM estimation is, in respect to the actual obtained by least squares.

The results of the simulation are presented in Figures (4.3) and (4.4). Here, each plot of the figures represents the density function of the coefficients obtained by least squares estimation (i.e., the estimated value from the simulation) and PLS-PM estimation (i.e., the estimated value from the PLS-PM algorithm). For the sake of simplicity, we present just the results obtained from a sample size of 400 and a *small* standard deviation of the measurement error ($\sigma_{MV} = 0.05$). We summarize the results under the other error's standard deviation levels and sample sizes in Tables (4.1) and (4.3).

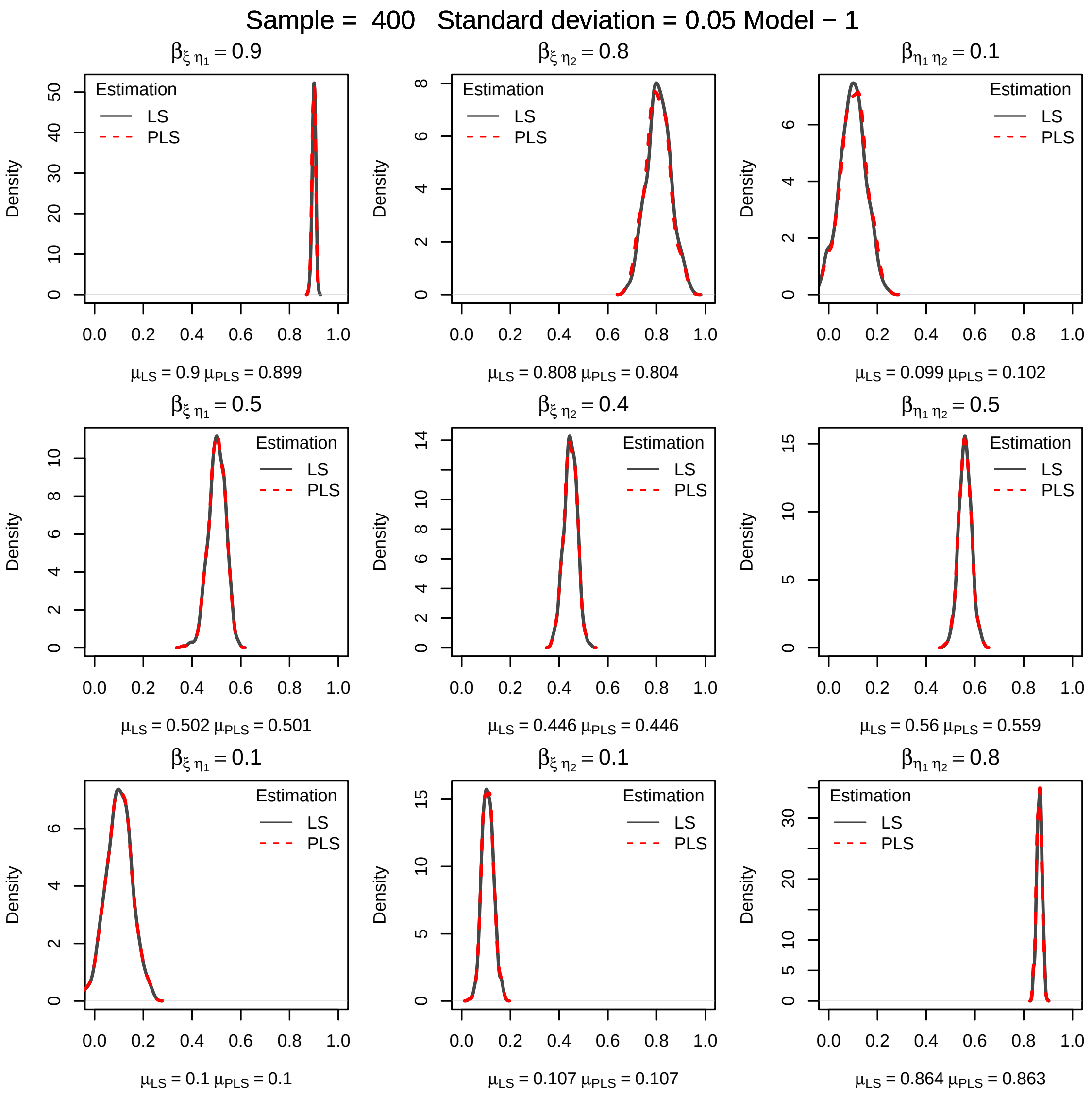


Figure 4.3: Density function of the coefficients estimated by least squares and PLS-PM algorithm - Model1

We can note that the least squares and the PLS-PM estimation are very similar: the curves overlap. In particular, the mean of the PLS-PM coefficient estimation is very similar to the theoretical value, including when we vary the coefficient values from high to small. Similar results were obtained for Model 2, with minor discrepancies for the PLS estimation.

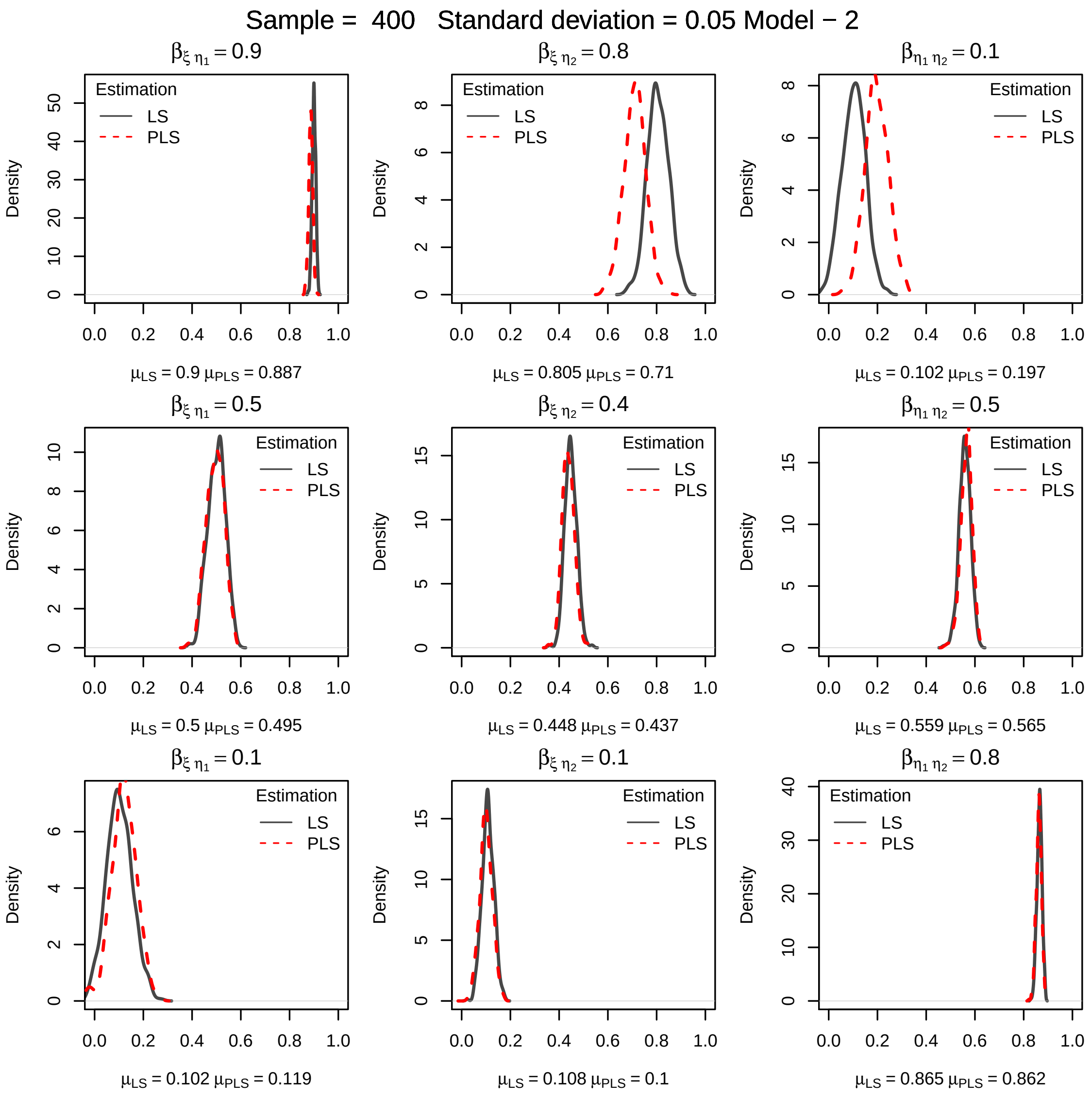


Figure 4.4: Density function of the coefficients estimated by least squares and PLS-PM algorithm - Model 2

In Tables (4.1) and (4.3), we present, in order, the theoretical value, the mean of the least squares estimation and the mean of PLS estimation, when we vary the error's standard deviation levels (from 0.05 to 1) and sample size (from 100 to 1000).
We can observe that, when we change the magnitude of the error's standard deviation, we note an increase in the discrepancies between the least squares and PLS estimations. In almost all cases, the difference is represented by an underestimation of the path coefficients fitted by the PLS algorithm. The coefficients estimated in Model 2 reproduce similar trends.

$\sigma_{\mathbf{MV}}$	$\beta_{\xi\eta_1}$	$\mu_{\mathbf{LS}}$	$\mu_{\mathbf{PLS}}$	$\beta_{\xi\eta_2}$	$\mu_{\mathbf{LS}}$	$\mu_{\mathbf{PLS}}$	$\beta_{\eta_1\eta_2}$	$\mu_{\mathbf{LS}}$	$\mu_{\mathbf{PLS}}$
Sample: 100									
	0.9000	0.9011	0.8999	0.8000	0.8092	0.8035	0.1000	0.0972	0.1023
$\sigma_{MV} = 0.05$	0.5000	0.4967	0.4959	0.4000	0.4519	0.4517	0.5000	0.5557	0.5553
	0.1000	0.1034	0.1038	0.1000	0.1130	0.1132	0.8000	0.8654	0.8641
	0.9000	0.9003	0.7982	0.8000	0.8092	0.5792	0.1000	0.0989	0.2695
$\sigma_{MV} = 0.2$	0.5000	0.5005	0.4500	0.4000	0.4488	0.4227	0.5000	0.5577	0.5051
	0.1000	0.1020	0.1023	0.1000	0.1083	0.1105	0.8000	0.8653	0.7669
	0.9000	0.9003	0.6007	0.8000	0.8065	0.4205	0.1000	0.1027	0.3009
$\sigma_{MV} = 1$	0.5000	0.4926	0.3445	0.4000	0.4454	0.3472	0.5000	0.5624	0.4076
	0.1000	0.0967	0.0867	0.1000	0.1065	0.1027	0.8000	0.8657	0.5790
Sample: 400									
	0.9000	0.9001	0.8990	0.8000	0.8083	0.8039	0.1000	0.0986	0.1024
$\sigma_{MV} = 0.05$	0.5000	0.5019	0.5015	0.4000	0.4463	0.4460	0.5000	0.5597	0.5590
	0.1000	0.0996	0.0995	0.1000	0.1073	0.1073	0.8000	0.8645	0.8634
	0.9000	0.9001	0.7979	0.8000	0.8071	0.5769	0.1000	0.1005	0.2732
$\sigma_{MV} = 0.2$	0.5000	0.4999	0.4427	0.4000	0.4469	0.4185	0.5000	0.5594	0.5087
	0.1000	0.1004	0.0914	0.1000	0.1084	0.1066	0.8000	0.8641	0.7637
	0.9000	0.9000	0.5942	0.8000	0.8097	0.4137	0.1000	0.0976	0.2999
$\sigma_{MV} = 1$	0.5000	0.4997	0.3337	0.4000	0.4440	0.3445	0.5000	0.5619	0.4051
	0.1000	0.1035	0.0838	0.1000	0.1072	0.0950	0.8000	0.8649	0.5720
Sample: 1000									
	0.9000	0.8998	0.8986	0.8000	0.8078	0.8027	0.1000	0.0995	0.1041
$\sigma_{MV} = 0.05$	0.5000	0.4992	0.4985	0.4000	0.4483	0.4481	0.5000	0.5584	0.5577
	0.1000	0.1004	0.1001	0.1000	0.1080	0.1081	0.8000	0.8640	0.8628
	0.9000	0.9000	0.7959	0.8000	0.8050	0.5747	0.1000	0.1024	0.2749
$\sigma_{MV} = 0.2$	0.5000	0.5002	0.4423	0.4000	0.4478	0.4207	0.5000	0.5585	0.5052
	0.1000	0.0993	0.0898	0.1000	0.1075	0.1044	0.8000	0.8648	0.7650
	0.9000	0.8999	0.5914	0.8000	0.8071	0.4119	0.1000	0.1002	0.3019
$\sigma_{MV} = 1$	0.5000	0.4996	0.3305	0.4000	0.4470	0.3483	0.5000	0.5591	0.3997
	0.1000	0.0989	0.0720	0.1000	0.1084	0.0918	0.8000	0.8650	0.5715

Table 4.1: Theoretical value, mean of least squares and PLS estimation by the distinct error's standard deviation levels of the manifest variables and distinct sample size - Model 1

$\sigma_{\mathbf{MV}}$	$\beta_{\xi\eta_1}$	$\mu_{\mathbf{LS}}$	$\mu_{\mathbf{PLS}}$	$\beta_{\xi\eta_2}$	$\mu_{\mathbf{LS}}$	$\mu_{\mathbf{PLS}}$	$\beta_{\eta_1\eta_2}$	$\mu_{\mathbf{LS}}$	$\mu_{\mathbf{PLS}}$
Sample: 100									
	0.9000	0.9007	0.8879	0.8000	0.8014	0.7138	0.1000	0.1064	0.1927
$\sigma_{MV} = 0.05$	0.5000	0.5063	0.5104	0.4000	0.4464	0.4367	0.5000	0.5582	0.5601
	0.1000	0.0963	0.1326	0.1000	0.1069	0.0864	0.8000	0.8664	0.8617
	0.9000	0.8999	0.8771	0.8000	0.8088	0.7265	0.1000	0.0977	0.1678
$\sigma_{MV} = 0.2$	0.5000	0.4963	0.4977	0.4000	0.4465	0.4409	0.5000	0.5619	0.5427
	0.1000	0.0993	0.1316	0.1000	0.1106	0.0961	0.8000	0.8659	0.8387
	0.9000	0.9005	0.6926	0.8000	0.8086	0.6583	0.1000	0.0989	0.0530
$\sigma_{MV} = 1$	0.5000	0.4998	0.4075	0.4000	0.4485	0.4459	0.5000	0.5590	0.3023
	0.1000	0.0946	0.1046	0.1000	0.1088	0.1167	0.8000	0.8655	0.5185
Sample: 400									
	0.9000	0.9001	0.8866	0.8000	0.8055	0.7099	0.1000	0.1025	0.1967
$\sigma_{MV} = 0.05$	0.5000	0.5004	0.4950	0.4000	0.4477	0.4374	0.5000	0.5585	0.5647
	0.1000	0.1021	0.1192	0.1000	0.1080	0.0997	0.8000	0.8647	0.8620
	0.9000	0.9002	0.8760	0.8000	0.8063	0.7211	0.1000	0.1008	0.1726
$\sigma_{MV} = 0.2$	0.5000	0.5021	0.4920	0.4000	0.4456	0.4403	0.5000	0.5603	0.5461
	0.1000	0.0973	0.1141	0.1000	0.1099	0.1026	0.8000	0.8649	0.8397
	0.9000	0.8995	0.6867	0.8000	0.8050	0.6536	0.1000	0.1026	0.0465
$\sigma_{MV} = 1$	0.5000	0.5010	0.3901	0.4000	0.4469	0.4424	0.5000	0.5598	0.3021
	0.1000	0.0995	0.0952	0.1000	0.1085	0.1110	0.8000	0.8655	0.5168
Sample: 1000									
	0.9000	0.9006	0.8872	0.8000	0.8041	0.7088	0.1000	0.1033	0.1971
$\sigma_{MV} = 0.05$	0.5000	0.5001	0.4939	0.4000	0.4470	0.4365	0.5000	0.5596	0.5662
	0.1000	0.0991	0.1058	0.1000	0.1073	0.1025	0.8000	0.8647	0.8629
	0.9000	0.9005	0.8757	0.8000	0.8072	0.7203	0.1000	0.1001	0.1737
$\sigma_{MV} = 0.2$	0.5000	0.5002	0.4882	0.4000	0.4468	0.4402	0.5000	0.5590	0.5462
	0.1000	0.1003	0.1047	0.1000	0.1089	0.1055	0.8000	0.8652	0.8410
	0.9000	0.9002	0.6868	0.8000	0.8078	0.6508	0.1000	0.0995	0.0479
$\sigma_{MV} = 1$	0.5000	0.5005	0.3831	0.4000	0.4488	0.4418	0.5000	0.5576	0.2972
	0.1000	0.1028	0.0869	0.1000	0.1072	0.1111	0.8000	0.8651	0.5152

Table 4.2: Theoretical value, mean of least squares and PLS estimation by the distinct error's standard deviation levels of the manifest variables - Model 2

Observing the results, we can draw two principal conclusions:

1. The simulated data approximate the theoretical data well when the causal-relationship model is "easy"

2. The PLS path estimation approximates the real value (simulated value) if the model presents a small standard deviation error of the manifest variables and strong relationship between latent variables.

4.5 Software and Computational Aspects

For the computational aspects, we employed the statistical software **R**, version 2.15.2. We used its pseudo-random generator function 'rbeta' for the latent variables and the 'mvnorm' function for the manifest variables of the formative block of Model 2. For calculating PLS path models, we utilized the function 'plspm'; whereas for the PATHMOX approach, we programmed the required algorithms .

4.6 Performed Simulations

The simulations present three distinct objectives. In the first one, we evaluate the coherence of the F-block and the F-coefficient statistics in respect to the F-global, when two path models are compared. We are interested in verifying if a significant difference found by the F-global is the same difference that is identified by the other two statistics. In the second, our aim is to compare the three F statistics with those obtained by the LAD approximation; in this case we want to check if the two approaches identify the same differences. In the third one, we investigate the performance of the invariance measurement test. Here, we want to determine the different conditions under which this test is suitable for understanding if the measurement model can be considered the same for all terminal nodes of the tree, or if it is specific for each one of them.

4.6.1 Experimental Factors

We have evaluated the performance of the tests under different experimental conditions. The factors of the experimental design are the following: sample size, difference between coefficients, the distinct levels of standard deviation of measurement error . In the invariance measurement test simulation study, we also take into account the difference between measurement model coefficients, because in this case we are interested in verifying if the test is suitable for identifying the difference in the measurement model of the terminal nodes obtained by PATHMOX. The impact of this factor is not considered in the other simulation studies, because PATHMOX does not make any restrictions on the measurement model during the split process: PATHMOX just takes into account the inner model, whereas the outer model can vary freely.

4.6.1.1 Simulation Factors Description

I. SIZE. We consider three sample sizes as the total number of cases: {100, 400, and 1000}.

II. DIFFERENCE BETWEEN COEFFICIENTS. The model has been estimated in two segments, A and B, by varying the level of the difference between path coefficients (see Figure (4.5)). In other words, they can be *EQUAL* in both segments, or the difference can be *SMALL*, *MEDIUM* and *LARGE*. Due to the restriction on the standard deviation of the error of the endogenous latent variables η_1 and η_2, the path coefficient values have been fixed by taking into account the conditions of admissibility, which are:

$$\begin{aligned} SD(\zeta_{\eta_1}) &\Rightarrow 1-\beta_{\xi\eta_1} > 0 \\ SD(\zeta_{\eta_2}) &\Rightarrow \beta_{\xi\eta_2} < 1-\beta_{\eta_1\eta_2} \end{aligned} \tag{4.10}$$

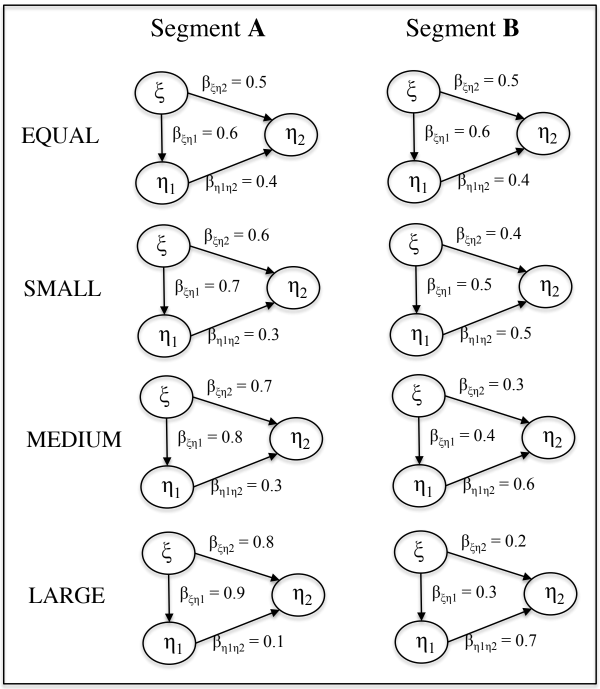

Segment **A**
Segment **B**
EQUAL
$\beta_{\xi\eta 2} = 0.5$
$\beta_{\xi\eta 1} = 0.6$
$\beta_{\eta 1\eta 2} = 0.4$
$\beta_{\xi\eta 2} = 0.5$
$\beta_{\xi\eta 1} = 0.6$
$\beta_{\eta 1\eta 2} = 0.4$
SMALL
$\beta_{\xi\eta 2} = 0.6$
$\beta_{\xi\eta 1} = 0.7$
$\beta_{\eta 1\eta 2} = 0.3$
$\beta_{\xi\eta 2} = 0.4$
$\beta_{\xi\eta 1} = 0.5$
$\beta_{\eta 1\eta 2} = 0.5$
MEDIUM
$\beta_{\xi\eta 2} = 0.7$
$\beta_{\xi\eta 1} = 0.8$
$\beta_{\eta 1\eta 2} = 0.3$
$\beta_{\xi\eta 2} = 0.3$
$\beta_{\xi\eta 1} = 0.4$
$\beta_{\eta 1\eta 2} = 0.6$
LARGE
$\beta_{\xi\eta 2} = 0.8$
$\beta_{\xi\eta 1} = 0.9$
$\beta_{\eta 1\eta 2} = 0.1$
$\beta_{\xi\eta 2} = 0.2$
$\beta_{\xi\eta 1} = 0.3$
$\beta_{\eta 1\eta 2} = 0.7$
ξ
η_1
η_2

Figure 4.5: Comparison by path coefficients between segments *A* and *B*

III. STANDARD DEVIATION OF MEASUREMENT ERRORS. We assume that the error terms ε follow a Normal distribution with zero expectation and three levels of standard deviation: small noise $\sigma = 0.05$, moderate noise $\sigma = 0.2$ and high noise $\sigma = 1$.

IV. DIFFERENCE BETWEEN COEFFICIENTS OF MEASUREMENT MODEL. As in the case of the difference between the path coefficients, we have considered different levels of difference between the measurement model coefficients (see Figure (4.6)). In other words, they can be *EQUAL* in both segments, or the difference can be *SMALL MEDIUM* and *LARGE*.

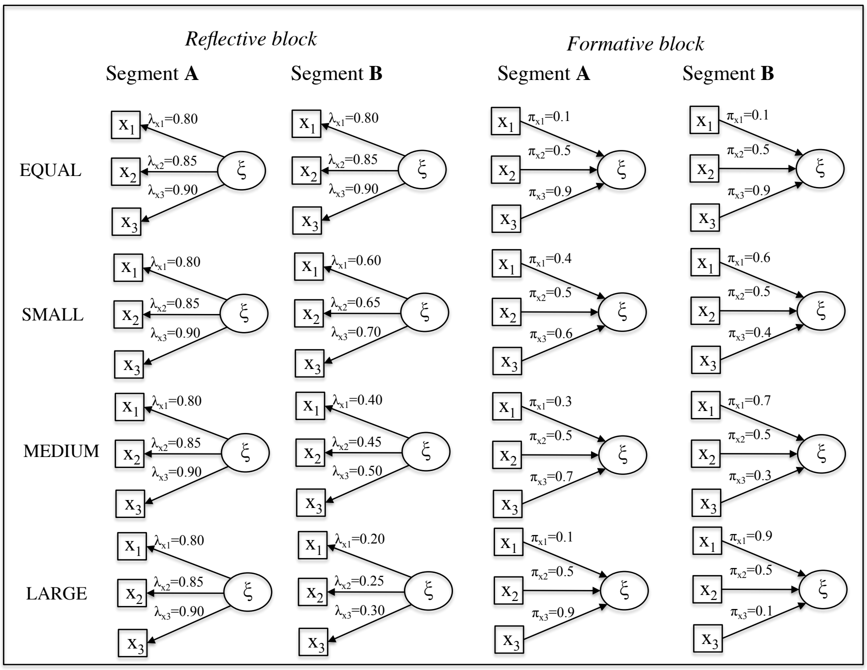


Figure 4.6: Comparison by measurement model coefficients of a *reflective* and *formative* block between segments *A* and *B*

4.6.2 Simulation Study to Asses the Coherence of the F-block and the F-coefficient Statistics Respect the F-global

In this simulation study, we want to assess the performance of the F-block and the F-coefficient statistics in respect to: the F-global regarding the different sample sizes (100, 400, and 1000); the different levels in error's standard deviation terms of the manifest variables; and the various levels of difference between the path coefficients. Our aim is to compare our simulation with that realized by Sánchez in 2009; in this way we can verify if the same difference is identified by the other two statistics when a significant difference is found by the F-global. We perform this study only on Model 1, since the PATHMOX approach only compares the path coefficients of the inner model without taking into account the definition of the blocks (reflective or formative).

In total, we have 3 × 3 × 4 = 36 scenarios, which are the number of possible combinations of sample sizes, noise levels and differences between coefficients. We run 50 repetitions for each experimental condition and present the mean of each 50 repetitions as an aggregate result.

Observing the results, we can note that the three tests present very similar behaviors. We can see that:

1. There is a clear effect of sample size: the larger the sample size is, the more sensitive the tests are.
2. There is a clear effect of the level of noise: the larger the level of noise is, the less sensitive are the tests.
3. There is a clear effect of the difference in the path coefficients in the two segments: the greater the difference in the path coefficients, the more sensitive the tests are.

These results are graphically illustrated in Figure (4.7). There are nine plots which represent each of the trends mentioned above: from left to the right we graph first the F-global's p-value, second the F-block and finally the F-coefficient's p-value. This allows us to obtain a direct comparison of the three tests. For the sake of interpretation, we have included in each plot the LOWESS (Cleveland, 1979) regression line of the p-value with respect to the evaluated experimental condition. In the first three plots, it is possible to observe how the p-values decrease (i.e., they become more significant) as the sample size increases. In the second set of plots, we can see that a lower sensitivity occurs when the level of noise in the standard deviation of manifest variable error term becomes higher. The influence of the difference between the path coefficients can be clearly appreciated in the last three plots, in which the p-values decrease as the difference between path coefficients increases.

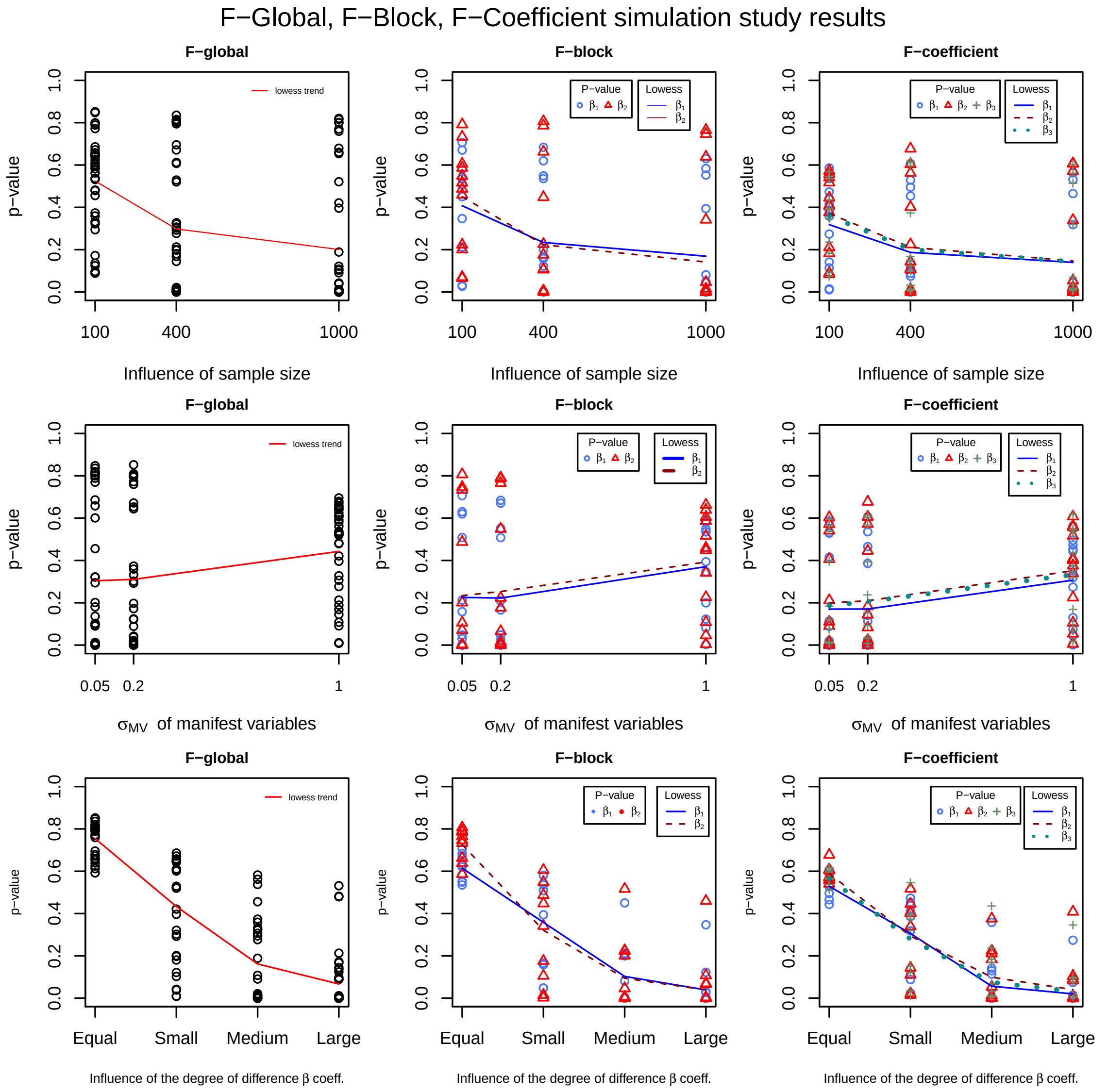


Figure 4.7: Comparison of the F-block and the F-coefficient statistics in respect to the F-global by the distinct simulation factors

Some difference between our simulation and Sánchez's work is evidenced by the factor different levels of variance of the endogenous latent variables. We didn't include this in our study, because controlling the path coefficient values in the simulation process causes the path coefficients to absorb the effect of the variance of the endogenous latent variable, meaning that high values of path coefficients produce lower variability in the residuals. Conversely, lower values will cause residuals with more variability.

Summarized results

The results of the simulations are summarized in Figure (4.8), (4.9) and (4.10). Each figure corresponds to a different sample size. For instance, Figure (4.8) shows the summarized results for different levels of error variance terms, as well as the various levels of difference between path coefficients obtained with a sample size of 100. We can see that, when the sample size is small (see Figure (4.8)), the trend of the p-value for the three statistics is not so marked when the standard deviation error terms of the manifest variables are equal to 1 (the last three plots in Figure (4.8)): the p-value decreases, as we expected, but it never goes below 0.2. This effect disappears when the sample size increases to 400 and 1000 (see Figure (4.9) and (4.10)). As in the aggregated results presented in Figure (4.7), the F-global, the F-block and the F-coefficient present very similar results.

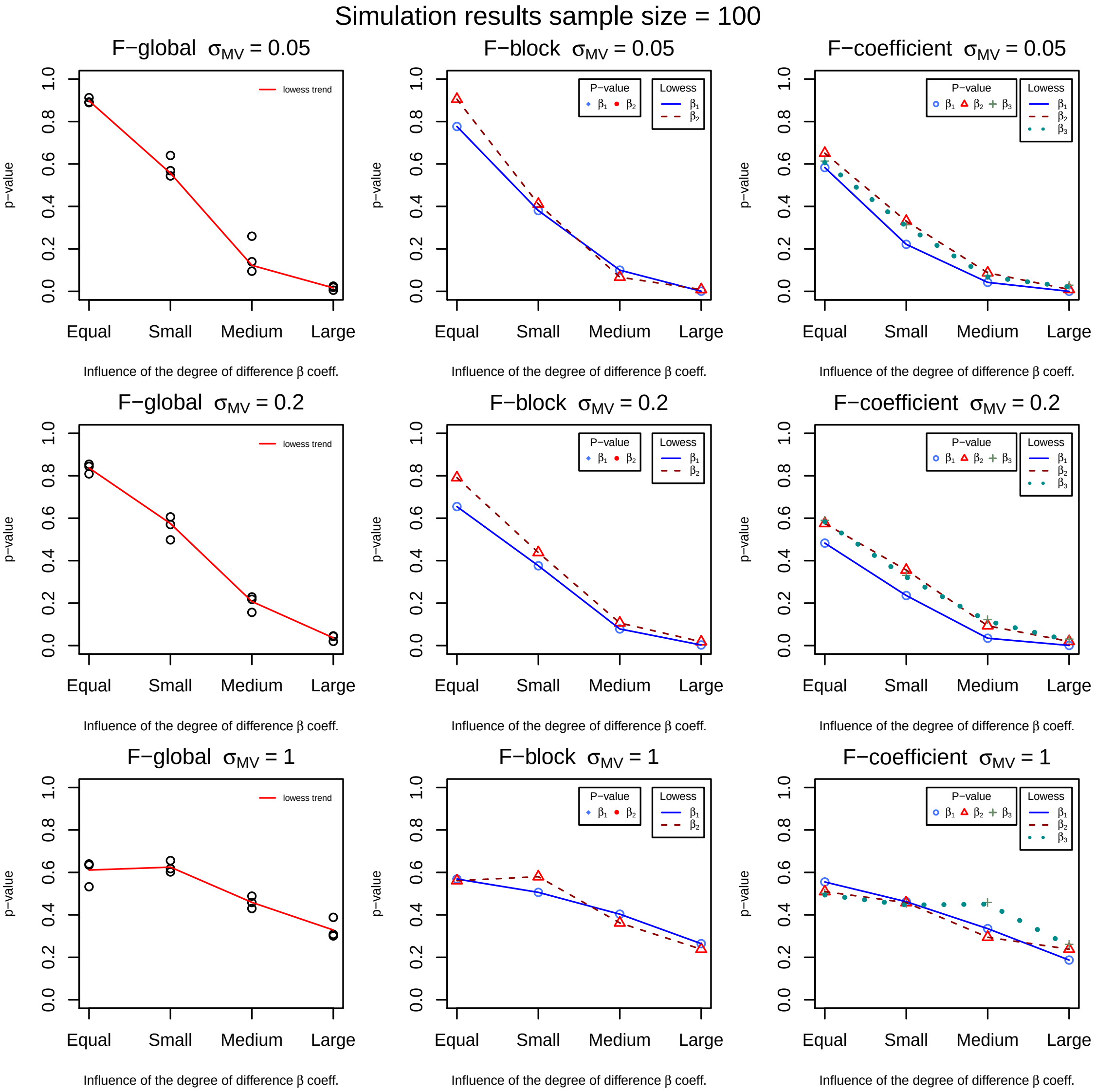


Figure 4.8: Comparison of the F-block and the F-coefficient statistics in respect to the F-global. Sample size 100

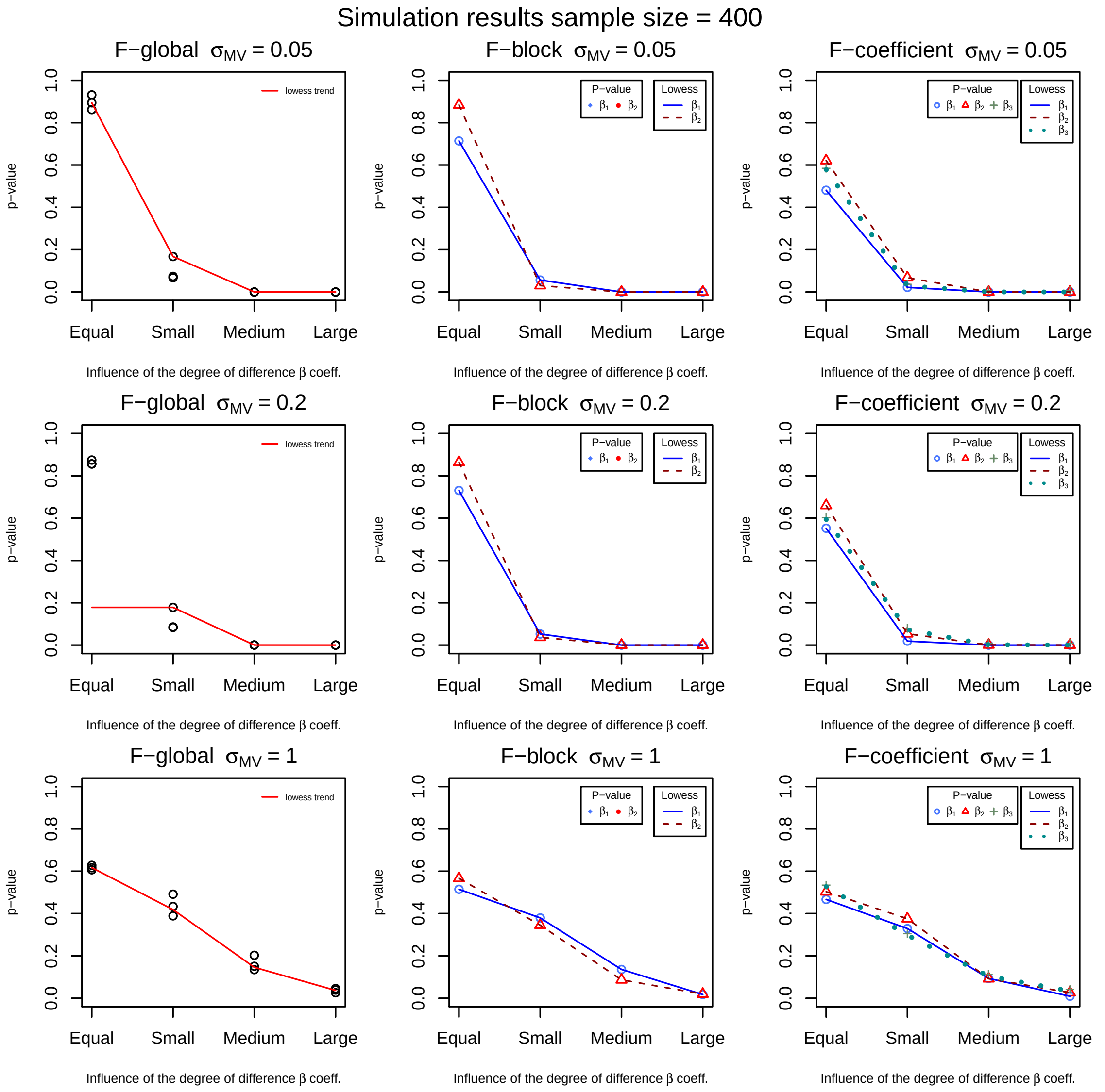


Figure 4.9: Comparison of the F-block and the F-coefficient statistics in respect to the F-global. Sample size 400

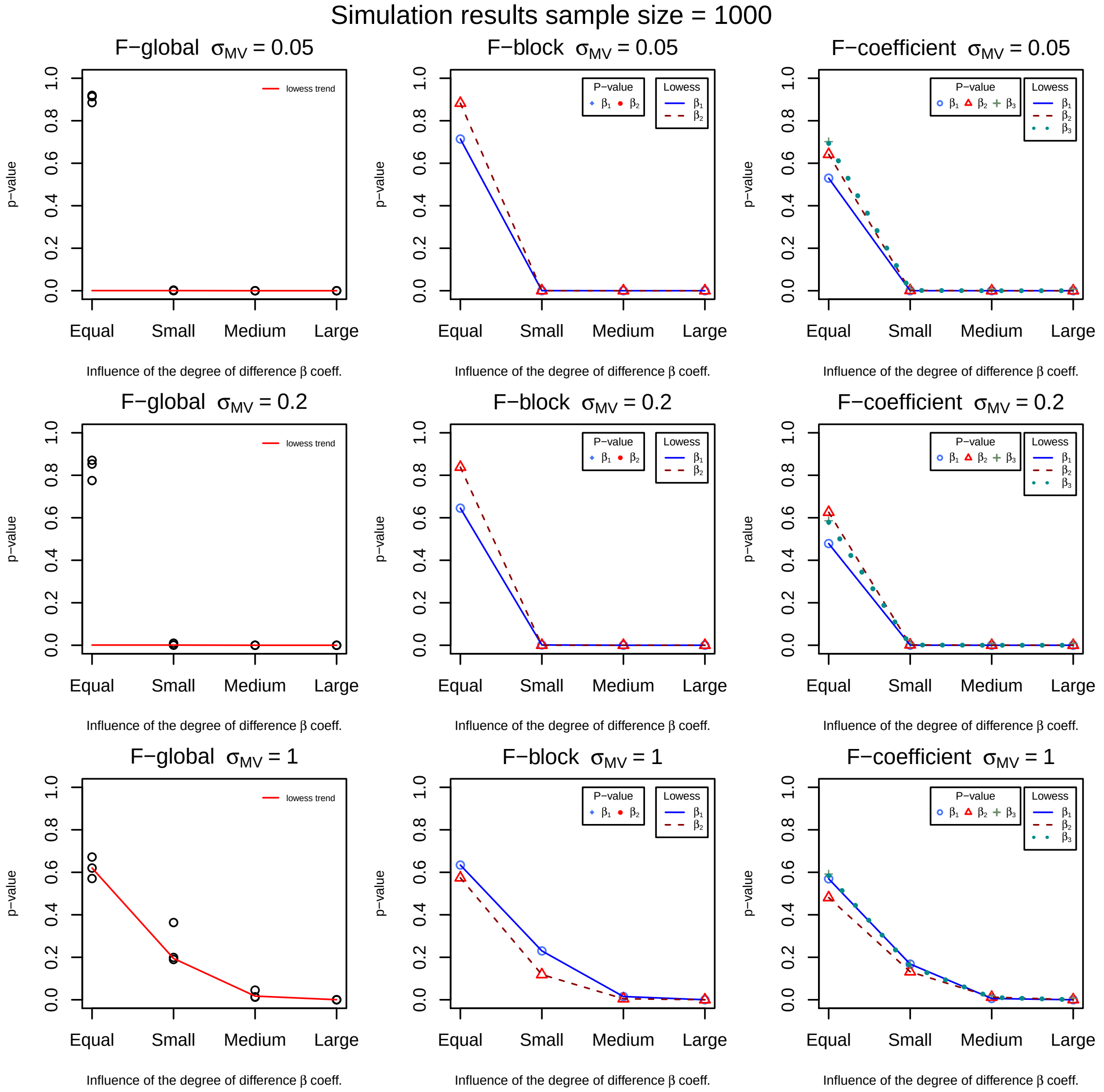


Figure 4.10: Comparison of the F-block and the F-coefficient statistics in respect to the F-global. Sample size 1000

4.6.3 Simulation Study to Compare the F Statistics with the F_{LAD} Approach

In this second study we want to compare the three F statistics obtained by least squares estimation with those obtained by the LAD approximation. The goal is to evaluate the degree of sensitivity in order to verify if the two approaches identify the same differences. In this way, we can establish under which conditions the LAD approximation is as suitable as the classic F-statistics are. To do this, we have used the same data generated by the first simulation study. Again, the performance of the tests is analyzed by the different sample sizes (100, 400, and 1000), the different levels of the standard deviation terms of the manifest variables, and the various levels of difference between the path coefficients. In total, 3 × 3 × 4 = 36 scenarios, which are the number of possible combinations of sample sizes, noise levels, and the differences between coefficients. We run 50 repetitions for each experimental condition and present the mean of each 50 repetitions as an aggregate result.

Observing Figure (4.11)), we can observe the same patterns shown by the F statistics that were obtained by least squares estimation (Figure (4.7)):

1. The larger the sample size is, the more sensitive the tests are.
2. The larger the level of noise is, the less sensitive the tests are.
3. The greater the difference in the path coefficients, the more sensitive the tests are.

As we expected, the trend of the three statistics is similar to those obtained by the least squares estimation; the only difference is that the p-values calculated with the LAD approximation are more conservative. This aspect is fully explainable by the nature the LAD approximation statistics.

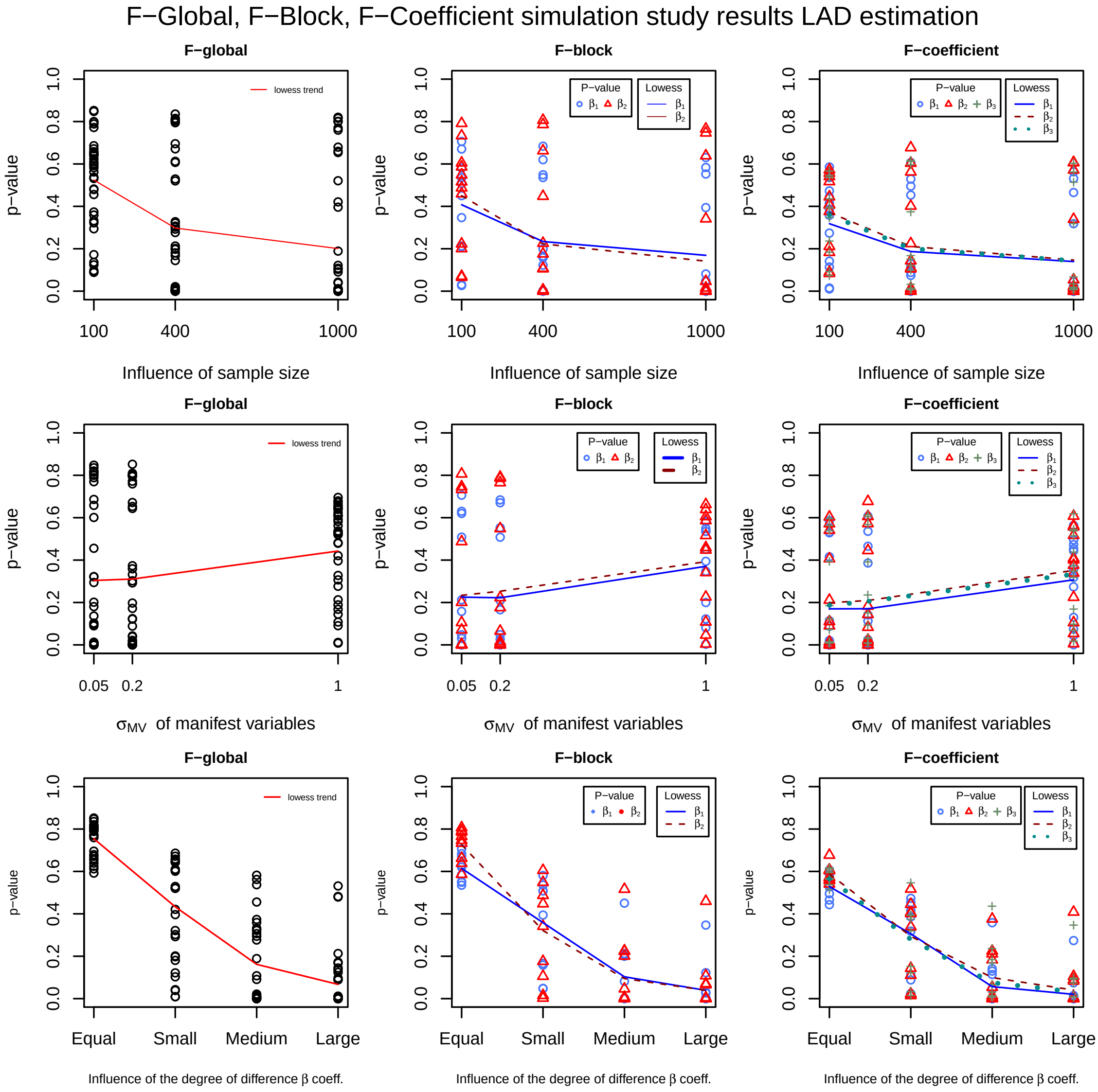


Figure 4.11: Comparison of the F-block and the F-coefficient statistics in respect to the F-global, by the distinct simulation factors - LAD approach

Summarized results

The summarized results from the simulations are contained in Figures (4.12), (4.13) and (4.14). As before, each figure corresponds to a different sample size. Again, we find the same pattern observed with least squares estimation. When the sample size is small (see Figure (4.8)), and the error terms of the manifest variables are equal to 1, the trend of the p-value for the three statistics is not so marked. But, again, this effect disappears when the sample size increases.

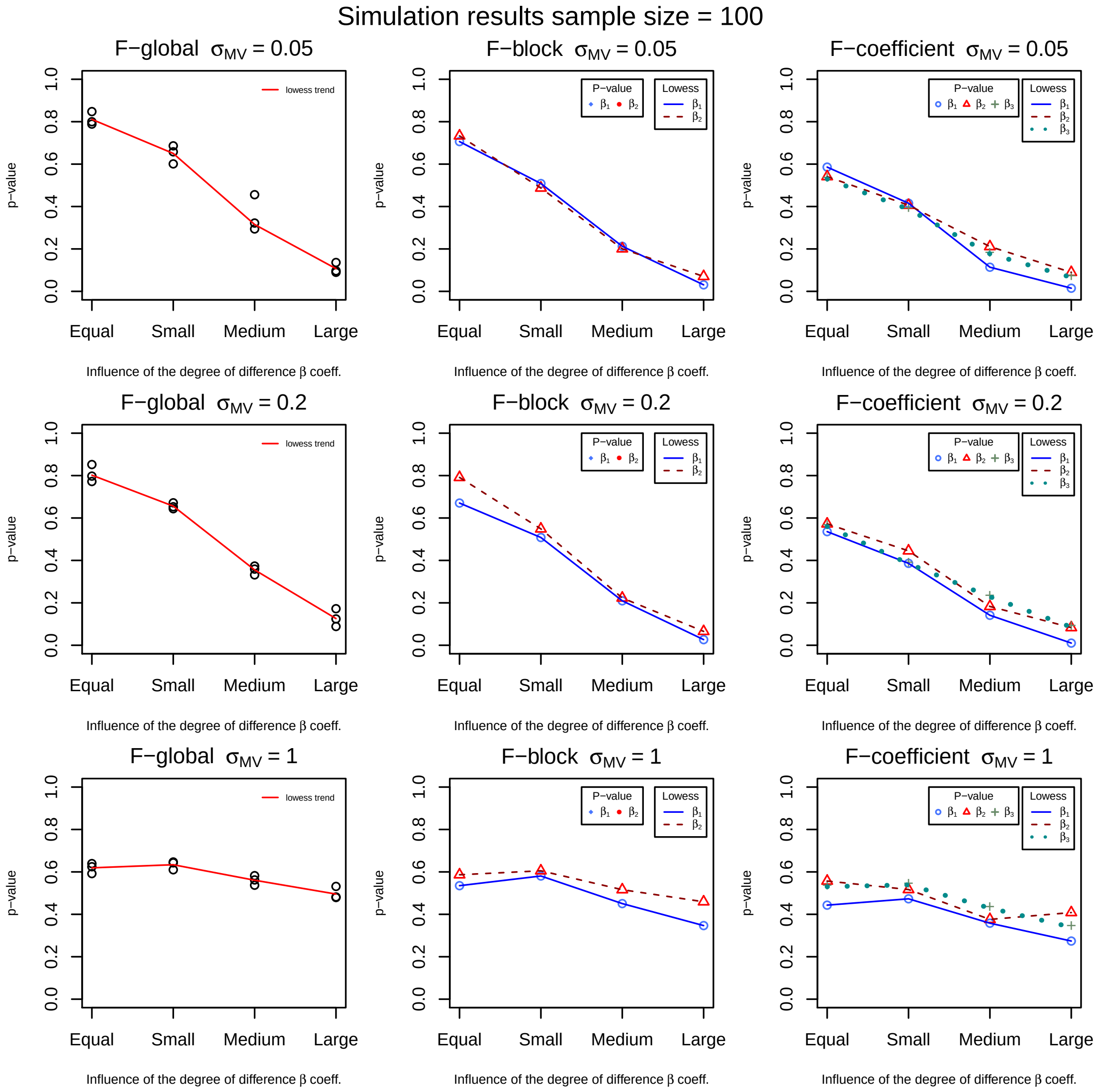


Figure 4.12: Comparison of F-global, F-block, and the F-coefficient statistics, LAD approach. Sample size 100

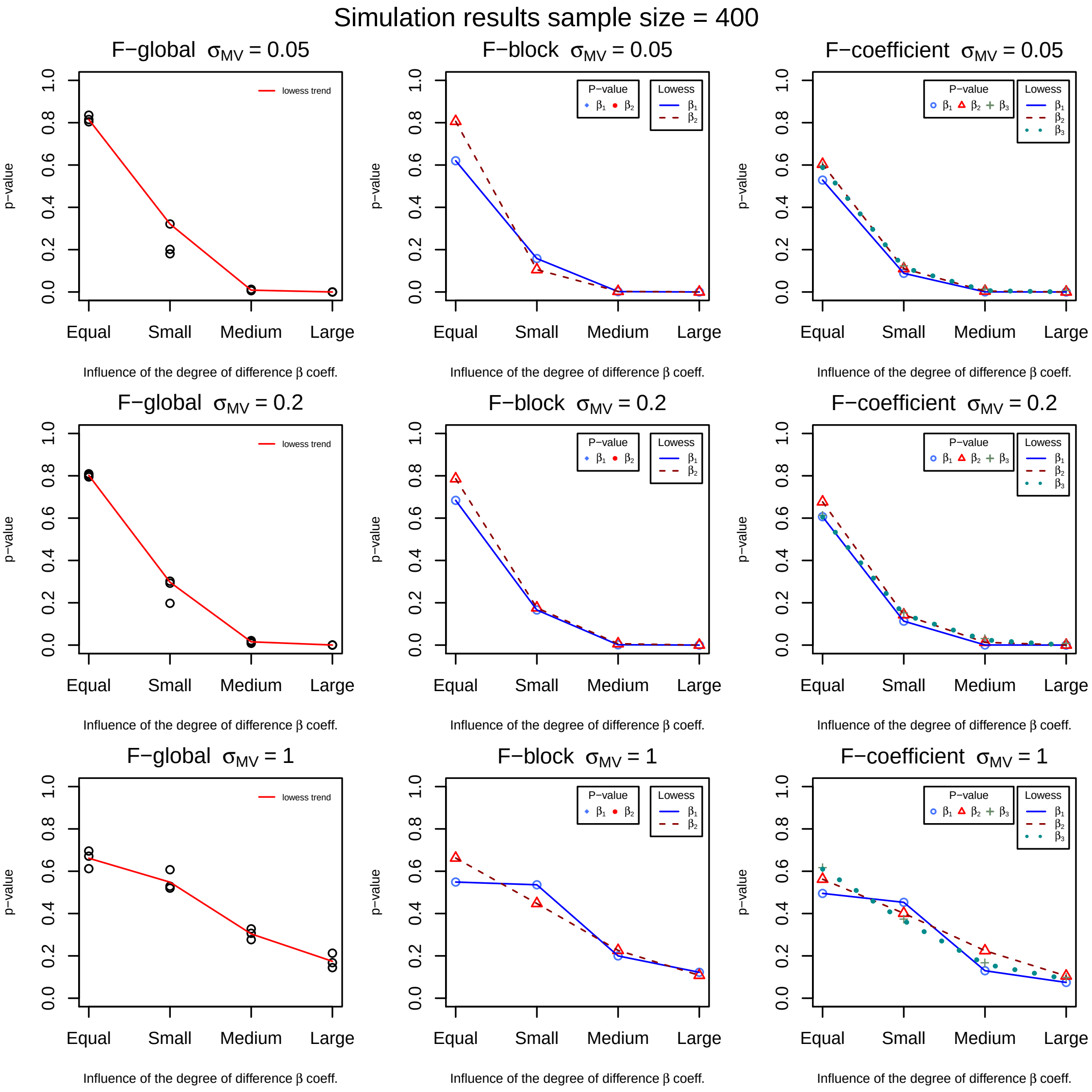


Figure 4.13: Comparison of F-global, F-block, and the F-coefficient statistics, LAD approach. Sample size 400

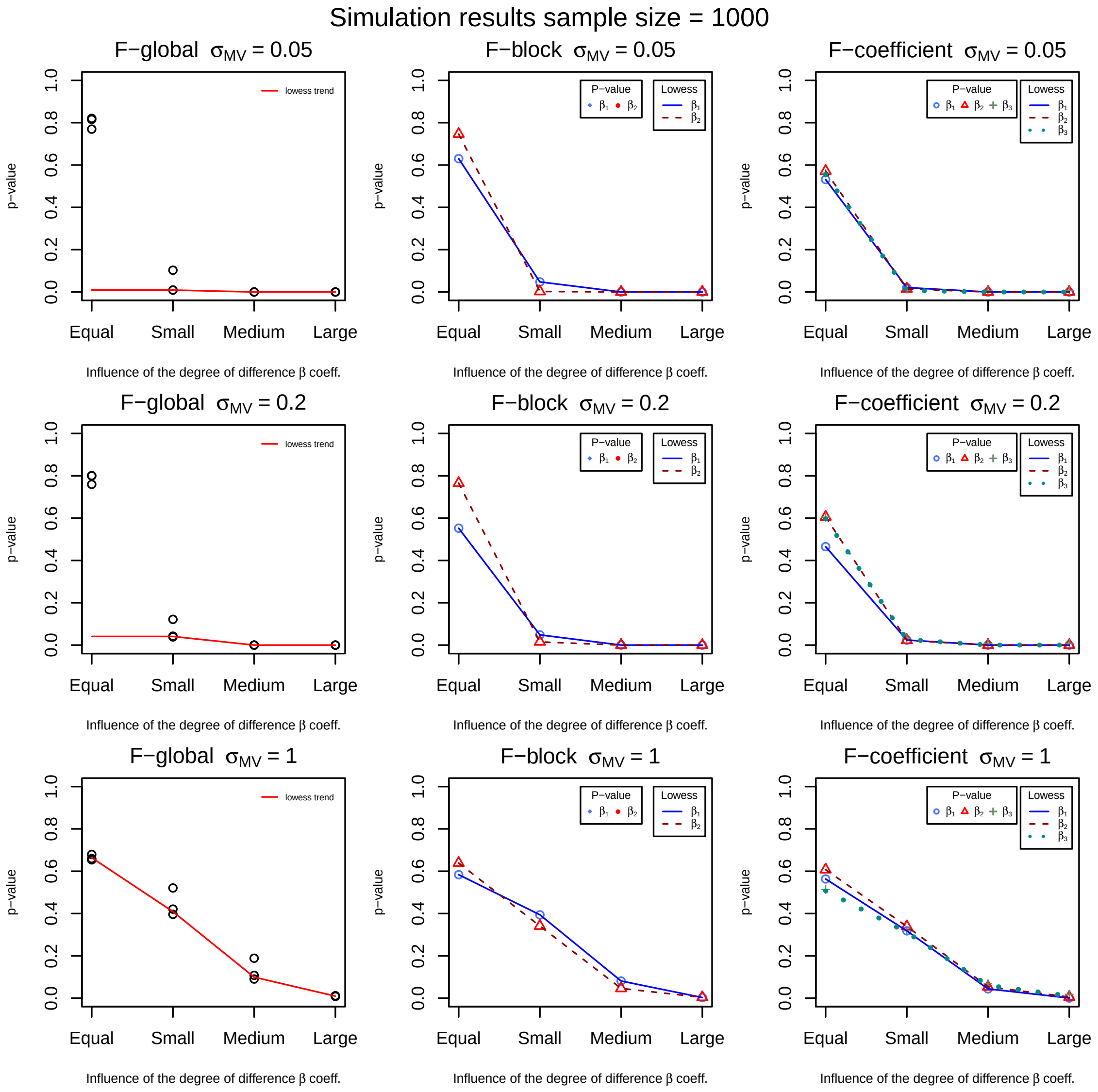


Figure 4.14: Comparison of F-global, F-block, and the F-coefficient statistics, LAD approach. Sample size 1000

4.6.3.1 Comparing the Least Squares and LAD P-value's Test

As we have seen, the least squares and the LAD have very similar behavior, the only difference being that the LAD statistics are more conservative. In Tables (4.3), (4.4) and (4.5) we compare the performance of the p-value of both tests under normal data and considering: the difference between coefficients, the sample size and the standard deviation error of the manifest variables . We can observe that in all cases the LAD p-values are higher, even if there is a difference between the path coefficients; whereas the p-values are lower when the path coefficients are equal in both segments. When we vary the standard deviation error of the manifest variables and the sample size (from 100 to 1000), we continue to observe the same pattern. These results suggest that the LAD statistics will be more adequate if the PLS-PM model contains outliers, or if the data are clearly non-normal.

	Factors	**F-global p-value**		**F-block eq_1 p-value**		**F-block eq_2 p-value**		**F-coefficient β_1 p-value**		**F-coefficient β_2 p-value**		**F-coefficient β_3 p-value**	
σ_{MV}	**β Difference**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**
0.05	Equal	0.8978	0.8120	0.8892	0.8476	0.9124	0.7887	0.5832	0.5859	0.6509	0.5410	0.6066	0.5299
	Small	0.5841	0.6484	0.6402	0.6862	0.5684	0.6010	0.2215	0.4155	0.3307	0.4062	0.3061	0.3873
	Medium	0.1649	0.3574	0.2598	0.4558	0.1398	0.3222	0.0423	0.1139	0.0874	0.2112	0.0681	0.1769
	Large	0.0163	0.1077	0.0053	0.0900	0.0249	0.1362	0.0001	0.0149	0.0091	0.0896	0.0208	0.0661
0.2	Equal	0.8978	0.8073	0.8892	0.7975	0.9124	0.8521	0.5832	0.5357	0.6509	0.5723	0.6066	0.5601
	Small	0.5841	0.6562	0.6402	0.6717	0.5684	0.6530	0.2215	0.3861	0.3307	0.4451	0.3061	0.3850
	Medium	0.1649	0.3549	0.2598	0.3735	0.1398	0.3591	0.0423	0.1414	0.0874	0.1829	0.0681	0.2290
	Large	0.0163	0.1284	0.0053	0.0887	0.0249	0.1241	0.0001	0.0102	0.0091	0.0833	0.0208	0.0856
1	Equal	0.8978	0.6186	0.8892	0.6244	0.9124	0.5920	0.5832	0.4432	0.6509	0.5568	0.6066	0.5301
	Small	0.5841	0.6332	0.6402	0.6100	0.5684	0.6429	0.2215	0.4731	0.3307	0.5169	0.3061	0.5389
	Medium	0.1649	0.5606	0.2598	0.5373	0.1398	0.5622	0.0423	0.3581	0.0874	0.3760	0.0681	0.4292
	Large	0.0163	0.4976	0.0053	0.4819	0.0249	0.5315	0.0001	0.2740	0.0091	0.4083	0.0208	0.3396

Table 4.3: Sample size 100

	Factors	F-global p-value		F-block eq_1 p-value		F-block eq_2 p-value		F-coefficient β_1 p-value		F-coefficient β_2 p-value		F-coefficient β_3 p-value	
σ_{MV}	**β Difference**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**
0.05	Equal	0.8958	0.8183	0.8941	0.8151	0.8618	0.8355	0.4807	0.5287	0.6207	0.6037	0.5784	0.5883
	Small	0.1029	0.2344	0.1676	0.3213	0.0737	0.1806	0.0219	0.0883	0.0675	0.1103	0.0295	0.1152
	Medium	0.0000	0.0090	0.0000	0.0128	0.0000	0.0079	0.0000	0.0005	0.0001	0.0043	0.0003	0.0065
	Large	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0002
0.2	Equal	0.8958	0.8022	0.8941	0.7946	0.8618	0.8096	0.4807	0.6067	0.6207	0.6770	0.5784	0.6057
	Small	0.1029	0.2641	0.1676	0.3022	0.0737	0.2924	0.0219	0.1132	0.0675	0.1433	0.0295	0.1442
	Medium	0.0000	0.0153	0.0000	0.0085	0.0000	0.0161	0.0000	0.0005	0.0001	0.0130	0.0003	0.0242
	Large	0.0000	0.0004	0.0000	0.0001	0.0000	0.0004	0.0000	0.0000	0.0000	0.0001	0.0000	0.0005
1	Equal	0.8958	0.6597	0.8941	0.6123	0.8618	0.6956	0.4807	0.4951	0.6207	0.5623	0.5784	0.6101
	Small	0.1029	0.5522	0.1676	0.6073	0.0737	0.5201	0.0219	0.4530	0.0675	0.4012	0.0295	0.3662
	Medium	0.0000	0.3037	0.0000	0.3277	0.0000	0.3067	0.0000	0.1298	0.0001	0.2246	0.0003	0.1606
	Large	0.0000	0.1750	0.0000	0.2128	0.0000	0.1678	0.0000	0.0745	0.0000	0.1051	0.0000	0.0907

Table 4.4: Sample size 400

	Factors	**F-global p-value**		**F-block eq_1 p-value**		**F-block eq_2 p-value**		**F-coefficient β_1 p-value**		**F-coefficient β_2 p-value**		**F-coefficient β_3 p-value**	
σ_{MV}	**β Difference**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**	**LS**	**LAD**
0.05	Equal	0.9060	0.8018	0.8853	0.8202	0.9193	0.7696	0.5298	0.5316	0.6419	0.5716	0.6940	0.5551
	Small	0.0013	0.0403	0.0030	0.1029	0.0006	0.0089	0.0001	0.0209	0.0012	0.0146	0.0014	0.0073
	Medium	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Large	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.2	Equal	0.9060	0.7868	0.8853	0.7595	0.9193	0.8009	0.5298	0.4653	0.6419	0.6058	0.6940	0.5966
	Small	0.0013	0.0675	0.0030	0.1216	0.0006	0.0387	0.0001	0.0239	0.0012	0.0227	0.0014	0.0264
	Medium	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Large	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	Equal	0.9060	0.6644	0.8853	0.6793	0.9193	0.6539	0.5298	0.5634	0.6419	0.6075	0.6940	0.5071
	Small	0.0013	0.4461	0.0030	0.5210	0.0006	0.4212	0.0001	0.3183	0.0012	0.3391	0.0014	0.3113
	Medium	0.0000	0.1292	0.0000	0.1889	0.0000	0.0907	0.0000	0.0442	0.0000	0.0537	0.0000	0.0605
	Large	0.0000	0.0099	0.0000	0.0118	0.0000	0.0086	0.0000	0.0009	0.0000	0.0055	0.0000	0.0105

Table 4.5: Sample size 1000

4.6.4 Simulation Study to Evaluate the Performance of the Invariance Measurement Test

The goal of the third analysis is to investigate the performance of the invariance measurement test. Here, we want to determine the different conditions under which this test is suitable for understanding if the measurement model can be considered the same for all terminal nodes of the tree, or if it is specific for each one of them. We will take into account: different sample sizes (100, 400, and 1000); the different levels in the error's standard deviation terms of the manifest variables; the various levels of difference between the path coefficients; and the various levels of difference between the coefficients (loadings in Model 1 and weights in Model 2) of the measurement model. In total, we have $3 \times 3 \times 4 \times 4 = 36$ scenarios, which are the number of possible combinations of sample sizes, noise levels, differences between path coefficients and differences between measurement coefficients. We run 50 repetitions for each experimental condition and present the mean of each of the 50 repetitions as an aggregate result.

We provide the results for both Model 1 and Model 2 in order to verify the effectiveness of the invariance measurement test when the relationship between the manifest variables and the construct is both *reflective* and *formative*.

Observing the results for both models, we can conclude that:

1. There is a clear effect of sample size: the larger the sample size is, the more sensitive the tests are.

2. There is a clear effect of the difference between the loadings in the two segments: the greater the difference in the loadings, the more sensitive the test is.

3. There is no difference in the invariance measurement test between Models 1 and 2.

In the case of the level of noise and the difference between the path coefficients, the trend is not so marked.

These results are graphically illustrated in Figure (4.15). There are four plots which represent each of the trends mentioned above. Also, for the sake of interpretation in this case, we have included in each plot the LOWESS (Cleveland, 1979) regression line of the p-value with respect to the evaluated experimental condition. In the first plot, it is possible to observe how the p-values decrease (i.e., they become more significant) as the sample size increases. In the second plot, we can see that lower sensitivity occurs when the level of noise becomes higher for the standard deviation of the error terms of the manifest variables. In this case the effect is almost null. As in the case of the factor *level of noise*, the influence of the difference between the path coefficients does not produce a significant trend. We can explain this by the fact that the test of invariance only takes into account the measurement model of the terminal nodes. The influence of the difference between loadings can be clearly appreciated in the last plot, where the p-values decrease as the difference between path coefficients increases.

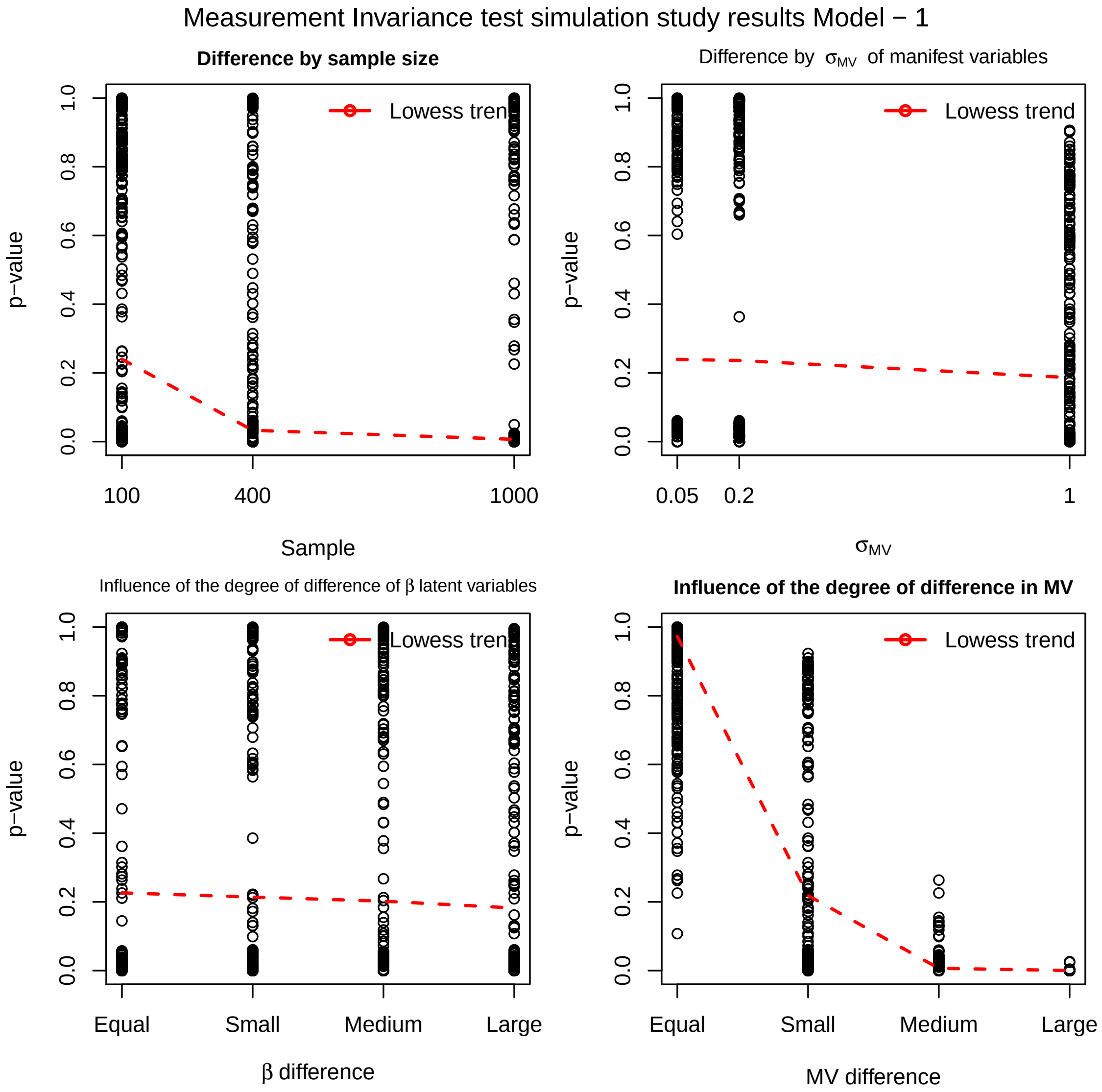


Figure 4.15: Comparison of the Invariance Measurement test by the Distinct Simulation Factors Model 1

Summarized results

The summarized results from the simulations are contained in the Figure (4.16). Here, we graph the different levels in error's standard deviation terms of the manifest variables, the different level of difference between the path coefficients and different level of difference between the measurement model coefficients taking in account the three sample size: 100, 400, 1000. We can see that again, there is no significative trend when different levels of errors of standard deviation, and different levels of differences between the path coefficients are considered. The only significative result is obtained by the factor different level of difference between the measurement model coefficients (last three plots in the Figure (4.16)).

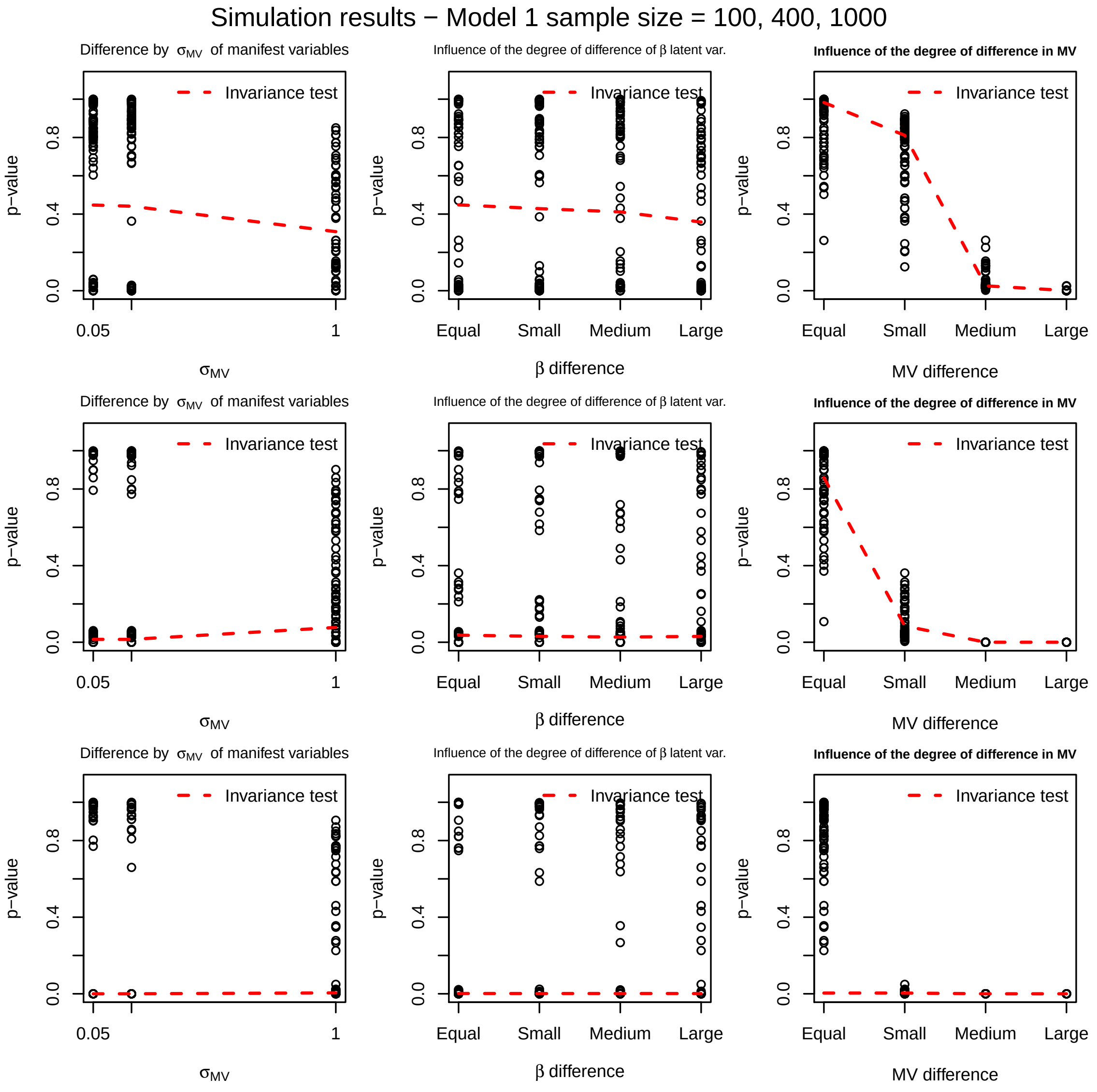


Figure 4.16: Comparison of the invariance measurement test by the distinct simulation factors in Model 1 - Sample size: 100, 400, 1000

The results of Model 2 are presented in Figure (4.17).The outcomes show that when the relationship between latent variables and manifest variables is also formative (ξ blocks in our case), the test detects very well the differences between measurement model coefficients.

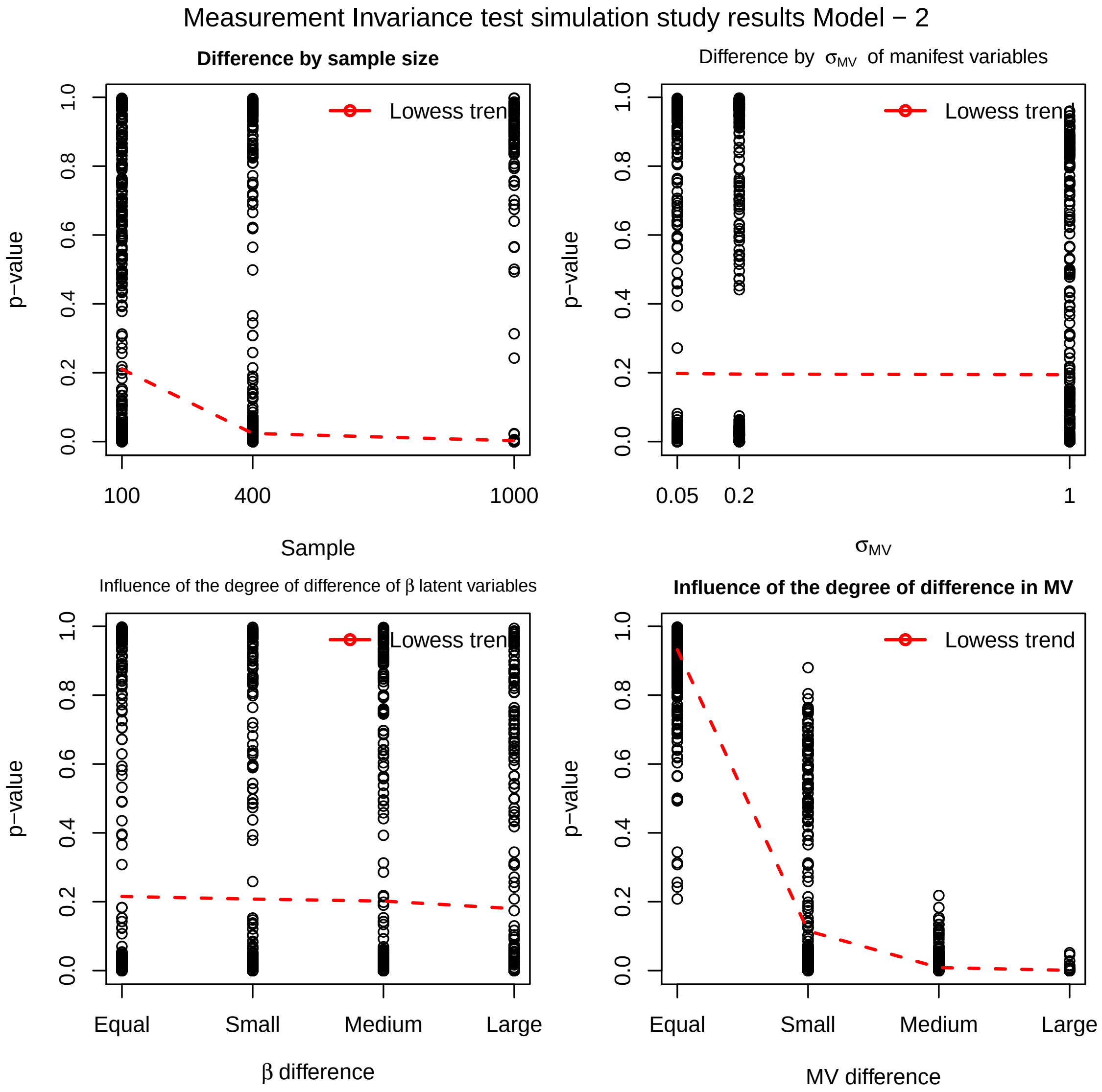


Figure 4.17: Comparison of the Invariance measurement test by the distinct simulation factors Model 2

Summarized results

The summarized results from the simulations for Model 2 are presented in Figure (4.18). Again we can find the same trends shown in Figure (4.16)

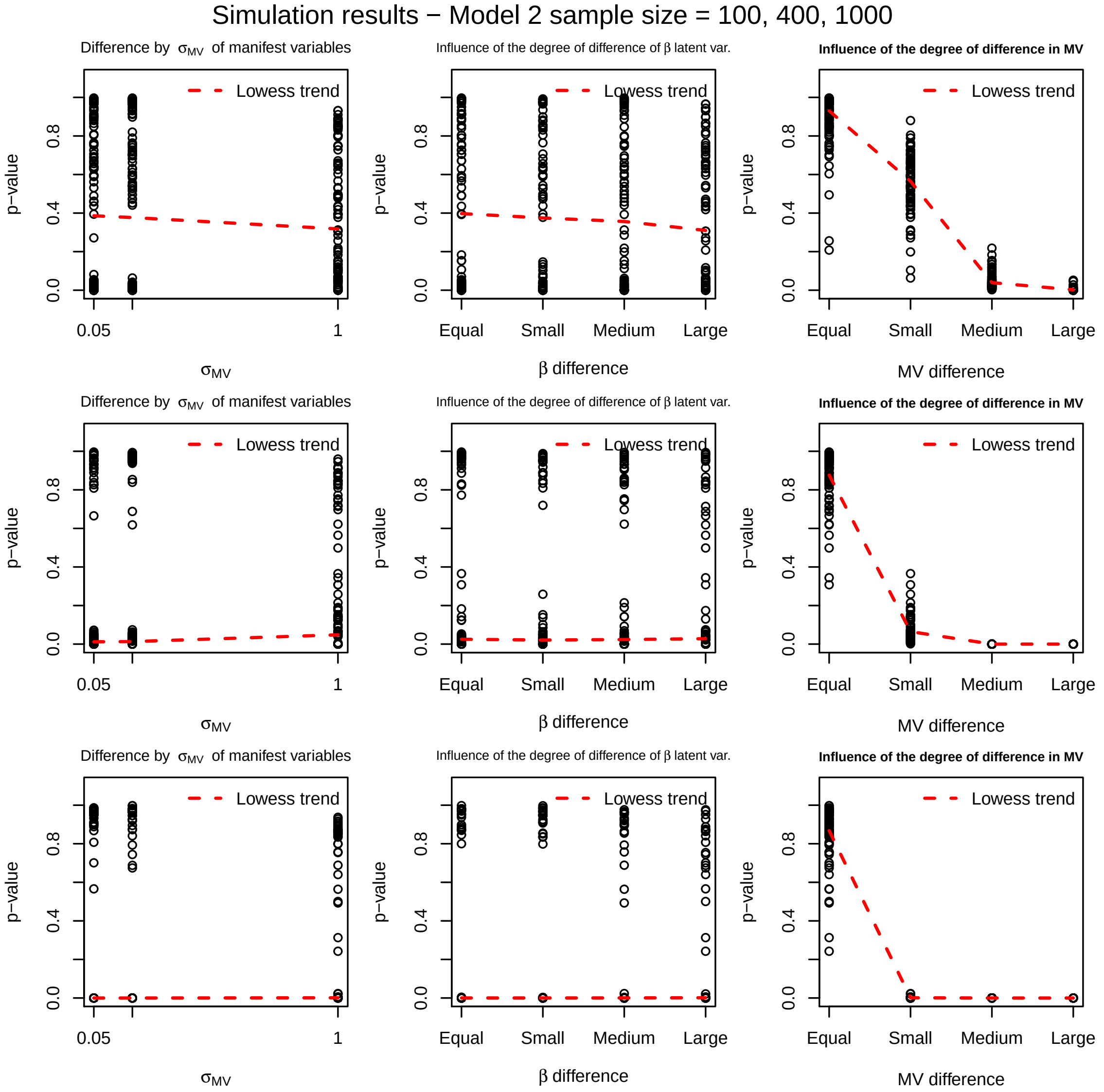


Figure 4.18: Comparison of the Invariance measurement test by the distinct simulation factors Model 2 - Sample size: 100, 400, 1000

Chapter 5

PATHMOX applications

In this chapter we present an application of PATHMOX with real data for each technique treated in Chapter 3. To this end, we consider a data-set on the evaluation of alumni Satisfaction and its comparison in two main ICT schools: the Informatics and Telecommunications schools of the UPC (Universitat Politécnica de Catalunya). In the context of partial least squares path modeling, we also present an application based on the mental health data-set, where we analyze three mental disorders that are common in elderly populations: Dementia, Delirium and Depression. We can divide this chapter into three main sections: in the first one, we describe data (section 5.1); in the second one, we present four applications of PATHMOX with multiple linear regression (section 5.2), robust absolute regression (section 5.3), logistic regression (section 5.5) and principal component analysis (section 5.6); in the last one, we consider two real-world study with partial least squaress path modeling, using data-sets on both, alumni satisfaction (section 5.7) and mental health (section 5.9). All the results have been obtained by the “genpathmox” **R** package (Lamberti, 2014) available on CRAN.

5.1 Data

To illustrate the application of PATHMOX, we will use two data-sets on alumni satisfaction and mental health. The first one is based on the evaluation of the alumni satisfaction and its comparison in two main ICT schools at UPC (Universitat Politécncia de Catalunya) : Informatics and Telecommunications schools; the second one contains information related to three mental disorders that are typically found in elderly populations: Dementia, Delirium and Depression.

5.1.1 Alumni Satisfaction Data-set

The data come from a survey of 147 students realized in 2008. It consists of a total of 32 variables collected by a questionnaire based on the ECSI model questions (see Table (5.1)). The students were asked to provide measures on a 11-point ordinal scale, ranging from very satisfied (10) to very dissatisfied (0). The goal of the study was to explain the Satisfaction from its drivers, Image of the school, Expectations on generic skills [1], Expectations on technical skills [2], Perceived Quality on generic skills, Perceived Quality on technical skills and Value or Profit obtained after graduation. This variables are defined in the following way:

- **Satisfaction** Degree of alumni satisfaction about the formation in school respect to their actual work conditions;
- **Image** Generic alumni perception of ICT schools: (internationally recognition, ranges of courses, leadership in research, . . .);
- **Specific Expectation** Perceived Expectation on specific skills (technical or applied skills);
- **Generic Expectation** Perceived Expectation on generic skills (abilities in problem solving, communication skills);
- **Generic Quality** Perception about achieved quality on the generic skills in the school (abilities in solving problem, communication skills);
- **Specific Quality** Perception about the achieved quality on the specific skills in the school;
- **Value** The advantage or profit that the alumni may draw from the school degree (well paid job, motivating job, prospectives for improvement and promotion).

[1] We understand by general skills a broad - spectrum of capabilities not specific to a profession or organizational environment, such as the ability of problem solving, communication, time management team working initiative . . .

[2] The technical skills refers to the knowledge and abilities, specific to a profession, either mathematical or engineering based, or specific to accomplish a technical tasks.

Table 5.1: Description of the manifest variables for each of the latent constructs - alumni satisfaction dataset

Latent variables	**Description**
Image	1 - It is the best college to study IE 2 - It is internationally recognized 3 - It has a wide range of courses 4 - The professors are good 5 - Facilities and equipment are good 6 - It is a leader in research 7 - It is well regarded by companies 8 - It is oriented to new needs and technologies
Specific expectations	1 - Basic skills 2 - Specific Technical skills 3 - Applied skills
Generic expectations	1 - Achieved abilities in solving problem 2 - Training in business management 3 - Written and oral communication skills 4 - Planning and time management acquired 5 - Team-work skills
Specific Quality	1 - Basic skills 2 - Specific Technical skills 3 - Applied skills
Generic Quality	1- Achieved abilities in solving problem 2 - Training in business management 3 - Written and oral communication skills 4 - Planning and time management acquired 5 - Team-work skills
Value	1 - It has allowed me to find a well paid job 2 - I have good prospectives for improvement and promotion 3 - It has allowed me to find a job that motivates me 4 - The training received is the basis on which I will develope my career
Satisfaction	1 - I am satisfied with the training received 2 - I am satisfied with my current situation 3 - I think I'll have a good career 4 - How prestigious is your job?

10 segmentation variables were also collected to serve as observed sources of heterogeneity variables. A description is shown in Table (5.2).

Table 5.2: Codification of segmentation variables according to their type of scale and levels

Segmentation variables			
Name	**Scale**	**N. levels**	**Levels description**
Career	nominal	3	EI / ETS / TEL
Gender	binary	2	female / male
Age	ordinal	4	25-26 years / 27-28 years / 29-30 years / more than 30 years
Studying	binary	2	yes studying / no studying
Contract	nominal	3	fixed / temporary / other types
Salary	ordinal	5	less than 18k / 25k / 35k / 45k / more than 45k
Firm-type	binary	2	private / public
Access grade	ordinal	3	acc-note less than 7 / acc-note 7-8 / more than 8
Grade	ordinal	4	note less than 6.5 / note 6.5-7 / note 7-7.5 / more than 7.5
Start-work	binary	2	after graduating / before graduating

5.1.2 The Mental Health Data-set

In the mental Health data set, we analyze data on 138 elderly patients from seven Quebec hospitals, observed between July 2005 and January 2007. The data were assembled from a team of St. Mary's Hospital Research Centre and were previously analyzed to answer a number of specific research questions. Our aim was, to use the partial least squares methodology, to investigate the relationship between three constructs representing three mental disorders that are common in elderly populations: Dementia, Delirium and Depression. A total of 27 variables were available, divided into two groups; one group was formed by 24 manifest variables; and the other by 3 segmentation variables: patient's gender, duration of hospitalizations, patient's age. We defined a measurement model for Dementia based on the items of two well-known instruments: the Hierarchical Dementia Scale (HDS) (Cole, 1988) and the Mini Mental State Examination MMSE (Folstein *et al.*, 1975). Similarly, to define the measurement models for Depression and Delirium, we used, respectively the Cornell scale for assessing Depression (Alexopoulos *et al.*, 1988) and of the Delirium Index (Pompei *et al.*, 1995) as measure of Delirium severity.

A description of the manifest variables used to estimate the latent variables is showed in the following table:

Table 5.3: Description of the manifest variables for each of the latent construct - Mental health data-set

LV	MV	Item
MMSE	$mmse_3$	What month of the year is this?
	$mmse_8$	What city are we in?
	$mmse_{11}$	I am going to say 3 words. After I have said all three, I want you to repeat them
	$mmse_{12}$	Spell the word “world”
	$mmse_{16}$	Repeat the following phrase: “no ifs, ands or buts”
	$mmse_{17}$	Take this paper in your right/left hand, fold the paper in half and put it on the floor
	$mmse_{19}$	Copy this design
HDS	hds_2	Prefrontal subscale
	hds_6	Denomination subscale.
	hds_7	Verbal comprehension subscale
	hds_{11}	Reading subscale.
	hds_{12}	Recent memory subscale
	hds_{18}	Motor subscale
	hds_{20}	Writing subscale
	hds_{22}	Similarities subscale
Depression	$corn_5$	Agitation: restlessness, hand wringing, hair pulling
	$corn_8$	Loss of interest: less involved in usual activities
	$corn_{15}$	Early morning awakening: earlier than usual for this individual
	$corn_{19}$	Mood congruent delusions: delusions of poverty, illness or loss
Delirium	del_4	Disorganized thinking
	del_5	Altered level of consciousness
	del_7	Memory impairment
	del_8	Perceptual disturbances
	del_{10-11}	Psychomotor agitation and Psychomotor retardation

5.2 PATHMOX Approach Applied to the Context of Linear Regression Model

As a first application, we consider the linear multiple regression model. In this case, using the alumni satisfaction data-set will allow us to explain with a simple additive model, the Satisfaction from its drivers, Image of the school, Expectations on general skills, Expectations on technical skills, Perceived Quality on general skills, Perceived Quality on technical skills and Value or Profit obtained after graduation.

5.2.1 Global Linear Model

We start our analysis with the estimation of one global model for all students. In Table (5.4) we show, from the left to right, the coefficient estimate, the standard deviation, the value of the statistic t and the associated p-value. As we can see, the main drivers of Satisfaction are: Value with a coefficient of *0.575* and a p-value of *zero*; and Image, with a coefficient of *0.240* and a p-value of *0.002*; followed by the Generic Expectations (coefficient: *0.111*, and p-value: *0.043*) and Specific Quality (coefficient: *0.123* and p-value: *0.067*). The coefficients Specific Expectations (*-0.083*), and Generic Quality (*0.026*) are not significant. The model presents an R^2 of *0.649*, which is considered as adequate value for good predictive power of the model.

Table 5.4: Coefficient estimate and statistics, global model

LS regression	**Estimate**	**Std. error**	**t value**	**Pr(>$\|t\|$)**
Image	0.240	0.078	3.086	0.002
Generic Expectations	0.111	0.054	2.045	0.043
Specific Expectations	−0.083	0.054	−1.544	0.125
Generic Quality	0.026	0.071	0.373	0.710
Specific Quality	0.123	0.067	1.847	0.067
Value	0.575	0.055	10.415	0.000

As a second step in the analysis, we could repeat the estimation of the model without considering the non-significant coefficients: Specific Expectations and Generic Quality. But, since we utilize this data to show how PATHMOX works, we consider this model valid and we proceed to apply PATHMOX to our data.

5.2.2 Least Squares PATHMOX Approach

In order to calculate the PATHMOX segmentation tree, it is necessary to specify the scale (e.g., binary, ordinal, or nominal) of the segmentation variables (see Table (5.2)). In addition, we have to determine the parameters and stop conditions of the algorithm.

We have decided to establish a value of 0.05 for the threshold of the p-value in looking for those partitions that are highly significant. Given that we have a total sample of 147 students, it seems to us that 15 students (10% of the total sample) is a reasonable minimum number to stop the growth of a node. The depth level (= 2) has been selected with the aim of obtaining a simple segmentation tree with a possible maximum number of four final segments.

In Figure (5.1), we present the obtained tree, where we can observe that in fact, there are four distinct models. The F-global statistics, the p-values and the obtained partitions for each node are summarized in Table (5.5). At the first split, PATHMOX defines two different models one for students with a *“low"* access grade (< 7), and the other, for students with a *“high"* access grade (*7-8* and > 8). Let as define the students with a *“low"* access grade (< 7) as students with a*“low profile"*, and students with a *“high"* access grade (*7-8* and > 8) as students with a *“high profile"*. We see that the produced split is highly significant, giving a F-statistic of *2.823* with a p-value of *0.013*. The tree continues by splitting node two (students with a *“low"* profile) and three (students with a *"high"* profile). The most significant split for the node two is obtained by the variable *age*, giving a F-statistic of *2.529* with a p-value of *0.029*. A child node (node four) is obtained for the students with less than 30 years (≤ 30), and another (node five), for students of more than 30 years (> 30). Node three is split again by variable *acc-grade*, giving an F-statistic of *2.757* with a p-value of *0.020*. We divide now the students with a better profile, into students with a *"good"* profile (*acc-note 7-8*), and students with the *"best"* profile (*acc-note* $8 >$). This ends the splitting process, as the maximum depth of two levels has been reached. Hence, in the end, we have four final segments, each one corresponding to a distinct model: **node four** model of less or equal to than 30 years, with *"low"* profile students; **node five** model of more than 30 year, with *"low"* profile students; **node six** model of students with a *"good"* profile; **node seven** model of students with the *"best"* profile.

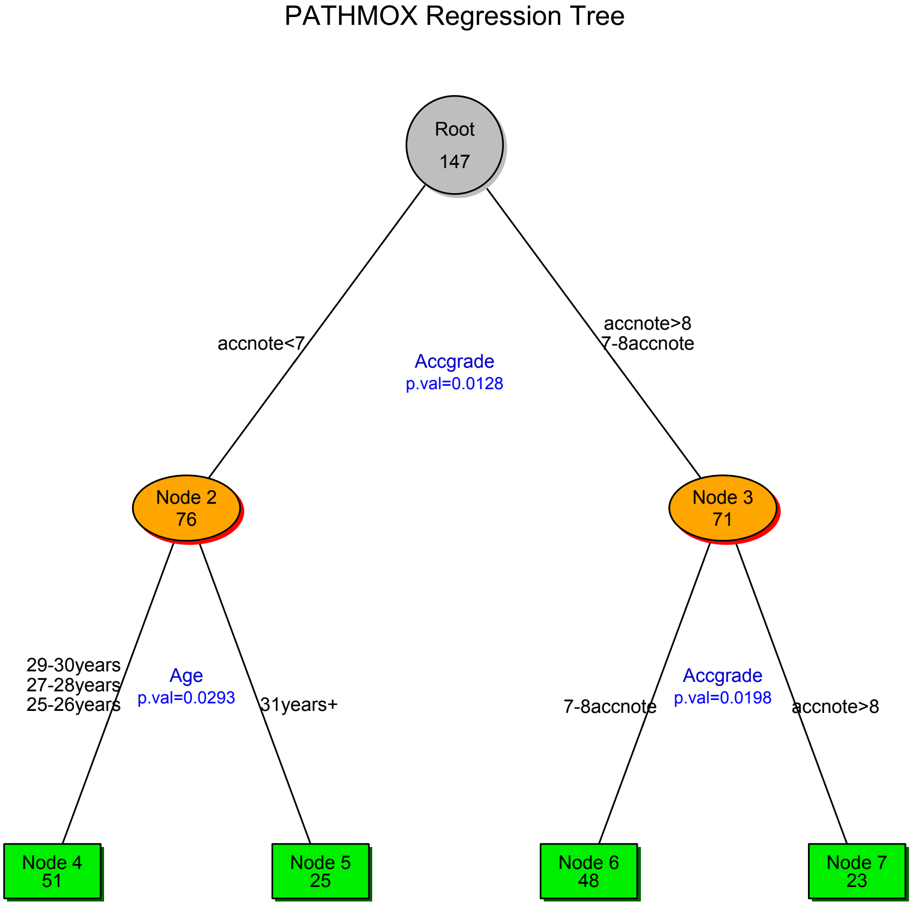


Figure 5.1: PATHMOX segmentation tree - least squares method

Table 5.5: F-global values and partitions - least squares method

Split	F_g statistic	F_g p-value	Variables	Mod G_1	Mod G_2
Split one	2.823	0.013	Access grade	accnote<7	7-8accnote/accnote>8
Split two	2.529	0.029	Age	25-26years/27-28years/29-30years	31years+
Split three	2.757	0.020	Access grade	7-8accnote	accnote>8

The last part of the analysis consists of the comparison between the terminal nodes of the tree. The coefficients and R^2 index calculated for each node are shown in Table (5.6), whereas the p-values associated to each coefficient are shown in the Table (5.7).

Table 5.6: Coefficients estimate & R^2 computed for each terminal node of the PATHMOX - least squares method

Coefficients & R^2 comparison					
Variables	**Root node**	**Node four**	**Node five**	**node six**	**Node seven**
Image	0.240	0.168	0.308	0.112	0.415
Generic Expectations	0.111	0.132	0.068	−0.011	0.302
Specific Expectations	−0.083	0.019	−0.602	−0.030	−0.216
Generic Quality	0.026	0.186	0.377	−0.005	−0.068
Specific Quality	0.123	0.059	0.031	0.174	0.375
Value	0.575	0.576	0.873	0.458	0.313
R^2	**0.688**	**0.759**	**0.767**	**0.742**	**0.812**

Table 5.7: P-value validation for the coefficients for each terminal node of PATHMOX

P-value coefficients comparison					
Variables	**Root node**	**Node four**	**Node five**	**Node six**	**Node seven**
Image	0.002	0.195	0.346	0.214	0.054
Generic Expectations	0.043	0.112	0.707	0.880	0.032
Specific Expectations	0.125	0.840	0.021	0.679	0.072
Generic Quality	0.710	0.163	0.139	0.956	0.654
Specific Quality	0.067	0.569	0.895	0.169	0.015
Value	0.000	0.000	0.002	0.000	0.064

In order to provide a visual way to appreciate the differences among segments, we have decided to show bar charts of the coefficients compared to the values obtained for the global model (see Figure (5.2)).

We can see that the node four (less or equal than 30 years,with *"low"* profile students), the coefficients also do no present a high degree of significance, which indicates Generic Quality as the most important effect in defining the Satisfaction. Node five (more than 30 years, with *"low"* profile students) indicates Value as the most significant effect. The model of students with a*"good"* profile (node six) is similar to the global model. Node seven (students with the *"best"* profile) is defined mainly by Image, Generic Expectations, and Specific quality. All the R^2 are highly significant. In particular node five and seven present an R^2 index, respectively, of *0.767* and *0.812*. One important consideration concerns the p-values of the coefficients of each terminal nodes: we can see, for example, that Value is significant in all groups. Also, if the coefficients are highly distinct in respect to the global model, they are sometimes not significant: for example node five has Generic Quality distinct in respect to the global model but this is not significant. Thus, we have to take care when we interpret the models of the terminal nodes.

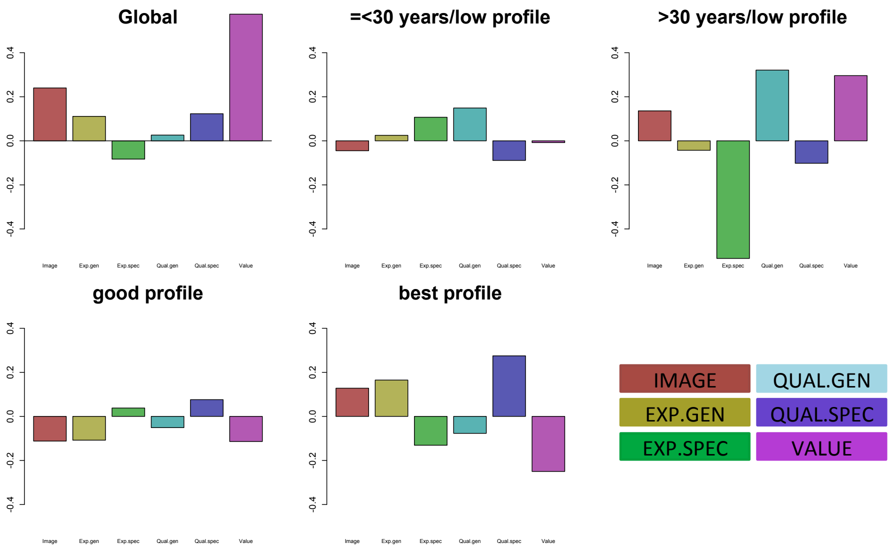


Figure 5.2: Effects on Satisfaction: comparison of the four segments with respect to the global model - least squares method

5.2.3 Extending the PATHMOX Approach for Detecting which Constructs Differentiate Segments

As we can see, the PATHMOX approach allows us to detect the existence of different models in a dataset without identifying segmentation variables beforehand: the different segments are revealed as branches of the segmentation tree. However, the F-test used in PATHMOX as split criterion is a global criterion: it allows us to assess whether all coefficients for two compared models are equal or not, but it does not indicate which particular coefficients are responsible for the difference. For instance, when PATHMOX detects a difference between the two groups: *good profile* students and *best profile* students, we don't know which ones of the coefficients are responsible for the detected difference. To identify the responsible coefficients of the split, we have introduced the F-coefficient test. Table (5.8) contains the statistic and the p-value of each coefficient for the three splits identified by PATHMOX through of the F-global criterion.

Table 5.8: F-coefficients values from least squares estimation: identification of the coefficients responsible for the partition

F-coefficient	**Split one: access grade**			**Split two: age**			**Split three: access grade**		
Variables	**Stat.**	**P-value**	**Signific.**	**Stat.**	**P-value**	**Signific.**	**Stat.**	**P-value**	**Signific.**
Image	0.493	0.484	NO	0.254	0.616	NO	2.411	0.126	NO
Generic Expectations	0.611	0.436	NO	0.159	0.691	NO	5.542	0.022	YES
Specific Expectations	0.000	0.996	NO	9.214	0.003	YES	2.337	0.132	NO
Generic Quality	2.175	0.143	NO	0.650	0.423	NO	0.157	0.694	NO
Specific Quality	4.574	0.034	YES	0.018	0.894	NO	1.311	0.257	NO
Value	5.471	0.021	YES	2.293	0.135	NO	0.879	0.352	NO

Starting from the split of the root node (split one), we can see that the significant difference identified by PATHMOX between students with a *"low"* profile and students with a *"high"* profile, depends on the coefficient Value (F-coefficient statistic: *5.471* and associated p-value: *0.021*) and Specific Quality (F-coefficient statistic: *4.574* and associated p-value: *0.034*). The partition of node two (split two) in the "≤ 30" and "> 30" students with *"low"* profile is imputed to Specific Expectations with an F-coefficient statistic of

9.214 and an associated p-value of *0.003*. In the last partition (split three), students with a *“good"* and *“best"* profile, the difference is imputed to Generic Expectations with an F-coefficient statistic of *5.542* and an associated p-value of *0.022*.

5.3 PATHMOX Approach Applied to the Context of LAD Robust Regression

The least squares linear regression estimator is well-known to be highly sensitive to the outliers. To overcome this drawback, we propose the least-sum of absolute deviations (LAD) regression (Koenker and Bassett (1978)) which is robust to the presence of outliers and proves to be the most efficient of least squares, in cases where the error term does not have a normal distribution. As in the least squares regression, we will use the example based on the evaluation of Alumni Satisfaction: we want to explain the satisfaction from its drivers, Image of the school, Expectations on general skills, Expectations on technical skills, Perceived Quality on general skills, Perceived Quality on technical skills and Value or profit obtained after graduation.

5.3.1 Global LAD Robust Regression Model

We start the analysis with the global model estimated for all students. In Table (5.9), we show, from left to right the coefficient estimate and an interval of confidence at 95% for each of them, which is useful for a validation purposes since, in this case, we don’t have any statistic to evaluate the coefficients. As we can see, that main drivers of Satisfaction are Value with a coefficient of *0.528*, and Image, with a coefficient of *0.142*. The other coefficients are not significant.

Table 5.9: Coefficient estimate and statistics from global LAD satisfaction model

LAD regression	**Coefficients**	**Lower bd**	**Upper bd**
Image	0.142	0.042	0.352
Generic Expectations	0.167	−0.059	0.299
Specific Expectations	−0.081	−0.238	0.088
Generic Quality	−0.002	−0.170	0.134
Specific Quality	0.196	−0.030	0.282
Value	0.528	0.434	0.684

5.3.2 LAD Robust Regression PATHMOX Approach

In order to calculate PATHMOX, we have to specify again the scale of the segmentation variables (see Table (5.2)). We choose 0.1 for the threshold of the p-value due to the conservative nature of the F-statistic calculated by the LAD approximation. We maintain the same stop conditions of the algorithm in the least squares example; these are: 15 students (10% of total sample) as the minimum number to stop the growth of a node and the depth level = 2.

In Figure (5.3), we present the obtained tree in which we can observe three distinct models. The F-global statistics, the p-value and the obtained partitions for each node are summarized in Table (5.10). At the first split, PATHMOX defines two different models one for students with a *“low"* profile, and the other, for students with a *“higher"* profile as in the least squares example; we see that the produced split is highly significant, giving an F-statistic of *2.008* with a p-value of *0.069*. The tree continues by splitting node two (students with *“low"* profile) and three (students with a *“higher"* profile). The most significant split for the node two is obtained by the variable *age*, giving an F-statistic of *1.907* with a p-value of *0.093*, obtaining the child node (node four) of the students with less than 30 years ($\leq$ 30), and child node (node five), with students of more than 30 years ($>$ 30). Node three is taken to be a final node, due to the fact that no significant partitions have been found. Hence, in the end, we have three final segments each corresponding to a distinct model: **node three** model of *"higher"* profile students; **node four** model of less or equal than to 30 years, with *"low"* profile students; **node five** model of more then 30 years, with *"low"* profile students.

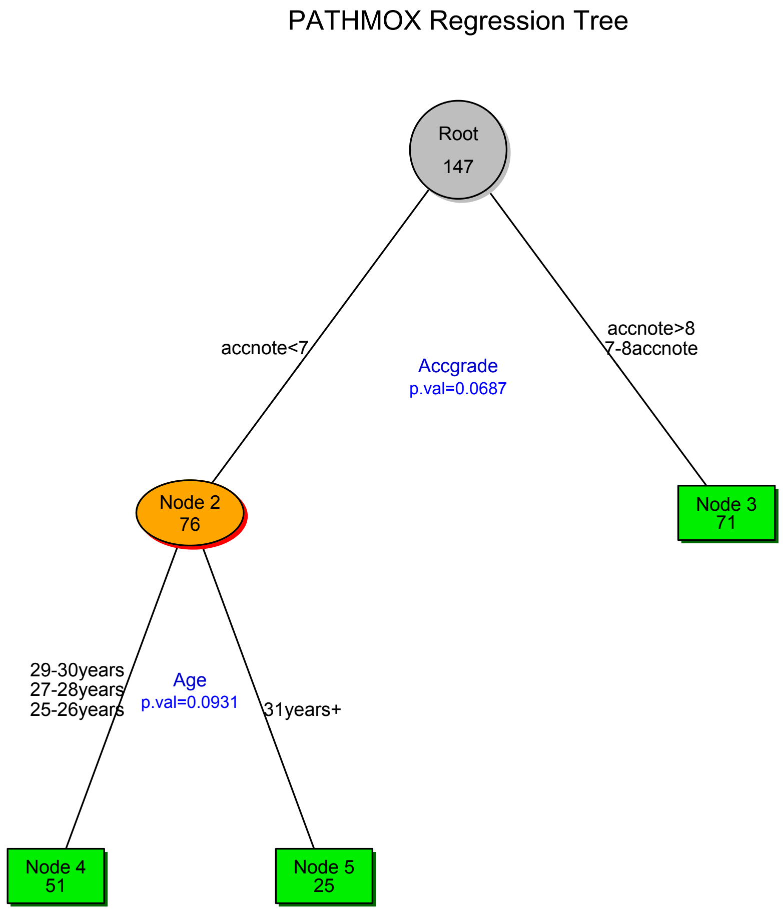


Figure 5.3: PATHMOX segmentation tree - LAD method

Table 5.10: F-global values and partitions - LAD method

Split	F_g **statistic**	F_g**pvalue**	**Variables**	**Mod** G_1	**Mod** G_2
Split one	2.008	0.069	Access grade	accnote<7	7-8accnote/accnote>8
Split two	1.907	0.093	Age	25-26years/27-28years/29-30years	31years+

The last part of the analysis consists of the comparison between the terminal nodes of the tree. The coefficients of each node are shown in Table (5.11). In order to provide a visual way to appreciate the differences among segments, we use again the bar charts to compare the coefficients with the values obtained for the global model (see Figure (5.4)).

Table 5.11: Coefficients estimate computed for each terminal node of the PATHMOX - LAD method

	Coefficients comparison			
Variables	**Root node**	**Node three**	**Node four**	**Node five**
Image	0.142	0.118	0.120	0.928
Generic Expectations	0.167	0.064	0.054	0.007
Specific Expectations	−0.081	−0.132	0.193	−0.407
Generic Quality	−0.002	−0.006	0.140	0.175
Specific Quality	0.196	0.286	0.018	0.007
Value	0.528	0.437	0.422	0.591

We can see that node three, model of students with *"higher"* profile, indicates that the most important effects are Value, and Specific quality. node four, model of less or equal than 30 years, among *"low"* profile students, indicates that Specific Expectations and Generic Quality are the most important effects. The model of more than 30 year, among *"low"* profile students (node five) is characterized, above all by the coefficient Image.

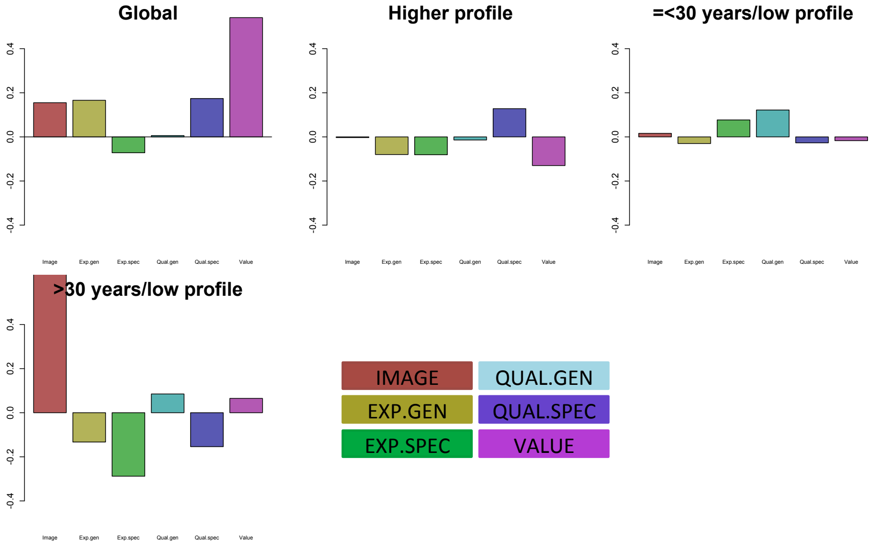


Figure 5.4: Effects on Satisfaction: comparison of the four segments with respect to the global model - LAD method

5.3.3 Extending the PATHMOX Approach for Detecting which Constructs Differentiate Segments: Robust Regression

As in the case of least squares linear regression, we can investigate which are the coefficients responsible for the difference by means of the F-coefficient statistic. We show the F-statistic as the associated p-value for each variable in Table (5.12). Starting from the split of the root node (split one), we can see that the significant difference identified by PATHMOX between the *“low"* and *“higher"* profile depends on the coefficient Value (F-coefficient statistic: *3.305* and associated p-value: *0.071*). In the case of students with less or equal than 30 and more than 30 years with a *“low"* profile (split 2), the difference depends above all, on the coefficients Specific Expectations (F-coefficient statistic: *4.393* and associated p-value: *0.040*) and Image (F-coefficient statistic: *3.089* and associated p-value: *0.083*).

Table 5.12: F-coefficients values from LAD estimation: identification of the coefficients responsible for the partition

F-coefficient	**Split one: access grade**			**Split two: age**		
Variables	**Stat.**	**P-value**	**Signific.**	**Stat.**	**P-value**	**Signific.**
Image	0.286	0.594	NO	3.089	0.084	YES
Generic Expectations	0.558	0.456	NO	0.056	0.814	NO
Specific Expectations	0.065	0.799	NO	4.393	0.040	YES
Generic Quality	0.118	0.732	NO	0.034	0.854	NO
Specific Quality	1.992	0.160	NO	0.003	0.958	NO
Value	3.305	0.071	YES	0.846	0.361	NO

5.4 Least Squares vs LAD PATHMOX

Least squares and LAD approaches can be compared with the segmentation tree developed by PATHMOX. Observing the two trees shown in Figure (5.1) and (5.3), we can see that the first split is the same. The root node is split by the variable *acc-grade* and produces the same partitions: students with a *“low"* profile (acc-note < 7) and students with a *“higher"* profile (acc-note ≥ 7). Conversely, when PATHMOX arrives to the second depth level of the tree, there is a difference in the partition of node three: with the least squares estimation, the node three is split again by variable *acc-grade*. Whereas, with the LAD estimation the node three is taken as a terminal node.

We can justify the difference considering the diversity the two statistics: the F-statistic obtained by least squares finds a significant difference, whereas the more robust F-statistic computed by LAD does not.

5.5 PATHMOX Approach Applied to Logistic Regression

As an example of generalized linear model, we have chosen logistic regression. To use the same data on the evaluation of alumni satisfaction, we have modified the dependent variable Satisfaction from continuous to binary (*0*: not satisfied, *1*: satisfied). As threshold for discriminating if a student was satisfied or not, we have utilized the value of the median: *0.16* of variable Satisfaction. Thus, we obtain *75* no satisfied students and *72* satisfied students. Again, we just consider a simple additive model: we want to explain the Satisfaction from its drivers, Image of the school, Expectations on general skills, Expectations on technical skills, Perceived Quality on general skills, Perceived Quality on technical skills and Value or profit obtained after graduation.

5.5.1 Global Logistic Regression Model

We start the analysis with the global model. In Table (5.13) we show, from left to right the coefficient's estimate, the standard deviation, the value of the statistic z and the p-value. As we can see, the main drivers of Satisfaction, in this case are also Value, with a coefficient of $\exp\beta = 6.652$ and a p-value of *zero*, and Image with a coefficient of $\exp\beta = 2.611$ and a p-value of *0.018*. The model present a null deviance of *203.72* and a residual deviance of *127.81*.

Table 5.13: Coefficient estimate and statistics of the global logistic satisfaction model

GLM regression	$\exp\beta$	**Std. error**	**z value**	**Pr(>$\|z\|$)**
Image	2.337	0.405	2.368	0.018
Generic Expectations	0.986	0.259	-0.191	0.848
Specific Expectations	0.892	0.261	-0.553	0.580
Generic Quality	1.315	0.331	0.817	0.414
Specific Quality	1.088	0.343	0.328	0.743
Value	5.344	0.414	4.571	0.000

5.5.2 Logistic Regression PATHMOX Approach

In order to calculate the PATHMOX, we have to specify the scale of the segmentation variables (see Tables (5.2)) and the stop conditions parameters of the algorithm. We have decided to establish a value of 0.05 for the threshold of the p-value to look for those partitions that are highly significant. Given that we have a total sample of 147 students and to avoid small segments that are no realistic in the context of the logistic regression, we have chosen a number of 28 students (20% of total sample) to stop the growth of a node. Also in this case, the depth level is $= 2$, with the aim of obtaining a simple segmentation tree with a possible maximum number of four final segments.

In Figure (5.5), we present the obtained tree. The Likelihood ratio (LRT) statistic, the p-value and the obtained partitions for each node are summarized in Table (5.14). As we can see, PATHMOX identifies two different models according to the variable *salary*: one for students with a *“low - middle"* salary (≤ 35) and another for students with a *“high"* salary (> 40). We see that the produced split is highly significant, giving a *LRT*-statistic of *16.546* with a p-value of *0.020*.

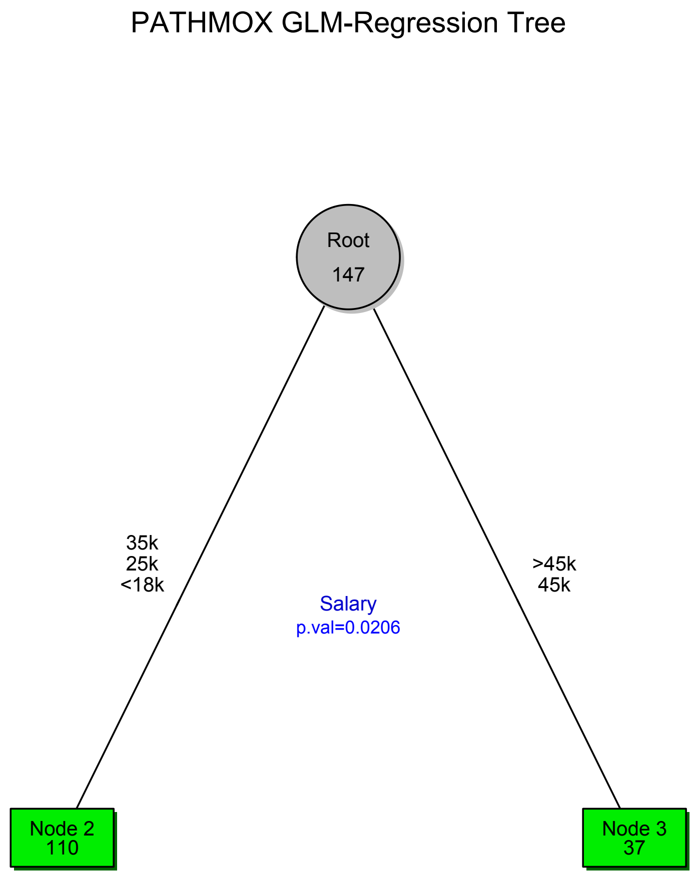


Figure 5.5: PATHMOX segmentation tree - GLM method

Table 5.14: Likelihood test values and partitions

Split	*LRT* **statistic**	*LRT* **p-value**	**Variables**	**Mod** G_1	**Mod** G_2
Split one	16.546	0.020	Salary	<18k/25k/35k	45k/>45k

The last part of the analysis consists of the comparison between the terminal nodes of the tree. The coefficients of each node, as the Akaike's criterion (AIC) are shown in Table (5.15), whereas the p-values are shown in Table (5.16). The bar charts are presented in Figure (5.6).

Table 5.15: Coefficients estimation & AIC computed for each terminal node of the PATHMOX - GLM method

Coefficients comparison			
Variables	**Root node**	**Node two**	**Node three**
Image	2.337	2.373	1.659
Generic Expectations	0.986	1.284	0.331
Specific Expectations	0.892	0.746	1.221
Generic Quality	1.315	1.168	4.914
Specific Quality	1.088	0.823	2.387
Value	5.344	6.746	9.994
AIC	**141.805**	**107.124**	**32.135**

Table 5.16: P-value validation for the coefficients for each terminal node of PATHMOX - GLM method

P-value comparison			
Variables	**Root node**	**Node two**	**Node three**
Image	0.018	0.034	0.657
Generic Expectations	0.848	0.605	0.039
Specific Expectations	0.580	0.379	0.165
Generic Quality	0.414	0.558	0.055
Specific Quality	0.743	0.589	0.640
Value	0.000	0.000	0.082

We can see that node two (students with a salary of less than 40) is similar to the global model. Conversely, node three indicates that the most important effect is Generic Quality. Both models, as in the case of the global one, indicate that the coefficient Value is highly significant. The AIC improves in both models with respect to the global, above all in node three, where the AIC decreases down to *32.135*.

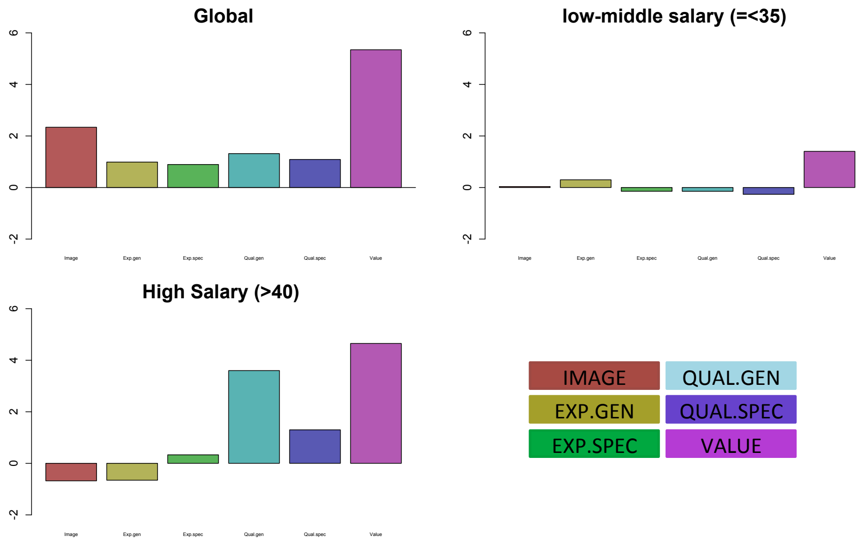


Figure 5.6: Effects on Satisfaction: comparison of the four segments with respect to the global model - GLM method

5.5.3 Extending the PATHMOX Approach to Detect which Constructs Differentiate Segments: Logistic Regression

Also in the case of logistic regression, we can investigate which are the coefficients responsible for the difference by means of the *LRT*-coefficient statistic. We show the *LRT*-statistic as the associated p-value for each variable in Table (5.17). Observing the table, we can see that the significant difference identified by PATHMOX between students with a *"low - middle"* salary (≤ 35) and students with a *"high"* salary (> 40) depends on the coefficients Generic Expectations (*LRT*-coefficient statistic: *15.786* and associated p-value: *0.046*) and Generic Quality (*LRT*-coefficient statistic: *13.898* and associated p-value: *0.084*)

Table 5.17: *LRT*-coefficients values: identification of the coefficients responsible for the partition

***LRT*-coefficient**	**Split one: salary**		
Variables	**Stat.**	**P-value**	**Signific.**
Image	3.829	0.872	NO
Generic Expectations	15.786	0.046	YES
Specific Expectations	6.739	0.565	NO
Generic Quality	13.898	0.084	NO
Specific Quality	0.892	0.999	NO
Value	0.054	1.000	NO

5.6 PATHMOX Approach Applied to Principal Component Analysis

In the context of Principal component Analysis (PCA), it makes sense to apply PATHMOX when we are interested in discovering different patterns of correlation between the variables of interest that define the principal components. As in the previous application, we will use the example based on the evaluation of alumni Satisfaction to show how PATHMOX works.

5.6.1 Global PCA Model

Let us start with the analysis of the global model. We have decided to consider two components that allow us to describe 65% of the variance in the swarm of points, which, we consider in this context to be an adequate value for good description power of the model. In Figure (5.7) we provide the correlation circle plot that is useful for investigating the relation between the original variable and the two principal components. In addition, we calculate the correlation, the squares cosine and the contribution values (see Table (5.18)). The correlation is the correlation value between the original value and the component. The squares cosine is the cosine of the angle formed by the variable and the component (the bigger it is, the higher the relation between component and variable is). The contribution is how much a variable contributes to explaining a component.

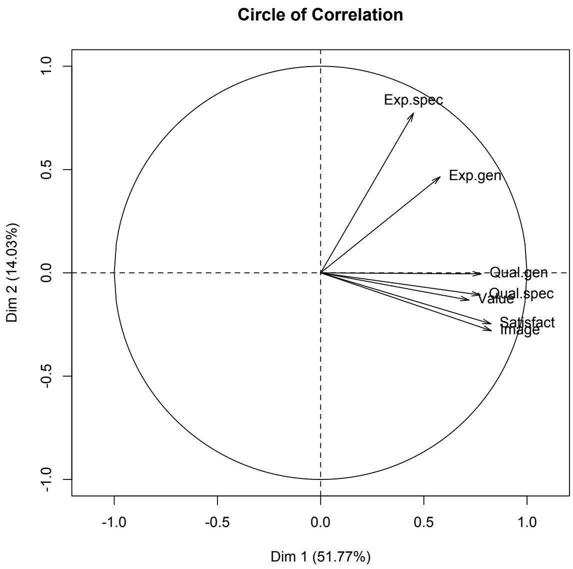


Figure 5.7: Correlation circle: global model

Table 5.18: Analysis of R^n space: coordinates, correlation, cos^2, contribution - component 1 & 2

Summary R^n space						
Variable	**Correlation**		**cos^2**		**Contribution**	
	Comp$_1$	**Comp$_2$**	**Comp$_1$**	**Comp$_2$**	**Comp$_1$**	**Comp$_2$**
Image	0.828	−0.280	0.685	0.079	18.907	8.009
Generic Expectations	0.579	0.465	0.336	0.216	9.260	22.006
Specific Expectations	0.450	0.773	0.203	0.598	5.588	60.847
Generic Quality	0.778	−0.005	0.605	0.000	16.692	0.003
Specific Quality	0.773	−0.107	0.598	0.011	16.501	1.165
Value	0.719	−0.132	0.516	0.017	14.249	1.767
Satisfaction	0.825	−0.247	0.681	0.061	18.802	6.204

As we can see, the first component is highly correlated, above all, with Image (*0.828*) and Satisfaction (*0.825*). Generic and Specific Quality follow, with a correlation value of *0.778* and *0.773*, then the variable

Value (*0.719*). Conversely, the second component, is highly related to Specific Expectations, with a correlation of *0.773*, and to Generic Expectations (*0.465*). The squares cosine and the contribution value confirm that Image and Satisfaction are the two most important variables for explaining the first component, whereas the second component is influenced only by Expectations. Thus, we can conclude that the first component can be defined as a *"global opinion"* of the students, and the second component as their *"expectations"*.

As we are interested only in analyzing different patterns of relationships among variables, we conclude here our analysis without considering the student projections on the components.

5.6.2 PATHMOX PCA

In order to calculate the PATHMOX, we have to specify the scale of the segmentation variables (Table (5.2)) and the stop conditions parameters of the algorithm. Due to the fact that in the PCA-PATHMOX we utilized a no-parametric criterion, we don't have a reference distribution and an associated p-value. Thus, to realize a split, we just take into consideration the average correlation value. We have chosen a value of 0.6 for the threshold of the statistic. We have fixed the limit of the tree depth = 1 to avoid a high number of terminal nodes, which are difficult to interpret in the context of PCA analysis. The limit of node size is the same as in the multiple linear regression application (15 students: 10% of total sample).

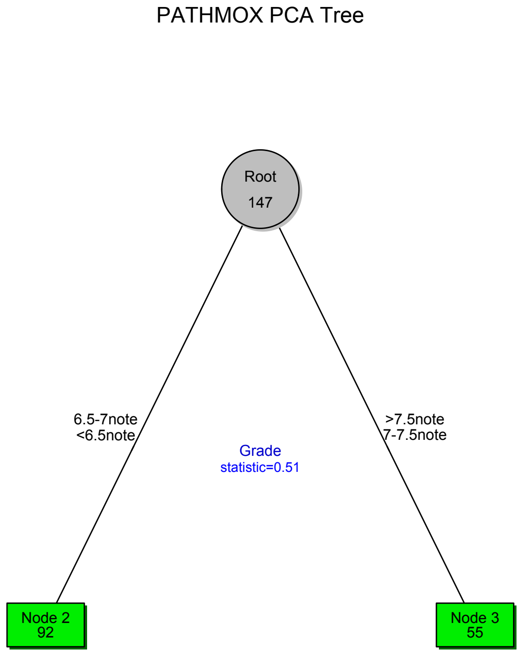


Figure 5.8: PCA PATHMOX segmentation tree - average correlation method

Table 5.19: Average correlation values and partitions

Split	Average correlation	Variables	Mod G_1	Mod G_2
Split one	0.510	Grade	<6.5note/6.5-7note	7-7.5note/>7.5note

In Figure (5.8), we present the obtained tree. The average correlation value and obtained partitions for each node are summarized in Table (5.19).

As we can see, PATHMOX identifies two different models according to the variable *grade*: one for students with a *"middle" grade* (grade ≤ 7), and another, for the *"good"* students (grade > 7). We see that the produced split is highly significant, giving a average correlation value of *0.510*.

As in the other applications, the last step is to analyze the differences among the terminal nodes. The correlation values, the squares cosine, and the contribution values are shown in Tables (5.20), (5.21) and (5.22).

Table 5.20: Correlation between component and variables, root and terminal nodes

Correlation node comparison						
		$\mathbf{Dim_1}$			$\mathbf{Dim_2}$	
Variables	**Root node**	**Node two**	**Node three**	**Root node**	**Node two**	**Node three**
Image	0.828	0.841	0.803	−0.280	−0.290	−0.387
Generic Expectations	0.579	0.549	0.656	0.465	0.614	−0.084
Specific Expectations	0.450	0.556	0.276	0.773	0.657	0.078
Generic Quality	0.778	0.817	0.691	−0.005	−0.053	−0.539
Specific Quality	0.773	0.723	0.846	−0.107	−0.179	−0.063
Value	0.719	0.750	0.669	−0.132	−0.122	0.658
Satisfaction	0.825	0.818	0.830	−0.247	−0.238	0.397

Table 5.21: cos^2 values, root and terminal nodes

Cos^2 node comparison						
		$\mathbf{Dim_1}$			$\mathbf{Dim_2}$	
Variables	**Root node**	**Node two**	**Node three**	**Root node**	**Node two**	**Node three**
Image	0.685	0.707	0.644	0.079	0.084	0.150
Generic Expectations	0.336	0.301	0.430	0.216	0.377	0.007
Specific Expectations	0.203	0.310	0.076	0.598	0.432	0.006
Generic Quality	0.605	0.668	0.478	0.000	0.003	0.291
Specific Quality	0.598	0.523	0.716	0.011	0.032	0.004
Value	0.516	0.562	0.447	0.017	0.015	0.433
Satisfaction	0.681	0.669	0.689	0.061	0.057	0.158

Table 5.22: Contribution values, root and terminal nodes

Contribution node comparison						
		$\textbf{Dim}_1$			$\textbf{Dim}_2$	
Variable	**Root node**	**Node two**	**Node three**	**Root node**	**Node two**	**Node three**
Image	18.907	18.904	18.505	8.009	8.437	14.285
Generic Expectations	9.260	8.054	12.347	22.006	37.680	0.670
Specific Expectations	5.588	8.276	2.191	60.847	43.254	0.579
Generic Quality	16.692	17.861	13.724	0.003	0.280	27.716
Specific Quality	16.501	13.985	20.582	1.165	3.190	0.380
Value	14.249	15.035	12.850	1.767	1.481	41.307
Satisfaction	18.802	17.885	19.801	6.204	5.678	15.061

In order to provide a visual way to appreciate the differences among segments, we have decided to show the correlation circles of the terminal nodes in comparison to those obtained for the global model (see Figure (5.9)).

Observing the graphic, we can note that node two, the model of students with a *“middle"* grade (≤ 7), is similar to the global node. We have all variables highly related to the first component (correlation with the first component: *satisfaction (0.818)*, Image (*0.841*), Value (*0.750*); Generic and Specific Expectations are highly correlated with the second component (correlation value with the second component: *0.614* and *0.657*). In the case of the model of *“good"* students (node three), we note now a contraposition between Value and Satisfaction on the one hand, and Image and Generic Quality on the other. To summarize, we can say that the students with a *“middle"* note (≤ 7) can differentiate just the Expectations by the other variables that are perceived as one latent construct: *“global opinion"*. The *“good"* students differentiate own opinion, identifying the Satisfaction and Value on the one hand, and Image and Generic Quality on the other.

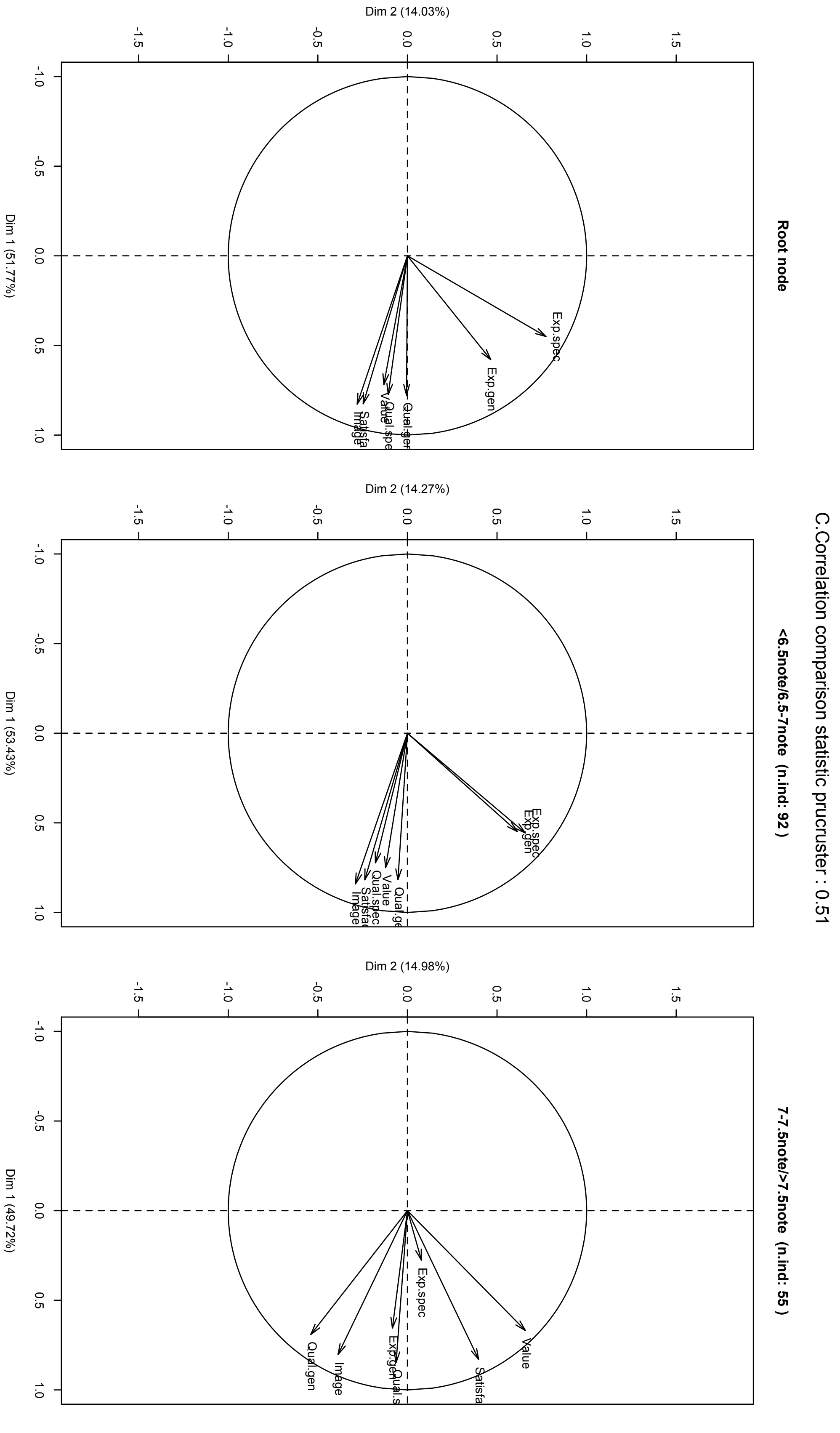


Figure 5.9: Correlation circles root and terminal nodes

5.7 Partial Least Squares PATHMOX Application

Traditionally, PLS models assume the homogeneity of all observations: it is supposed that a single model is able to represent all individuals. This assumption is often not met in real-data studies. In many cases, it is reasonable to think that different groups of people, characterized by heterogeneous behavior, are observable in data. The trouble of not considering the possible existence of segments among observations is that the conventional PLS models may lead the analyst to inadequate and inaccurate results. To solve these problems, it is necessary to assume the existence of different groups among the observations, which implies that more than one parameter set is necessary to describe the data. In this context, PATHMOX (Sánchez, 2009) represent a powerful tool for analyzing the heterogeneity. In this section we use the alumni satisfaction data-set to apply PATHMOX to partial least squares path modeling. The F-block, the F-coefficients and the invariance measurement test also have also been included, to offer a more complete analysis.

5.7.1 PLS-PM and Student Satisfaction

In accordance to Kerlin (2000), the research on student satisfaction was developed according to three different perspectives. One of the approaches considers the student as an employee. This approach measures the educational and career value for the time and effort spent, as well as the relationship of the student with the institution. Another approach is that considers the student as a person fully integrated in his or her environment (person - environment fit). In this case, what is measured is the congruence of the student's values with the institution's values. If incongruence occurs, the student will probably leave. The third approach is the one that views the student as a consumer or client. Following this last approach:“student-as-customer", student satisfaction has been approached as a customer satisfaction problem. In this context, PLS-PM has became a reference methodology for analyzing the students' preference. Martensen *et al.* (2000) and Eskildsen *et al.* (1999) did the initial work in adapting the ECSI concept to measuring student satisfaction and loyalty. Taking these works as a reference, almost all other applications were proposed with the aim of evaluating the most important drivers of student satisfaction, by adapting the customer satisfaction index (see Brown *et al.* (2009); Serenko (2010); Temizer *et al.* (2010)). Trinchera and Balzano (2011), using PLS-PM methodology, proposed a model in which four variables where analyzed: course organization, teaching, facilities, and interest and satisfaction; the study highlighted the importance of teaching quality in defining the student satisfaction. Grace *et al.* (2012) used the PLS-PM framework and the Course Experience Questionnaire (CEQ) to analyzed the student satisfaction of 164 Australian students attending one third-year course of a Bachelor degree in business; Duarte *et al.* (2012) also analyze the student satisfaction in Portugal Universities by PLS-PM, highlighting the importance of a double measure: the first during the students' courses and the second after some time had passed since they had finished.

5.7.2 Manifest and Latent Variables Relationship

In our model, we consider two main constructs: Value and Satisfaction. As we show in Figure (5.10), we have considered Image, Generic Quality and Specific Quality as antecedents of Value and Image, Generic Quality and Specific Quality and Value as antecedents of Satisfaction. All latent variables are reflective.

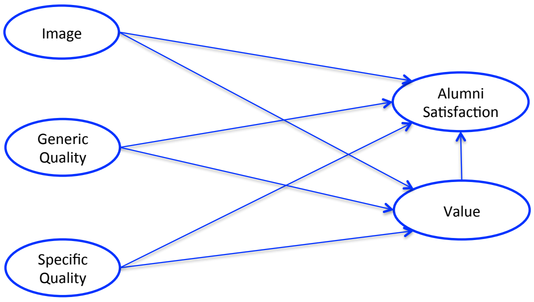


Figure 5.10: Path diagram of Alumni satisfaction latent variables relationship

5.8 PLS-PM Main Results

We present the main results of the fitted PLS-PM model, just to give an idea about the worth of the estimated model. As indicated by (Chin, 2010) we reported as brief summary of the main results:

1. Unidimensionality of the reflective latent constructs (Table (5.23)). All Dillon-Goldstein ρ are higher than *0.80*. Furthermore the sizes of the first and second eigenvalues confirm that every construct is unidimensional. Hence we can consider every composite factor as expression of a single concept;

Table 5.23: Different measures to assess unidimensionality of blocks of indicators

Unidimensionality blocks of indicators					
Constructs	**Indicators**	**Crombach's α**	**Dillon ρ**	**1-Eigenvalue**	**2-Eigenvalue**
Image	8	0.916	0.932	5.046	0.726
Specific Quality	3	0.739	0.854	1.993	0.727
Generic Quality	4	0.785	0.862	2.439	0.756
Value	4	0.788	0.864	2.455	0.650
Satisfaction	4	0.773	0.857	2.422	0.773

2. Reliability, measured by the average variance extracted by each construct respect of its indicators (Table (5.24)). In all cases is greater than *0.5*, meaning that each construct is measured by a set of coherent indicators;

Constructs	**Type**	**Measure type**	**Number of indicators**	**R.square**	**Av. communality**	**Av. redundance**	**AVE**
Image	Exogen	Rflct	8	0.000	0.629	0.000	0.629
Specific Quality	Exogen	Rflct	3	0.000	0.662	0.000	0.662
Generic Quality	Exogen	Rflct	4	0.000	0.608	0.000	0.608
Value	Endogen	Rflct	4	0.232	0.612	0.142	0.612
Satisfaction	Endogen	Rflct	4	0.653	0.584	0.382	0.584

Table 5.24: Summary results per blocks

3. The estimated inner model (Table (5.11) and Figure (5.25)). We can see that the main drivers of Satisfaction are Value with a path coefficient of *0.566*, and Image with a path coefficient of *0.299* whereas, in the case of Value, the most important effect is, above all, Image with a path coefficient of

0.363, followed by Specific Quality (path coefficient: *0.199*). The Generic Quality presents very low path coefficients either when it is related with Value (path coefficient: *-0.048*) and with Satisfaction (path coefficient: *-0.014*). In any case, we have decide to keep it in the model, because we suspect that it could become significant when we will take in account the heterogeneity.

Table 5.25: Results of the structural model (R^2 and path coefficients)

Results of the structural model (R^2 and path coefficients)			
Value		**Satisfaction**	
R2	0.232	R2	0.653
Intercept	2.086	Intercept	0.683
Image on Value	0.363	Image on Satisfaction	0.299
Specific Quality on Value	0.199	Specific Quality on Satisfaction	0.106
Generic Quality on Value	-0.048	Generic Quality on Satisfaction	-0.014
		Value on Satisfaction	0.566

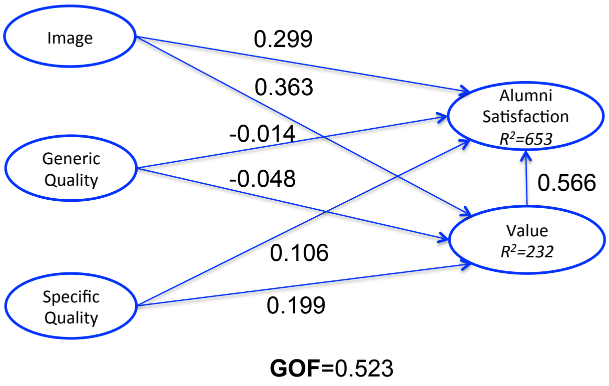


Figure 5.11: Path diagram of the students Satisfaction model

4. Significance of the path coefficients (Table (5.38)). We assess it by looking to the bootstrap confidence intervals of path coefficients , where all of them are significant except Specific Quality on Satisfaction, Generic Quality on Value and Generic Quality on Satisfaction.

Paths	**Original path value**	**Bootstrap mean**	**Bootstrap stand. error**	**Bootstrap perc. 0.025**	**Bootstrap perc. 0.975**
Image on Value	0.363	0.356	0.118	0.099	0.563
Image on Satisfaction	0.299	0.295	0.072	0.159	0.423
Specific Quality on Value	0.199	0.216	0.108	0.037	0.427
Specific Quality on Satisfaction	0.106	0.110	0.071	-0.025	0.256
Generic Quality on Value	-0.047	-0.038	0.104	-0.269	0.152
Generic Quality on Satisfaction	-0.014	-0.006	0.076	-0.175	0.116
Value on Satisfaction	0.566	0.558	0.063	0.448	0.668

Table 5.26: Bootstrap confidence intervals for path coefficients (*0.025* and *0.975* percentiles)

5. Predictability of the model (Table (5.11)) with a R^2 of *0.653* and *0.232* for the Satisfaction and Value

constructs respectively. The R^2 of Value is clearly too low for predictability purposes, however we consider it for the sake of comparison between potential subgroups of alumni.

5.8.1 PATHMOX PLS

Despite of the reduced sample size of the study, it seems that the obtained model is good enough to take conclusions about how Satisfaction and Value are formed in the alumni. However, the question is whether the effects we have detected are valid for the whole population or they are in fact artificial averages of underlying subpopulations. To answer to that question and to discover potential sources of heterogeneity we perform a PATHMOX analysis employing all available segmentation variables of (Table (5.2)). Necessarily, the tree we are interested in has to be small, since the sample size is not big at all and because, in model segmentation we are not looking for big trees with a large number of leaves, prone to incur in overfitting, instead we are interested in a moderate final number of segments which can be interpreted and made operational for management purposes. In that case we have limited the tree to a maximum depth of two levels (limiting to four at most the final number of segments). We have chosen a number of fifteen alumni (10% of total sample) as the minimum admissible size for nodes and a threshold significance for the split criterion set at 0.05. The obtained PATHMOX tree is shown in Figure (7.3). It is a tree with a total number of three nodes. The root node corresponds to the previous global model calculated over the entire sample of one hundred forty-seven alumni.

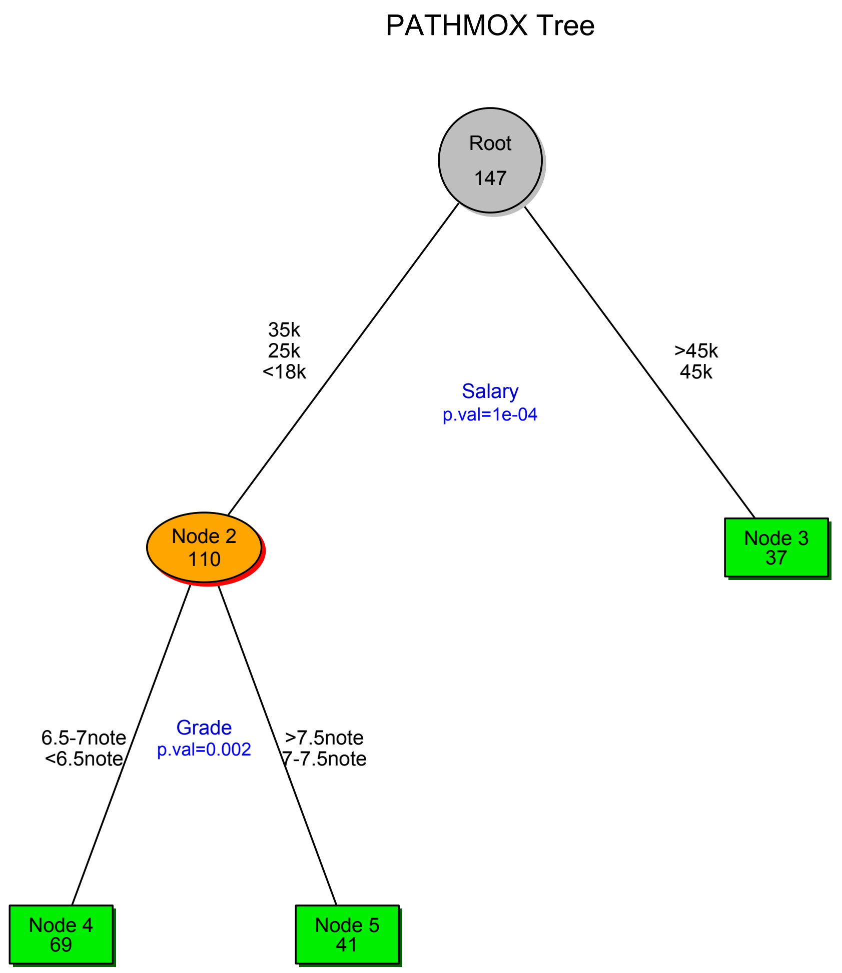


Figure 5.12: PLS PATHMOX segmentation tree - LS estimation

Table 5.27: F-global statistics values and partitions

Split	F_G **statistic**	F_G **pvalue**	**Variables**	**Mod** G_1	**Mod** G_2
Split one	4.056	0.000	Salary	<18k/25k/35k	45k/>45k
Split two	3.034	0.002	Grade	<6.5note/6.5-7note	7-7.5note/>7.5note

At the first split, PATHMOX defines two different models one for alumni with a lower salary ($\leq 35K$), and the other, for alumni with a higher salary ($> 35K$); we see that the produced split is highly significant, giving a F-statistic of *4.056* with a p-value of *zero*. The tree continues by splitting node two (alumni with lower salary); in this case the most significant split is obtained by the variable grade, giving a F-statistic of *3.034* with a p-value of *0.002*, obtaining two child nodes, one of the alumni with a "low" grade (≤ 7) and another with a "good" grade (> 7); the node three is taken as final node due to no significant partitions have been found. Hence, at the end, we have three final segments each corresponding to a distinct model: **node three** model of alumni with "high salary"; **node four** model of alumni with "regular salary and low grade"; **node five** model of alumni with "regular salary and good grade".

5.8.2 Extending the PATHMOX Approach: F-block and F-coefficient

The obtained tree allows us to identify which splits maximizes the difference between the PLS-PM models in the child nodes, for example, splitting the root node by salary ($> 35K$), we identify the two most different models. However, we don't know which of the equations forming the PLS-PM model (two in that case, one for satisfaction and the other for value) are really different; and in that case, which are the coefficients of the latent predictors in every equation responsible of the detected difference. To that purpose PATHMOX goes deeper in the split process by computing two statistics: F-block and the F-coefficient. The F-block allows us the detect which are the equations responsible of the split; the F-coefficient statistic allows us to identify the responsible predictors of the obtained partitions. The results of these tests as applied to our data are given in the Table (5.28) and Table (5.29).

5.8.2.1 F-block Results

Starting from the split of the root node (node one), PATHMOX identifies both constructs Value and Satisfaction as the ones responsible for the difference of the inner models: the F-block presents, in the case of value, a p-value of *zero* and for Satisfaction, a p-value of *0.045*. Looking at the split of node two: the partitions depends just from the construct Value; the F-block presents a high significant p-value equal to *zero*. Finally, for each of the equation identified by the F-block test as responsible of the difference, we can identify the responsible path coefficients for such difference by the F-coefficient test.

Split	**Value**			**Satisfaction**		
	$\mathbf{F_B}$	**P-value**	**Significance**	$\mathbf{F_B}$	**P-value**	**Significance**
Split one: salary	6.250	0.000	YES	2.301	0.045	YES
Split two: grade	5.304	0.000	YES	1.217	0.302	NO

Table 5.28: F-block statistics values

5.8.2.2 F-coefficient Results

Observing the F-coefficient results (see Table (5.29)), we can appreciate that the responsible path coefficients of the split of the root node are Image on Satisfaction (p-value: *0.044*), the Intercept on Satisfaction and the Intercept on Value (p-value: *0.004*), meaning that the difference between the node two and three is marked not only by a coefficient: Image on satisfaction, but also by a different level of the alumni satisfaction (Intercept on Satisfaction) and a different level of the alumni Value (Intercept on Value), in the two models. The F-coefficient's p-value of the path coefficient Generic Quality on Satisfaction is not significant; however, it is near the threshold of *0.05* (it is equal to *0.069*) thus, we suspect that this path coefficient can vary across the models of node two and three.

Regarding the split of the node two the F-coefficient identifies three path coefficients: Image on Value with a p-value of *zero*, Specific Quality on Value with a p-value of *0.005*, and the Intercept on Value with a p-value of *0.020*; we observe that any driver of Satisfaction is significant. Again, we can interpret these results. Thus, the difference between the node three and the node four depends on two path coefficient: Specific Quality on value and Image on Value, and on a different level of the alumni Value (Intercept on Value). The drivers of Satisfaction don't vary across the models of the two terminal nodes.

	Split one: salary			Split two: grade		
Paths	**F_C**	**P-value**	**Significance**	**F_C**	**P-value**	**Significance**
Intercept on Value	8.211	0.004	YES	5.515	0.020	YES
Image on Value	1.110	0.293	NO	18.060	0.000	YES
Specific Quality on Value	0.662	0.416	NO	8.108	0.005	YES
Generic Quality on Value	0.359	0.550	NO	0.145	0.704	NO
Intercept on Satisfaction	3.896	0.049	YES	1.138	0.287	NO
Image on Satisfaction	4.100	0.044	YES	0.001	0.980	NO
Specific Quality on Satisfaction	0.101	0.751	NO	0.718	0.398	NO
Generic Quality on Satisfaction	3.412	0.066	NO	0.432	0.512	NO
Value on Satisfaction	0.982	0.322	NO	1.339	0.249	NO

Table 5.29: F-coefficient statistics values

At this point we have clear idea of how and why we have obtained the three final segments shown in Figure (7.3). Thus, we can proceed to compare and to analyze the difference between the three sub-models.

5.8.3 Terminal Nodes Comparison

To better appreciate the differences between the structural models of the leaves nodes obtained, we visualize them graphically. In Figure (5.13) and Figure (5.14) we can see the differential models in each leaf corresponding to Satisfaction and Value intangibles. For both latent constructs we represent its global model and the difference in the paths coefficients of the local model respect to the global one in each terminal node.

5.8.3.1 Satisfaction

In the node three, the alumni with “high salary", we can observe, as principal difference respect to the other models are the Generic Quality giving a higher Value respect to the models of node four and five, also if the F-coefficients was not significant, and the Image with a lower value respect to the other models. As we have said before when we have analyzed F-coefficients p-values of the satisfaction drivers, there is no significant difference in the models of the nodes four and five.

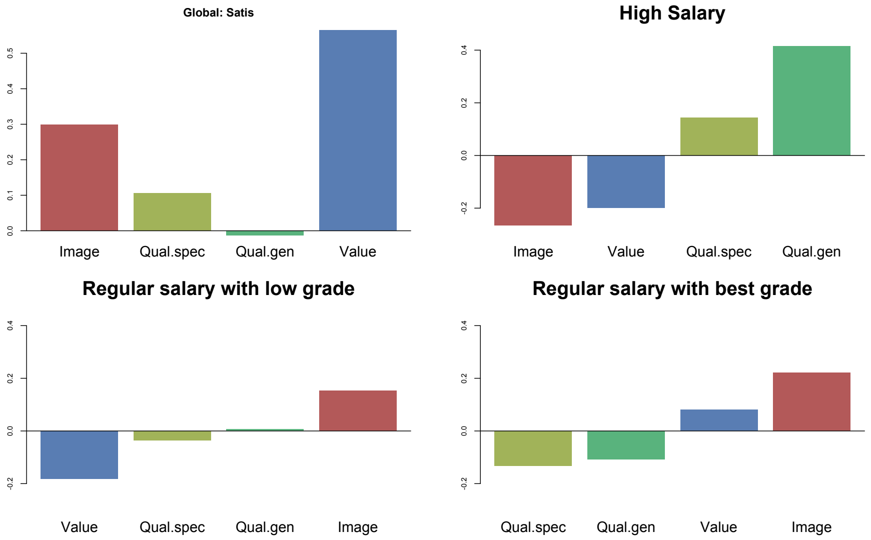


Figure 5.13: Effects on Satisfaction: comparison of the four segments with respect to the global model

5.8.3.2 Value

In the node three, the alumni with “high salary", we can observe that there are not significant difference respect the others model. We remember that the F-coefficient identified as responsible of the first split only the intercept on value meaning that the difference between the models was just for a different level of the alumni Value. Regarding the alumni with a “regular salary and low grade" (node four) and the alumni with “regular salary and good grade" (node five) we can appreciate important differences: the alumni with a “regular salary and low grade" (node four) gives more importance to the Image to define the Value whereas the alumni with “regular salary and good grade" (node five) consider more important the Specific Quality.

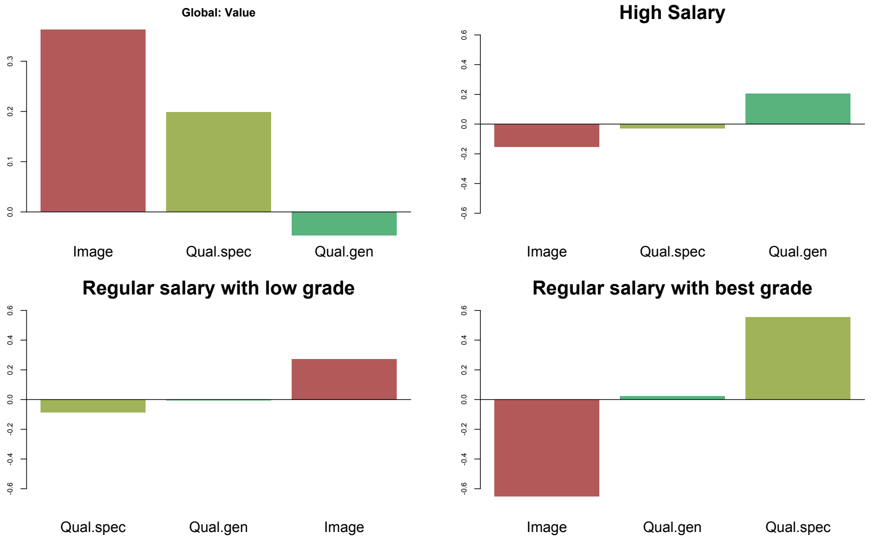


Figure 5.14: Effects on Value: comparison of the four segments with respect to the global model

5.8.4 Extending the PATHMOX Approach for Dealing with the Factor Invariance Problem

The Pathmox approach doesn't assume factor invariance, i.e., it does not impose the equality of the measurement model across nodes. Therefore, at different nodes the measurement models may also differ substantially. As a result, the meaning of the latent variables (inner model) may vary from node to node. Although such an occurrence is not necessarily undesirable, it is important to detect it in order to provide correct interpretations. To this end, we perform the invariance test. The obtained result (see Table (5.30)) is a Chi-square statistic of *190.9751* with *50* degrees of freedom, giving a p-value of *zero*, which is largely non-significant. This implies that the measurement models are different in the three terminal nodes, and that the meaning of every latent variable is specific to its segment.

Table 5.30: Invariance test results

Invariance test		
χ^2 **Statistic**	**df**	**p-value**
190.975	50	0

5.8.5 Extending the PATHMOX Approach for Overcoming the Parametric Hypothesis of the F-test

One criticism of the PATHMOX approach is that, when applied to partial least squares path modeling, it utilizes a parametric test based on the hypothesis that the residuals have a normal distribution for comparing two structural models. PLS-PM is generally utilized to model data that come from the survey analysis. These data are usually characterized by an asymmetric distribution: when an individual expresses an opinion, for example on a specific service, the opinion will likely be skewed due to social bias. This situation produces skewness in the distribution of data. As we well know, it does not matter when we utilize PLS methodology, because one goal of this method is the absence of assumptions about the distribution of data. However, it represents a limit when we compare PLS models across PATHMOX, because we cannot guarantee the normal distribution of our data, a condition needed to apply the F-test split criteria. To overcame this limit, we can extend the test to compare two robust LAD regressions to PLS path modeling.

Let us consider now the LAD approximation, and apply the PATHMOX the alumni satisfaction data-set. Let us assume the same stopping rules as in the the classic PATHMOX approach, which are:

- 0.05 as significant threshold;
- 10% as node size limit;
- 2 as tree depth level;

In Figure (5.15) we present the obtained tree. The F-global statistics calculated with the LAD approximation and the associated p-values of the obtained partitions for each node are summarized in Table (5.31). We can note that PATHMOX identifies just one significant partition by the segmentation variable salary with a F statistic of *2.5091* and a corresponding p-value of *0.0021*. In the end we obtain just two distinct models: **node two** model of students with *“regular"* salary; **node three** model of students with *“high"* salary;

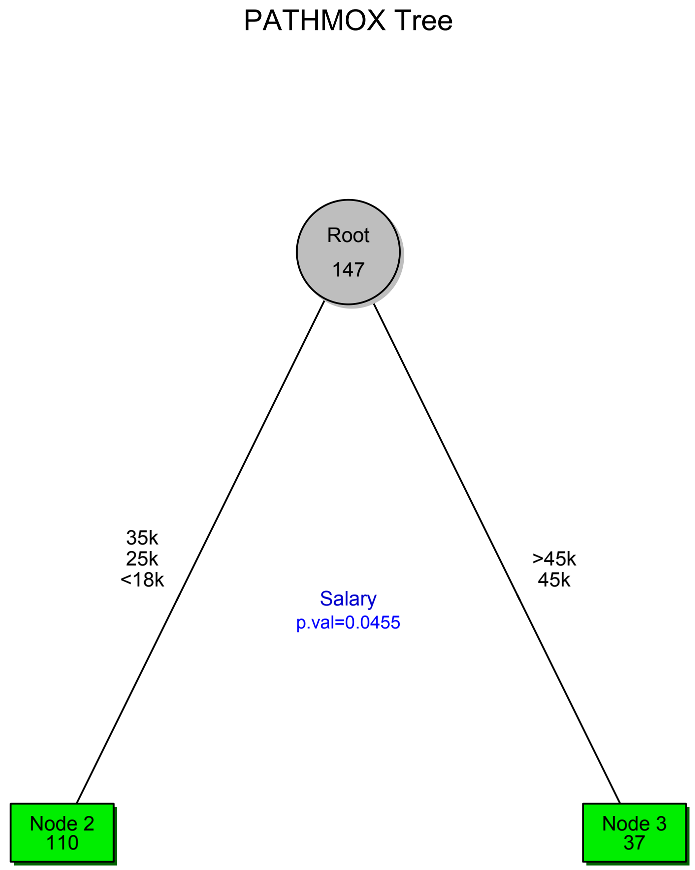


Figure 5.15: PLS PATHMOX segmentation tree - LAD approximation estimation

Table 5.31: F-global statistics values and partitions - LAD approximation

Split	F_G **statistic**	F_G **pvalue**	**Variables**	**Mod** G_1	**Mod** G_2
Split one	2.509	0.002	Career	TEL	EI/ETS

In Tables (5.32) and (5.33), we present the results obtained from applying the F-block and F-coefficient tests which are calculated by LAD approximation. We can see that the split between students with a *“regular"* and a *“high"* salary depends on the Value construct. The F-block presents in a F statistic of *2.546* and a p-value of *0.040*. Looking at the F-coefficient results (Table (5.33)), we can see that the path coefficient responsible for the split is just the Intercept on Value.

Table 5.32: F-block statistics values - LAD approximation

Split	**Value**			**Satisfaction**		
	F_B	**P-value**	**Significance**	F_B	**P-value**	**Significance**
Split one: career	2.546	0.040	YES	1.469	0.200	NO

Table 5.33: F-coefficient statistics values - LAD approximation

Split one: career			
Path coefficients	F_C	**P-value**	**Significance**
Intercept on Value	4.034	0.046	YES
Image on Value	0.461	0.498	NO
Specific Quality on Value	0.704	0.402	NO
Generic Quality on Value	0.625	0.430	NO
Intercept on Satisfaction	0.611	0.435	NO
Image on Satisfaction	2.281	0.132	NO
Specific Quality on Satisfaction	0.299	0.585	NO
Generic Quality on Satisfaction	1.862	0.174	NO
Value on Satisfaction	0.283	0.595	NO

5.9 PATHMOX Approach: Application to Mental Health Data-set

In the mental health data set, we analyze data on 138 elderly patients from seven Quebec hospitals. Our aim is, to use the partial least squares methodology and PATHMOX , to investigate the relationship between three constructs representing three mental disorders that are common in elderly populations: Dementia, Delirium and Depression. The variable description is reported in details in section (5.1.2). The manifest variables are re-propose in the following table:

Table 5.34: Description of the manifest variables for each of the latent construct - Mental health data-set

LV	MV	Item
MMSE	$mmse_3$	What month of the year is this?
	$mmse_8$	What city are we in?
	$mmse_{11}$	I am going to say 3 words. After I have said all three, I want you to repeat them
	$mmse_{12}$	Spell the word “world”
	$mmse_{16}$	Repeat the following phrase, “no ifs, ands or buts”
	$mmse_{17}$	Take this paper in your right/left hand, fold the paper in half and put it on the floor
	$mmse_{19}$	Copy this design
HDS	hds_2	Prefrontal subscale
	hds_6	Denomination subscale
	hds_7	Comprehension subscale Verbal
	hds_{11}	Reading subscale
	hds_{12}	Recent memory subscale
	hds_{18}	Motor subscale
	hds_{20}	Writing subscale
	hds_{22}	Similarities subscale
Depression	$corn_5$	Agitation; restlessness, hand wringing, hair pulling
	$corn_8$	Loss of interest; less involved in usual activities
	$corn_{15}$	Early morning awakening; earlier than usual for this individual
	$corn_{19}$	Mood congruent delusions; delusions of poverty, illness or loss
Delirium	del_4	Disorganized thinking
	del_5	Altered level of consciousness
	del_7	Memory impairment
	del_8	Perceptual disturbances
	del_{10-11}	Psychomotor agitation and Psychomotor retardation

5.9.1 Manifest and Latent Variables Relationship

In our model, the Dementia construct is treated as a second order latent variable estimated among HDS and MMSE (treated as first order latent variables). As we show in Figure (5.16), we have considered Dementia as an antecedent of Depression and Delirium, and Depression, as an antecedent of Delirium. All latent variables are formative as their own indicators describe different aspects of the diseases; Between HDS [3] and MMSE [4] and Dementia we have considered a reflective relation since these two indices reflect the presence or not of the disease in the patients.

[3] The HDS Hierarchic Dementia Scale (Cole and Dastoor, 1987) measures specific cognitive abilities and is designed to cover the broad range of performance seen in these patient groups. HDS ranks from 0 to 200; scores can indicate severe (0-39), moderate (40-99), mild (99-159) and minimal (160-200) loss of cognitive abilities.

[4] The MMSE mini-mental state examination (Folstein *et al.*, 1975) is a brief 30-point questionnaire test that is used to screen for cognitive impairment. It is also used to estimate the severity of cognitive impairment and to follow the course of cognitive changes in an individual over time; any score greater than or equal to 27 points indicates a normal cognition. Below this, scores can indicate severe (9 points), moderate (10-18 points) or mild (19-24 points) cognitive impairment.

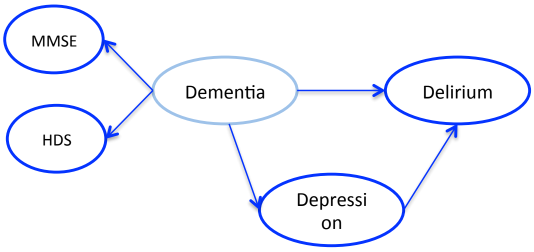


Figure 5.16: Path diagram of *3D* latent variables relationship

5.10 3 *D* PLS Global Model

The first step in PATHMOX analysis consists of specifying the global structural model that describes the relationship between the variables of interest. In this section we present the main results obtained with the classical PLS-PM approach. We first discuss the validation of the outer model and after we analyze the inner model.

For sake of interpretation we just present the results regarding the three latent variables: Dementia, Depression and Delirium.

Outer model

All constructs are specified as formative. As said previously, this can be verified in Figure (5.17), which shows for each construct the first and second eigen-values obtained by principal component analysis. For all constructs, the first two eigenvalues are similar in magnitude, which suggests that they describe different aspects of the latent variables. This result is confirmed by a correlation circle of variables for the PCA of each block, showing clear evidence that they are not unidimensional.

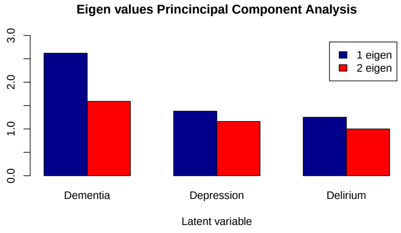


Figure 5.17: Unidimensionality validation

In Table (5.35), we present the weights, which express how each manifest variable contributes to the construct it is supposed to measure.

Dementia		**Depression**		**Delirium**	
DM_mmse_3	**0.232**	corn_5	-0.003	del_7	**0.620**
DM_mmse_8	0.029	corn_15	**0.581**	del_5	**0.342**
DM_mmse_12	0.158	corn_8	**0.578**	del_4	**0.497**
DM_mmse_19	**0.238**	corn_19	**0.462**	del_9_10	0.333
DM_mmse_17	0.1440			del_8	0.248
DM_mmse_11	0.100				
DM_mmse_16	0.151				
DM_hds_20	0.179				
DM_hds_12	**0.205**				
DM_hds_11	0.083				
DM_hds_7	0.051				
DM_hds_22	**0.357**				
DM_hds_2	0.087				
DM_hds_18	0.033				
DM_hds_6	0.053				

Table 5.35: Results of the outer model: outer weights

The most important weights are:

- Dementia: *DM_mmse*_3 (What month of the year is this?), *DM_mmse*_19 (Copy this design) and *DM_hds*_12 (Recent memory sub-scale) and *DM_hds*_22 (Similarities sub-scale);
- Depression: *corn*_8 (Loss of interest; less involved in usual activities), *corn*_15 (Early morning awakening; earlier than usual for this individual), *corn*_19 (Mood congruent delusions; delusions of poverty).
- Delirium: *del*_7(Memory impairment), *del*_4 (Disorganized thinking) and *del*_5 (Altered level of consciousness)

The Table (5.36), shows the correlations between manifest variables and constructs. It should be noted that, as we would expect, the manifest variables (MV) are more correlated with their own constructs (LV) than with the others

LV	MV	Dementia	Depression	Delirium
Dementia	DM_mmse_3	**0.567**	0.053	0.309
	DM_mmse_8	**0.211**	-0.074	0.125
	DM_mmse_12	**0.435**	-0.013	0.179
	DM_mmse_19	**0.561**	0.207	0.307
	DM_mmse_17	**0.353**	0.113	0.196
	DM_mmse_11	**0.286**	0.019	0.153
	DM_mmse_16	**0.428**	-0.024	0.147
	DM_hds_20	**0.473**	-0.046	0.183
	DM_hds_12	**0.545**	-0.022	0.280
	DM_hds_11	**0.316**	0.029	0.131
	DM_hds_7	**0.228**	-0.047	0.147
	DM_hds_22	**0.707**	-0.0310	0.441
	DM_hds_2	0.**135**	0.049	0.077
	DM_hds_18	**0.257**	-0.064	0.135
	DM_hds_6	**0.162**	0.019	0.172
Depression	corn_5	0.014	**0.167**	0.039
	corn_15	-0.032	**0.463**	0.123
	corn_8	0.047	**0.530**	0.123
	corn_17	0.017	**0.553**	0.135
	corn_19	0.056	**0.521**	0.119
Delirium	del_7	0.376	0.085	**0.650**
	del_5	0.212	0.054	**0.371**
	del_4	0.267	0.161	**0.533**
	del_9_10	0.157	0.191	**0.383**
	del_8	0.150	0.102	**0.308**

Table 5.36: Correlations between the manifest variables and the latent constructs

Inner model

The inner model validation is presented in Figure (5.18):

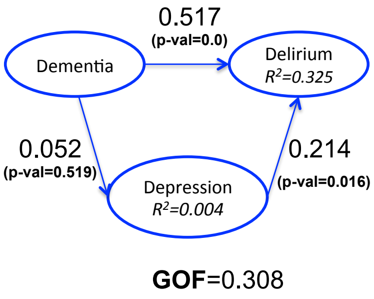


Figure 5.18: Path diagram of *3D* latent variable relationships. In parenthesis, we show the significance's p-value the path coefficients

We can see that when we analyze the Depression construct, the effect of Dementia on Depression is very low, (path coefficient: 0.052 not significant) and the R^2 is practically zero. When we consider the Delirium construct we find an important effect of Dementia on Delirium (path coefficient: 0.517), whereas the effect of Depression on Delirium is lower (path coefficient: 0.214); in this case the R^2 is 0.325. Both coefficients are significant. The goodness of fit, a global criterion for assessing the quality of PLS model, is 0.308.

5.11 PATHMOX Tree

We can now investigate by a tree analysis whether or not the global PLS model is valid for the whole population taking in account the segmentation variables: patientÕs gender, time of hospitalizations, patients' age. Our analysis suggests that the sample can be split into two subsamples, each with a distinct PLS model. The tree-structure obtained by PATHMOX is given in Figure (3.1).

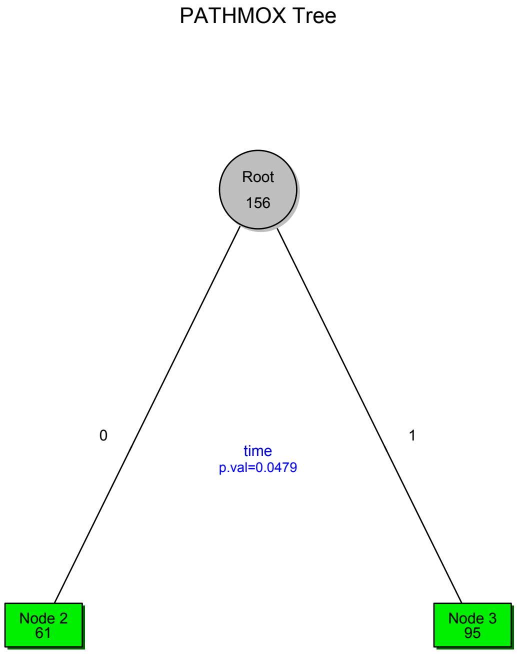


Figure 5.19: PATHMOX's Segmentation Tree

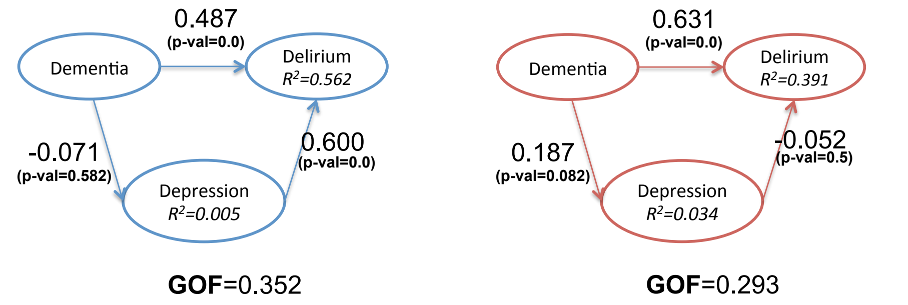
Table 5.37: F-global statistics values and partitions

Split	F_G **statistic**	F_G **pvalue**	**Variables**	**Mod** G_1	**Mod** G_2
Split one	4.056	0.000	time	less than one year (time = 0)	more than one year (time = 1)

The partition is obtained by the segmentation variable *duration of hospitalization* (indicated withe the codification "time" in the Figure (5.19)) with an F-global statistic of *1.912* and a corresponding p-value of *0.047*. The root node is split into two children nodes: the **node two** with *61* patients and a duration of hospitalization less than one year and the **node three** with *95* patients and a duration of hospitalization of more than one year.

5.11.1 Terminal Nodes Comparison

The last step in PATHMOX consists in the terminal nodes comparison. Further details are contained in the following Figure (5.20), which represents the inner models at the two terminal nodes.



PLS model node two: shorter term hospitalization patients **PLS model node three**: longer term hospitalization patients

Figure 5.20: Path diagram of the two terminal nodes identified by PATHMOX. From left to right, we find the path diagram of node two: *shorter term hospitalization patients* and the path diagram of node three: *longer term hospitalization patients*.

For the shorter hospitalization node, we can see there is a large difference in the path coefficient Depression on Delirium: *0.600* and significant (p-value: *zero*). For the longer term hospitalization patients the path coefficient Depression on Delirium is very small (*-0.052*) and not significant.

5.11.2 Extended PATHMOX Analysis

As discussed above, PATHMOX also provides very useful aid to interpretation through the F-block and F-coefficient tests. The results of these tests as applied to our data are given in Table (5.38):

F**-block test**			
Constructs	**Statistic**	**P-value**	**Significance**
Depression	2.192	0.110	NO
Delirium	3.792	0.011	YES
F**-coefficient test**			
Path coefficients	**Statistic**	**P-value**	**Significance**
Intercept on Depression	2.912	0.087	NO
Dementia on Depression	1.503	0.222	NO
Intercept on Delirium	1.568	0.218	NO
Dementia on Delirium	2.043	0.152	NO
Depression on Delirium	8.011	0.000	YES

Table 5.38: F-block and F-coefficients results

Note that the F-block test identifies the Delirium construct as the one responsible for the difference in the inner models at the two child nodes (F-block statistic = *3.792*, p-value: *0.011*) . The F-coefficient tests identify the path coefficient that links Depression to Delirium as significantly different across the two child nodes (F-coefficient statistic: *8.011*, p-value: *zero*). Hence we can conclude that model heterogeneity exists at the inner model level.

Now we turn to the measurement models. The invariance test is highly significant, with a Chi-squares statistic of *93.391* on *50* degrees of freedom (p-value: *zero*).

Table 5.39: Invariance test results

Invariance test		
χ^2 **Statistic**	**df**	**p-value**
93.391	50	0

This implies that the measurement models are different in the two terminal nodes and that the meaning of every latent variable is specific to its segment. This can be seen from the three bar-charts of Figure (5.21), which graph the difference of the weights of two PLS model nodes.The bars in red indicate the most significant difference.

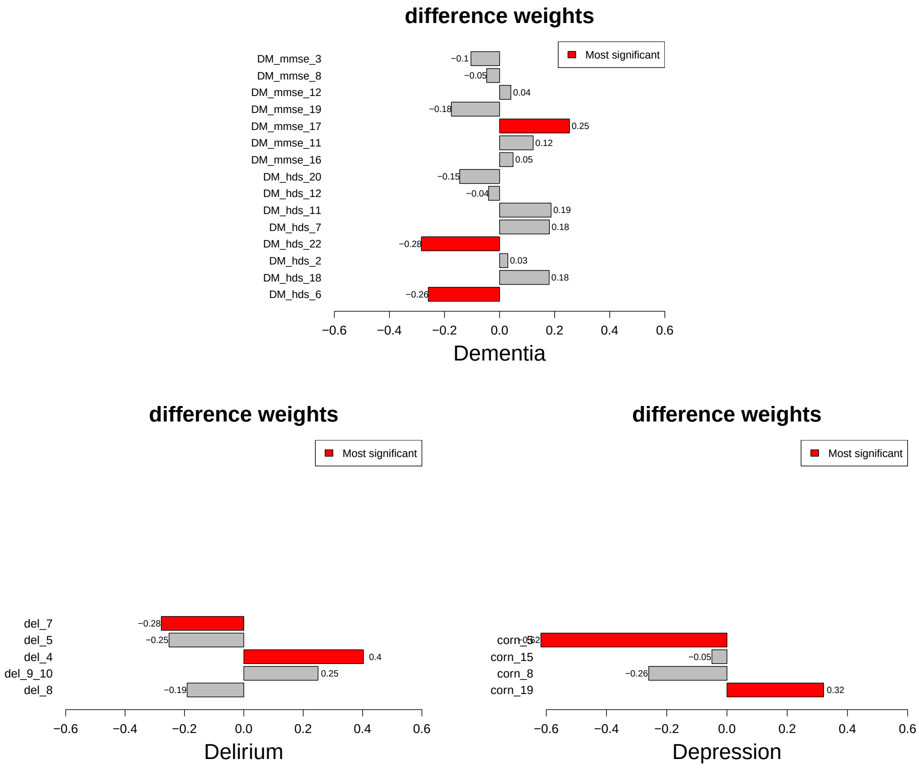


Figure 5.21: Bar-chart of the weights difference of the two PATHMOX's terminal nodes

Chapter 6

Conclusion and Further Works

This work deals with the problem of modeling heterogeneity. We conducted our research following the path of Gastón Sánchez's 2009 work on PLS Path Modeling. We started with two principal objectives: we wanted to improve the PATHMOX approach by going deeper into the analysis of the obtained partitions; and we wished to investigate the possibility of extending PATHMOX to other methodologies.

To achieve the first purpose, we considered three distinct aspects of PATHMOX that could be been improved:

- extend PATHMOX to detect which constructs differentiate segments;
- extend PATHMOX to deal with the factor invariance problem;
- extend PATHMOX to overcome the parametric hypothesis of the F-test.

The first aspect is related with the split criterion used in PATHMOX. Granted that the PATHMOX approach allows us to detect the existence of different models in a data-set without identifying segmentation variables beforehand, we have underlined that the F-test is a global criterion. We have also questioned if it allows assessing whether all the coefficients for two compared models are equal or not. We found that it does not indicate which particular equation and which coefficients are responsible for the difference. To asses the endogenous equation and path coefficients that are responsible for a split, we have proposed the F-block and the F-coefficient tests. Considering the same lemma of Lebart *et al.* (1979) that was also used by Sánchez in 2009 to justify the F-global test distribution, we have shown that we can readapt the F-global test to achieve these statistics. We have proof by simulation that the coherence of the F-block and the F-coefficient respects the F-global when two path models are compared. In the end, we reprogrammed the PATHMOX algorithm to include these two tests in order to verify by real-data whether the interpretation of the identified partitions can be improved. The second aspect refers to the restriction that PATHMOX applies to generate partitions. The PATHMOX approach doesn't assume factor invariance, i.e., it does not impose equality of the measurement model across nodes. Therefore, the measurement models may also differ substantially at different nodes. As a result, the meaning of the latent variables (inner model) may vary from node to node. Although such an occurrence is not necessarily undesirable, it is important to detect it in order to provide correct interpretations. To overcome the problem of factor invariance, we have suggested the invariance measurement test. We have used again the second Lemma of Lebart *et al.* (1979) to provide an argument to justify the χ^2 distribution of the test. We have shown by simulation the effectiveness of the test to detect differences in the measurement weights of the tree's terminal nodes. Again, we have included this statistic in the PATHMOX procedure, showing the importance of analyzing whether or not the weights in terminal nodes are equivalent in real data applications. The last aspect that we have considered is the parametric hypothesis of the F-test. The PATHMOX approach utilizes a parametric test to compare two structural models, based on the hypothesis that the residuals have a normal distribution. This imposes a limit when we compare PLS models across PATHMOX, because we cannot guarantee the normal distribution of our data. To overcame this limit, we can extend the test to compare two robust LAD regressions (Koenker and Bassett, 1982) in PLS path modeling. We proved by a simulation study that the F-tests calculated through LAD approximation provide the same results as the classic F-test, the only difference being that LAD approximation tests are more conservative. In the end, we programmed the PATHMOX approach with LAD approximation to test the functionality with real-data.

Starting with the consideration that even if it makes sense to compare models by taking into account a set of segmentation variables, and even if an appropriate split criterion is utilized for the comparison, we can apply PATHMOX to explore the hypothesis of heterogeneity in our data. We have shown that we can generalize PATHMOX to other methodologies. We have taken into account different scenarios regarding linear regression, robust regression, generalized linear regression and principal component analysis. We have individuated an appropriate split criterion in order to compare partitions in each one of them, and we have programmed PATHMOX to be applied to this distinct methodology. In the end, we have achieved improvements in the analysis and interpretation of real-data.

6.1 Further Works

Granted that the simulation showed that the invariance measurement test works well and is suitable for detecting differences in the terminal nodes at measurement levels, this test is a global criterion: we know if the weights of the terminal nodes are the same or not, but we do not know which nodes or weights are responsible for the difference. Thus, an interesting work would consider the possibility of extending the same logic of the F-block and the F-coefficient test to a measurement model. More simulations should be done to analyze the F-test LAD approximation. One interesting aspect would be to simulate data with outliers, in order to investigate changes in the effectiveness of the statistics. Again, as underlined by Sánchez (2009), more complex models should be considered in order to investigate more deeply the behavior of the three F-tests. More advances can be made by going deeper into the differences between formative and reflective models in PATHMOX.

Regarding the second issue of this dissertation, extending PATHMOX to other methodologies, we can argue that we are just at the beginning. We have proof that it is possible to apply it in different scenarios, but many other methodologies could certainly be considered. For example, we could extend PATHMOX to the mixture regression by taking the maximum likelihood as a criterion for model comparison. We could go deeper into the analysis of the GLM models by analyzing PATHMOX in the context of other models as a multinomial regression. We could also analyze the behavior of PATHMOX in the context of LISREL.

Chapter 7

The genpathmox R Package

7.1 Introduction

In the present annex we present the **genpathmox** package that provides a solution for handling segmentation variables in complex statistical methodology. It contains a generalized version of the PATHMOX algorithm to approach different methodologies: linear regression and least absolute regression models. Furthermore, it implement en extended version of the PATHMOX algorithm in the context of partial least square path modeling (Sánchez, 2009) including the F-block test (to detect the responsible latent equations of the difference), the F-coefficient (to detect the path coefficients responsible of the difference) and the invariance test (to realize a comparison between the sub-models' latent variables).

7.2 The PATHMOX Algorithm Description

The package **genpathmox** implements the PATHMOX. The algorithm starts with establishing a set of admissible partitions for each node of the tree using the segmentation variables information. Once considered all possible subgroups, the algorithm, following a global comparison approach (Aluja 2013b) (i.e split criterion), compares all sub-models calculated for each partition defining a degree of difference (i.e p-value) by a statistic test. The optimal partition is chosen by comparing the different p-values, selecting the lowest. The choice of the split criterion is related with both, the nature of the analyzed data, and the kind of methodology used to approach data. This step is repeated iteratively thus obtaining a ranking of the best segmentation variables and the different nodes of the tree. The last phase is to identify a criterion for discriminating whether the segment is an intermediate node or terminal (*stopping rule*). Here a pre-pruning rule is adopted. One node is considered terminal when: 1) the segmentation variables for that parent node have been exhausted and no more splits are feasible; 2) the selected best split has an associated p-value exceeding the significance threshold, 3) the number of elements in one node is less than or equal to some predefined minimum number of elements. The procedure that can be summarized in the following four main steps:

Algorithm 2 PATHMOX Algorithm

```
Step 1. Start with the global model at the root node
Step 2. Establish a set of admissible partitions for each segmentation variable in each node of the tree
Step 3. Detect the best partition by:
        3.1. Compare all binary partitions in all segmentation variables
        3.2. Apply the test, calculating for each comparison a p-value
        3.3. Sort the p-values in a descending order
        3.4. Chose as the best partition the one associated to the lowest p-value
Step 4. If (stop criteria^1 = false) than
        repeat by step 3
        1. Posible stop criteria:
        a. The number of individuals in the group falls below a fixed level
        b. The p-values test are not significant
        c. Maximum number of tree depth attained
```

7.2.1 Split Criterion Implemented in genpathmox

The package **genpathmox** implements different split criteria according the selected methodology (see Table (7.1)):

Table 7.1: genpathmox split criterion

Methodology	genpathmox split criteria	References
OLS regression	F-global test, F-coefficient test	Lebart *et al.* (1979)
LAD regression	F_{LAD}-global, F_{LAD}-coefficient	Birkes (1993)
PLS-PM	F-global, F-block, F-coefficient	Sánchez (2009); Aluja (2013)

7.3 genpathmox Installation and Usage

genpathmox is freely available from the Comprehensive **R** Archive Network, better known as CRAN, at: http://cran.r-project.org/web/packages/genpathmox/

7.3.1 Installation

The main version of the package is the one hosted in CRAN. You can install it like you would install any other package in **R** by using the function `install.packages()`. In your **R** console simply type:

```
> # installation
> #
> install.packages("genpathmox")
```

Once **genpathmox** has been installed, you can use one of the functions `library()` or `require()` to load the package.

7.4 What's in genpathmox

genpathmox comes with a number of functions to perform a series of different types of analysis. The functions `pls.pathmox()` and `pls.treemodel()` are designed to run the pathmox analysis in the context of the PLS-PM including some specific tools as the F-block and the F-coefficients tests that allow to detect the responsible equations and the responsible coefficients of the difference, and the invariance test that allows to compare the latent variables' weights of the identified sub-models. The functions `reg.pathmox()` and `reg.treemodel()` generalize the PATHMOX algorithm to the context of the multiple linear and least absolute deviation regression. The accessory functions of `pls.pathmox()` and `reg.pathmox()` are the plotting, the summary and print functions, whereas `pls.treemodel()` and `reg.treemodel()` just have the plotting function.

7.5 genpathmox Data-sets

To illustrate the different tools of the **genpathmox** we will consider an example on two ICT schools. It comes from a survey realized in 2008, collected on 147 students. It consists in a total of 39 variables collected in a questionnaire inspired on the ECSI model questions. The goal of the study was to explain the satisfaction from its drivers, image of the school, expectations on generic skills [1], expectations on technical skills [2], perceived quality on generic skills, perceived quality on technical skills and value or profit obtained after graduation. This variables are defined in the following way:

- **Satisfaction** Degree of alumni satisfaction about the formation in school respect to their actual work conditions;
- **Image** Generic alumni perception of ICT schools: (internationally recognition, ranges of courses, leadership in research, . . .);
- **Specific Expectation** Perceived Expectation on specific skills (technic or applied skills);
- **Generic Expectation** Perceived Expectation on generic skills (abilities in problem solving, communication skills);
- **Generic Quality** Perception about achieved quality on the generic skills in the school (abilities in solving problem, communication skills);
- **Specific Quality** Perception about the achieved quality on the specific skills in the school;
- **Value** The advantage or profit that the alumni may draw from the school degree (well paid job, motivated job, prospectives in improvement and promotion).

10 segmentation variables, which serve as observed sources of heterogeneity variables, were also collected. A description is showed in Table (7.2).

[1] We understand by general skills a broad - spectrum of capabilities not specific to a profession or organizational environment, such as the ability of problem solving, communication, time management team working initiative . . .

[2] The technical skills refers to the knowledge and abilities, specific to a profession, either mathematical or engineering based, or specific to accomplish a technical tasks.

Table 7.2: Codification of segmentation variables according to their type of scale and levels

Name	Scale	N. levels	Levels description
Career	nominal	3	EI / ETS / TEL
Gender	binary	2	female / male
Age	ordinal	4	25-26years / 27-28years / 29-30years / more than 30
Studying	binary	2	yes studying / no studying
Contract	nominal	3	fixed / temporary / others types
Salary	ordinal	5	less then 18k / 25k / 35k / 45k / more then 45k
Firm-type	binary	2	private / public
Access grade	ordinal	3	acc-note less then 7 / acc-note 7-8 / more then 8
Grade	ordinal	4	note less than 6.5 / note 6.5-7 / note 7-7.5 / more then 7.5
Start-work	binary	2	after graduated / before graduated

In **genpathmox** we have two data-sets. The first one contains an estimation of the variable: satisfaction, image, specific expectation, generic expectation, generic quality, specific quality and the value, and the 10 segmentation variables showed in Table (7.2). We use it to show how PATHMOX works in the context of regression models. The second one, called `fibtele` contains the survey's questions and again the 10 segmentation variables. We handle it to run an example of PATHMOX in the partial least square path-modeling context.

7.6 An genpathmox Example in Regression Model

In order to calculate the PATHMOX segmentation tree, it is necessary to specify the scale (e.g., binary, ordinal, or nominal) of the segmentation variables (see Table (7.2)). In addition, we have to indicate in the function `reg.pathmox()`, by the code `method="lm"`, that we want to run PATHMOX in the context of the multiple linear model and we have to fix the stop conditions of the algorithm. We have decided to establish a value of 0.05 (`signif=0.05`) for the threshold of the p-value to look for those partitions that are highly significant. Given that we have a total sample of 147 students, a number of 22 students (15% of total sample) (`size=0.15`), seems us to be a reasonable minimum number to stop the growth of a node. The depth level = 2 (`deep=2`), has been selected with the aim to obtain a simple segmentation tree with a possible maximum number of four final segments. We can easily read our data and indicate the scale of our factors by the code:

```
> #load data set
> #
> data(fibtelereg)
> #
> data.fib = fibtelereg
> #
> #Identifying the segmentation variables
> #
> segvar = data.fib[,2:11]
> #
> #Identifying the manifest variables
> #
> data.fib = data.fib[,12:18]
> #
> #Rescaling the segmentation variables
> #
> segvar$Age = factor(segvar$Age, ordered=T)
> #
> segvar$Salary = factor(segvar$Salary,
+ levels=c("<18k","25k","35k","45k",">45k"), ordered=T)
> #
> segvar$Accgrade = factor(segvar$Accgrade,
+ levels=c("accnote<7","7-8accnote","accnote>8"), ordered=T)
> #
> segvar$Grade = factor(segvar$Grade,
+ levels=c("<6.5note","6.5-7note","7-7.5note",">7.5note"), ordered=T)
```

Then, we can perform the PATHMOX analysis by using function `reg.pathmox()`.

```
> #Run the PAthmox analysis
> #
> fib.reg.pathmox = reg.pathmox(Satisfact~.,data=data.fib,segvar,signif=0.05,
+ deep=2,method="lm",size=0.15, tree=FALSE)

REGRESSION SEGMENTATION TREE

---------------------------------------------
Info Parameters Algorithm
  parameters Algorithm value
1     Threshold signif  0.05
2   Node size limit(%)  0.15
3     Tree depth level     2
4               Method    lm

---------------------------------------------
Info Segmentation Variables
          Nlevels Ordered Treatment
Career          3   FALSE   nominal
Gender          2   FALSE    binary
Age             4    TRUE   ordinal
Studying        2   FALSE    binary
Contract        3   FALSE   nominal
Salary          5    TRUE   ordinal
Firmtype        2   FALSE    binary
Accgrade        3    TRUE   ordinal
Grade           4    TRUE   ordinal
Startwork       2   FALSE    binary
```

We can see that the first results that we obtain are some informations about the algorithm parameters (imputed in the function `reg.pathmox()`) and a summary of the segmentation variables; in order we have the name of the segmentation variables, the number of levels of each one of them, if they are ordered and the the way in which we treat them (i.e if we consider them as ordinal, binary or nominal).
Let us to use now the function `plot()` to obtain a graphic visualization of the tree.

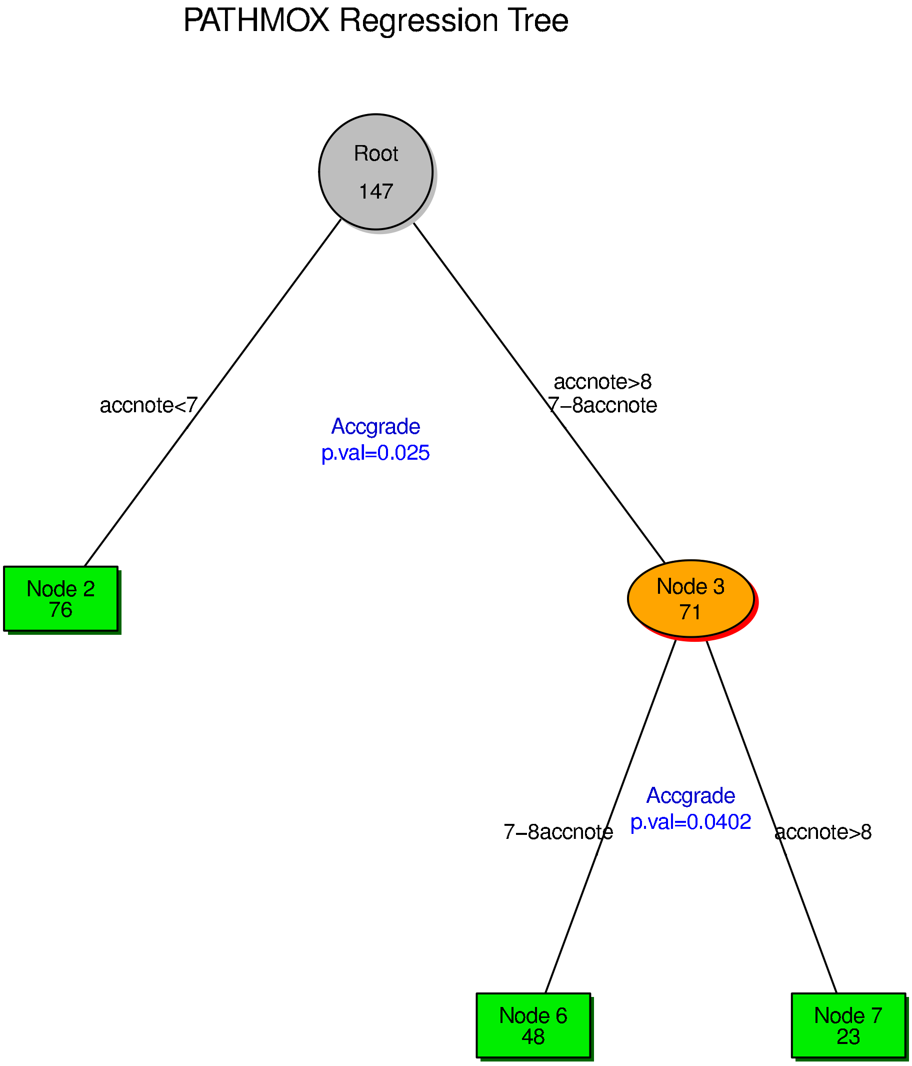


Figure 7.1: Output function `plot()` for the `fib.reg.pathmox` object. PATHMOX segmentation tree

In Figure (7.1), we present the obtained tree, where we can observe that in fact, there are three distinct models. At the first split, PATHMOX defines two different models one for students with a *"low"* access grade (< 7), and the other, for students with a *"higher"* access grade (*7-8* and > 8). Let as define the students with a *"low"* access grade (< 7) as students with *"low profile"*, and the students with a *"higher"* access grade (*7-8*) and (> 8) as students with *"higher profile"*; we see that the produced split is significant, giving a F-statistic of *2.384* with a p-value of *0.025*. The node two is taken as final because no significative partitions has been found. The tree continues by splitting the node three. The most significant partition is obtained, again by variable *acc-grade*, giving a F-statistic of *2.283* with a p-value of *0.040*, dividing now the students with a better profile, in students with a *"good profile"* (*acc-note 7-8*), and students with the *"best profile"* (*acc-note* > 8). This ends the splitting process, as the maximum depth of two levels has been reached. Hence, at the end, we have three final segments each one corresponding to a distinct model: **node 2** model of students with a *"low profile"*; **node 6** model of students with a *"good profile"*; **node 7** model of students with the *"best profile"*.

Running the function `summary()` we have a complete information on the obtained PATHMOX tree.

The first result (Info Parameters Algorithm) is a table with the parameters fixed to run the PATHMOX analysis: the threshold of the p-value, minimum node's size, depth level and the method.

The second information that we achieve is the Info Tree, that informs about the tree depth and the number of the terminal nodes.

The third result is the Info nodes, that contains the description of each node of the tree; here the algorithm provides the corresponding number of the node (variable node), what is the parent node of each node (variable parent), which is the depth corresponding to a node (depth), which nodes are root, intermediates and leaves (type and terminal), the number of the individuals (size) in each node, the % size respect the total (variable percent) and the segmentation variable used to generate the node (variable and category).

The last information (Info Splits) allows to investigate the reasons of the splits. Here the PATHMOX provides the segmentation variables, the *F* statistic and the p-value of the obtained partitions (*F*-global test). Next in Info Splits we find the *F-coefficient* test results (Lebart *et al.* 1979) that allows to go deeper and to identify the responsible coefficients of the splits.

Looking this last result, we can see that the significant difference identified by PATHMOX between students with a *"low profile"*, and students with a *"higher profile"*, depends on the coefficients Value (*F*-coefficient statistic: *5.389* and associated p-value: *0.021*) and Specific Quality (*F*-coefficient statistic: *4.507* and associated p-value: *0.036*); in the partition of node 3: students with a *"good"* and *"best profile"*, the difference is due to Generic Expectations with a *F*-coefficient statistic of *5.534* and an associated p-value of *0.024*.

```
> summary(fib.reg.pathmox)

REGRESSION SEGMENTATION TREE

---------------------------------------------
Info Parameters Algorithm:
  Parameters Algorithm value
1     Threshold signif  0.05
2   Node size limit(%)  0.15
3     Tree depth level     2
4               Method    lm
---------------------------------------------
Info Tree:
        Parameters Tree value
1             Deep tree     2
2 Number terminal nodes     3
---------------------------------------------
Info nodes:
  Node Parent Depth Type Terminal Size Percent Variable             Category
1    1      0     0 root       no  147  100.00     <NA>                 <NA>
2    2      1     1 leaf      yes   76   51.70 Accgrade            accnote<7
3    3      1     1 node       no   71   48.30 Accgrade 7-8accnote/accnote>8
4    6      3     2 leaf      yes   48   32.65 Accgrade           7-8accnote
5    7      3     2 leaf      yes   23   15.65 Accgrade            accnote>8
---------------------------------------------
Info Splits:

Variable:
  node variable     mod.g1               mod.g2
1    1 Accgrade  accnote<7 7-8accnote/accnote>8
2    3 Accgrade 7-8accnote            accnote>8

F.statistic global:
  node fg.statistic fg.pvalue
1    1       2.3838   0.02502 *
2    3       2.2832   0.04024 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F.statistic coefficient:

Node 1 :
          fc.statistic fc.pvalue
Intercept       0.2098   0.64764
Image           0.4855   0.48714
```

```
Exp.gene        0.6015   0.43939
Exp.spec        0.0000   0.99636
Qual.gen        2.1425   0.14563
Qual.spec       4.5067   0.03561 *
Value           5.3897   0.02178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Node 2 :
          fc.statistic fc.pvalue
Intercept       0.0042    0.9487
Image           2.3290    0.1325
Exp.gene        5.3539    0.0243 *
Exp.spec        2.2575    0.1385
Qual.gen        0.1514    0.6987
Qual.spec       1.2665    0.2651
Value           0.8490    0.3607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The last part of the analysis consists in the comparison between the terminal nodes of the tree. We can realize this last step by the function `reg.treemodel()`. This function allows to visualize the main results of the analysis for each one of the identified sub-models. In order, we can appreciate the the method used to fit the regression models (`$method`), the estimated coefficients (`$coefficients`) the standard deviation (`$Std.`), the p-value significance (`$pval.coef`) and R^2 (`$r2`). In the function `reg.treemodel()` we can also use the code `label=TRUE` and `label.nodes` to give a specific name to each node of the tree. In our example we have used the label “global model” for the root node, “low profile” for the node two, “good profile” for the node six, and “best profile” for the node seven.

```
> fib.treemodel=reg.treemodel(fib.reg.pathmox, label=TRUE,
+ label.nodes=c("Global model","low profile","good profile", "best profile"))
> fib.treemodel

$method
[1] "lm"

$coefficients
              global model low profile good profile best profile
predImage            0.240       0.271        0.128        0.368
predExp.gene         0.111       0.154        0.003        0.276
predExp.spec        -0.083      -0.136       -0.045       -0.214
predQual.gen         0.026       0.171       -0.025       -0.051
predQual_spec        0.123       0.016        0.199        0.398
predValue            0.575       0.660        0.461        0.325

$Std.
              global model low profile good profile best profile
predImage            0.078       0.124        0.089        0.186
predExp.gene         0.054       0.081        0.070        0.122
predExp.spec         0.053       0.096        0.071        0.110
predQual.gen         0.071       0.121        0.091        0.145
predQual_spec        0.066       0.099        0.125        0.132
predValue            0.055       0.086        0.073        0.154

$pval.coef
              global model low profile good profile best profile
predImage            0.002       0.031        0.156        0.064
predExp.gene         0.042       0.063        0.961        0.037
predExp.spec         0.123       0.160        0.532        0.069
predQual.gen         0.709       0.162        0.786        0.728
predQual_spec        0.066       0.870        0.119        0.008
predValue            0.000       0.000        0.000        0.050

$r2
global model  low profile good profile best profile
   0.688       0.7167       0.728       0.8056

attr(,"class")
[1] "treemodelreg"
```

In order to provide a visual way to appreciate the differences among segments, we can use the `plot()` function. It provides the bar charts of the coefficients of each terminal nodes (see Figure7.2).

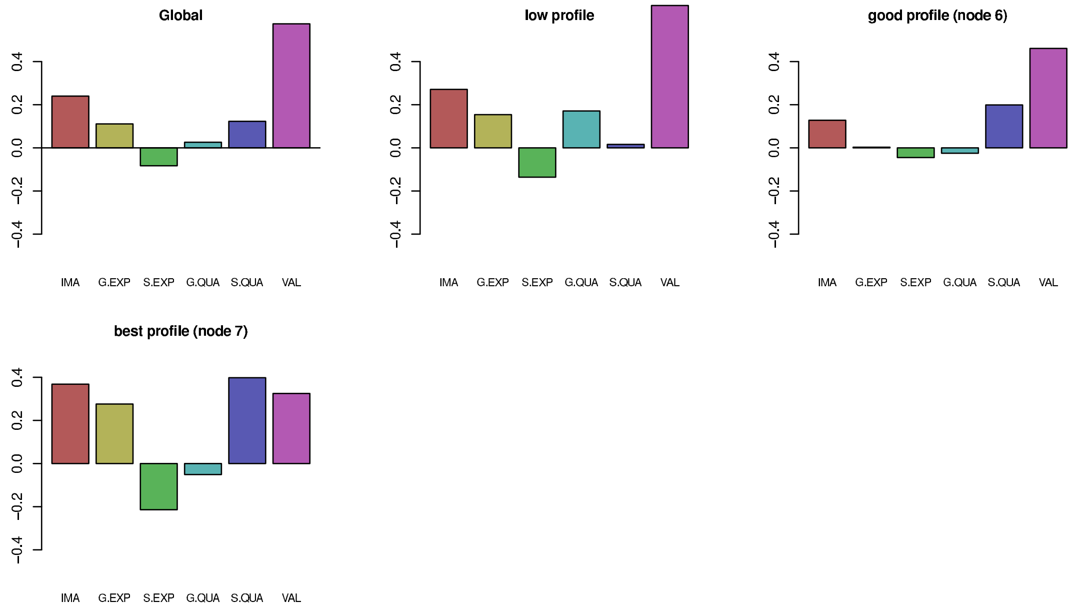


Figure 7.2: Output function `plot()` for the `fib.treemodel` object. Effects on Satisfaction: comparison of the three segments with respect to the global model

We can see that the difference between the node 2 (*"low profile"* students) and the other two terminal nodes depends, above all, on the Value as underlined by the F-coefficients results; the difference between the model of students with *"good profile"* (node six) and the model of students with *"best profile"* (node seven) is defined mainly by the Generic Expectations confirming the F-coefficients results. All the R^2 are high significant. In particular the node six and seven present a R^2 index respectively of *0.767* and *0.812*. An important consideration is about the p-values of the coefficients of each terminal node. We can see for example, that the Value is significant in all groups; sometimes, also if the coefficients are highly distinct respect the other models, they could be not significant: for example node two has Specific Quality distinct respect the other two models but not significant; the node six is different from the node seven also for the coefficient Generic Expectations but again it is significant in node seven. Thus, we have to take care when we interpret the models of the terminal nodes.

7.7 An genpathmox Example in Partial Least Squares Path Modeling PLS-PM

Partial least squares path modeling (PLS-PM) is one of the methods from the broad family of PLS techniques. It was originally developed by Herman Wold and his research group during the 1970s and the early 1980s. Around the PLS community, the term path modeling is preferred to that of structural equation modeling, although both terms can be found within the PLS literature. Our preferred definition of PLS-PM is based on three fundamental concepts:

1. It is a multivariate method for analyzing multiple blocks of variables
2. Each block of variables plays the role of a latent variable
3. It is assumed that there is a system of linear relationships between blocks

In other words: PLS-PM provides a framework for analyzing multiple relationships between a set of blocks of variables (or data tables). It is supposed that each block of variables is represented by a latent construct or theoretical concept; the relationships among the blocks are established taking into account previous knowledge (theory) of the phenomenon under analysis. There are plenty of references about PLS-PM, but we will only mention one from Wold, and two more recent ones: (Vinzi 2010; Tenhenaus 2005; Wold 1982).

7.7.1 Partial Least Squares Segmentation Tree: PATHMOX Analysis

In order to calculate the PATHMOX segmentation tree in the context of PLS-PM, we have to fit first the global model. We realize this step by using the R function `plspm()` of the **plspm** package (Sánchez 2009). The **genpathmox** package works integrating the `plspm()`, thus, to run a PATHMOX analysis in PLS-PM by **genpathmox** we have to use the `plspm()` function to calculate the global model.

To run the `plspm()` function we have define at least four different parameters:

- the `data` parameter. It is a matrix or data frame containing the manifest variables.
- the `path_matrix` parameter. It is a square (lower triangular) boolean matrix representing the inner model (i.e. the path relationships between latent variables).
- the `blocks` parameter. It is a list of vectors with column indices or column names from Data indicating the sets of manifest variables forming each block (i.e. which manifest variables correspond to each block).
- the `modes` parameter. It is character vector indicating the type of measurement for each block. Using the code `modes=A` we indicate a reflective block, whereas by the code `modes=B`, we indicate a formative block.

We recommend to examine the **plspm** reference manual at: http://cran.r-project.org/web/packages/plspm/plspm.pdf to understand in details the different tools of the function. In our example we consider a simple inner model. We want to investigate the alumni satisfaction considering the following causal model:

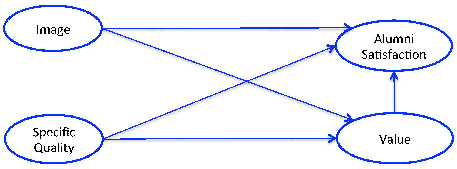


Figure 7.3: Path diagram representing the casual relationship between the latent variables of the model

We have considered as antecedent of the value the image and the specific quality, and, as antecedent of satisfaction, the value, the image and the specific quality. We have specified our model by the `path_matrix` parameter. The measurement model of the latent variables is taken in a reflective way (`modes.fib = rep("A", 4)`) in all cases, that is, we assume that every intangible represents a concept being their indicators, reflects of that concept. We can run the `plspm()` function by the following code:

```
> data(fibtele)
> #
> data = fibtele[,12:35]
> #
> data.fib = data[,-c(12:16)]
> #
> #Defining the inner model
> #
> Image = rep(0,4)
> Qual  = rep(0,4)
> Value = c(1,1,0,0)
> Satis = c(1,1,1,0)
> #
> inner.fib = rbind(Image,Qual, Value, Satis)
> #
> colnames(inner.fib) = rownames(inner.fib)
> #
> #Defining the blocks of indicators (outer model)
> #
> outer.fib = list(1:8,9:11,12:15,16:19)
> #
> #Defining the modes
> #
> modes.fib = rep("A", 4)
> #
> #Calculating the global model
> #
> pls.fib = plspm(data.fib, inner.fib, outer.fib, modes.fib)
```

Computed the global model, we can start the PATHMOX analysis by specifying the scale (e.g., binary, ordinal, or nominal) of the segmentation variables and by fixing the parameters of the function `pls.pathmox()`. We have decided to establish a value of 0.05 (`signif=0.05`) for the threshold of the p-value to look for those partitions that are highly significant. Given that we have a total sample of 147 students, we fix, as reasonable minimum number to stop the growth of a node, the 20% of total sample (`size=0.2`). The depth level = 1 (`deep=1`), has been selected with the aim to obtain a simple segmentation tree with a possible maximum number of two final segments. We indicate the scale of our factors by the code:

```
> #Identifying the segmentation variables
> #
> segvar = fibtele[,2:11]
> #
> #Rescaling the segmentation variables
> #
> segvar$Age = factor(segvar$Age, ordered=T)
> #
> segvar$Salary = factor(segvar$Salary,
+ levels=c("<18k","25k","35k","45k",">45k"), ordered=T)
> #
> segvar$Accgrade = factor(segvar$Accgrade,
+ levels=c("accnote<7","7-8accnote","accnote>8"), ordered=T)
> #
> segvar$Grade = factor(segvar$Grade,
+ levels=c("<6.5note","6.5-7note","7-7.5note",">7.5note"), ordered=T)
```

Then, we can perform the PATHMOX analysis by using function `pls.pathmox()`.

```
> #Run the PAthmox analysis
> #
> fib.pls.pathmox=pls.pathmox(pls.fib,segvar,signif=0.05, deep=1,
+ size=0.2,n.node=20)

PLS-PM SEGMENTATION TREE

---------------------------------------------
Info Parameters Algorithm
  parameters Algorithm value
1     Threshold signif  0.05
2   Node size limit(%)  0.20
3     Tree depth level  1.00

---------------------------------------------
Info Segmentation Variables
          Nlevels Ordered Treatment
Career          3   FALSE   nominal
Gender          2   FALSE    binary
Age             4    TRUE   ordinal
Studying        2   FALSE    binary
Contract        3   FALSE   nominal
Salary          5    TRUE   ordinal
Firmtype        2   FALSE    binary
Accgrade        3    TRUE   ordinal
Grade           4    TRUE   ordinal
Startwork       2   FALSE    binary
```

Again, we obtain, as in the case of the function `reg.pathmox()`, the algorithm parameters and a summary of the segmentation variables as first output. We use the function `plot()` to provide a graphic visualization of the tree.

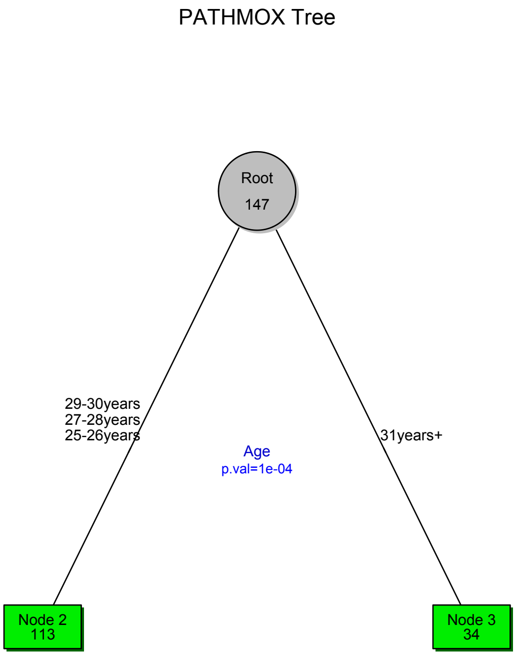


Figure 7.4: Output function `plot()` for the `fib.pls.pathmox` object. PATHMOX segmentation tree

In Figure (7.4), we present the obtained tree, where we can observe that in fact, there are two distinct models. PATHMOX splits the root node in two partitions (the node two and three) by the segmentation variables age of alumni; in the node two we have the *"younger alumni"* and in the node three the *"older alumni"*. The produced split is highly significant, giving a F statistic of *4.679* with a p-value of *0*. As in the case of regression example, we can run the function `summary()` to visualize all the information about he obtained tree.

```
> summary(fib.pls.pathmox)

PLS-PM SEGMENTATION TREE

---------------------------------------------
Info Parameters Algorithm:
  Parameters Algorithm value
1     Threshold signif  0.05
2   Node size limit(%)   0.2
3     Tree depth level     1
---------------------------------------------
Info Tree:
        Parameters Tree value
```

```
1            Deep tree     1
2 Number terminal nodes     2
---------------------------------------------
Info nodes:
  Node Parent Depth Type Terminal Size Percent Variable
1    1      0     0 root       no  147  100.00     <NA>
2    2      1     1 leaf      yes  113   76.87      Age
3    3      1     1 leaf      yes   34   23.13      Age

                          Category
1                             <NA>
2 25-26years/27-28years/29-30years
3                         31years+
---------------------------------------------
Info Splits:

Variable:
  node variable                          g1.mod   g2.mod
1    1      Age 25-26years/27-28years/29-30years 31years+

F.statistic global:
  node fg.statistic fg.pvalue
1    1       4.6797 5.584e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F.statistic block:
$fbtable
          Value    Satis
node 1 9.309329 1.207548

$pfbtable
              Value     Satis
node 1 6.908909e-06 0.3077828

F.statistic coefficient:

Node 1 :
               fc.statistic fc.pvalue
int -> Value         0.9172    0.3390
Image -> Value      17.9325 3.108e-05 ***
Qual -> Value       23.7933 1.802e-06 ***
int -> Satis         1.8905    0.1702
Image -> Satis       0.1514    0.6975
Qual -> Satis        0.0056    0.9404
Value -> Satis       0.8007    0.3716
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `summary()` function presents the same structure of the `summary()` function implemented for the `reg.pathmox()`. In order we visualize: the algorithm parameters, the tree structure, the nodes' information, and the obtained splits. In the context of PLS-PM, we can go further the F-global test by using *F-block* and *F-coefficient* test . The F-block allows us to detect which are the equations responsible of the split; the F-coefficient statistic allows us to identify the responsible predictors of the obtained partitions. Starting by the F-block we can see that the difference identified by PATHMOX between *"younger"* and *"older alumni"* depends on the Value construct (p-value equal to zero); looking for the responsible path coefficients of the split PATHMOX identifies the Image on Value and Quality on Value; in both cases the *F-coefficient* test gives p-values highly significant (i.e. equal to 0). We can interpret these results; the difference between the two PLS-PM of the terminal nodes depend on two path coefficients: Image on Value and Quality on Value, whereas the satisfaction doesn't vary across the models.

The last part of the analysis consists in the comparison between the terminal nodes of the tree. We can realize this last step by the function `pls.treemodel()`. This function allows us to visualize the most important results of the analysis for each one of the identified sub-models. The function `pls.treemodel()` also provides the invariance test that allows to verify the factor invariance, i.e. equality of the measurement model across nodes. Calling the function we can appreciate the considered inner model (`$IDM`), the invariance test (`$invariance.test`), the weights (`$weights`), the loadings (`$loadings`), the path coefficients (`$paths`), the R^2 (`$r2`) and the significance of the path coefficients assesed by t-test (`$sign`).

Again, we can use the code `label=TRUE` and `label.nodes` to give a specific name to each node of the tree. We start by comparing the the sub-models and we finished interpreting the invariance test.

```
> fib.pls.treemodel=pls.treemodel(pls.fib,fib.pls.pathmox,labe=TRUE,
+ label.nodes=c("Global model","younger alumni","older alumni"))
> #
> fib.pls.treemodel

$IDM
      Image Qual Value Satis
Image     0    0     0     0
Qual      0    0     0     0
Value     1    1     0     0
Satis     1    1     1     0

$invariance.test
  chisq.statistic p.value dfH0 dfH1
1        2.583691       1  565  542

$weights
      global model younger alumni older alumni avg.weights
ima1         0.194          0.196        0.194       0.199
ima2         0.146          0.136        0.161       0.140
ima3         0.181          0.184        0.168       0.187
ima4         0.153          0.160        0.149       0.154
ima5         0.145          0.163        0.118       0.152
ima6         0.139          0.125        0.154       0.134
ima7         0.140          0.145        0.123       0.139
ima8         0.159          0.173        0.128       0.162
quaf1        0.438          0.412        0.561       0.432
quaf2        0.379          0.415        0.237       0.396
quaf3        0.420          0.428        0.374       0.413
val1         0.366          0.424        0.210       0.390
val2         0.337          0.347        0.310       0.327
val3         0.291          0.230        0.406       0.271
val4         0.281          0.274        0.321       0.290
sat1         0.500          0.526        0.439       0.513
sat2         0.309          0.358        0.237       0.346
sat3         0.288          0.287        0.272       0.287
sat4         0.203          0.167        0.253       0.166

$loadings
      global model younger alumni older alumni
ima1         0.804          0.789        0.845
ima2         0.812          0.791        0.860
ima3         0.875          0.877        0.883
ima4         0.718          0.683        0.823
ima5         0.727          0.727        0.749
ima6         0.799          0.786        0.825
ima7         0.811          0.797        0.841
ima8         0.793          0.780        0.839
quaf1        0.732          0.680        0.882
quaf2        0.847          0.845        0.873
quaf3        0.851          0.863        0.799
val1         0.778          0.796        0.744
val2         0.863          0.880        0.776
val3         0.764          0.708        0.864
val4         0.718          0.709        0.784
sat1         0.774          0.785        0.787
sat2         0.835          0.834        0.843
sat3         0.787          0.736        0.867
sat4         0.633          0.457        0.867

$paths
              global model younger alumni older alumni
Image->Value         0.320          0.098        0.857
Image->Satis         0.290          0.238        0.231
Qual->Value          0.210          0.473       -0.334
Qual->Satis          0.117          0.149        0.171
Value->Satis         0.575          0.608        0.582

$r2
      global model younger alumni older alumni
Image        0.000          0.000        0.000
Qual         0.000          0.000        0.000
Value        0.235          0.295        0.467
Satis        0.681          0.712        0.691
```

```
$sign
$sign$Value
          global model younger alumni older alumni
Intercept        1.000          1.000        1.000
Image            0.001          0.364        0.000
Qual             0.031          0.000        0.066

$sign$Satis
          global model younger alumni older alumni
Intercept        1.000          1.000         1.00
Image            0.000          0.001         0.21
Qual             0.069          0.049         0.24
Value            0.000          0.000         0.00


attr(,"class")
[1] "treemodel"
```

In order to provide a visual way to appreciate the differences among segments, we can use the `plot` function. It provides the differential models in each leaf corresponding to satisfaction and value intangibles (see Figure (7.5) and Figure (7.6)). For both latent constructs we represent its global model and the difference in the paths coefficients of the local model respect to the global one in each terminal node.

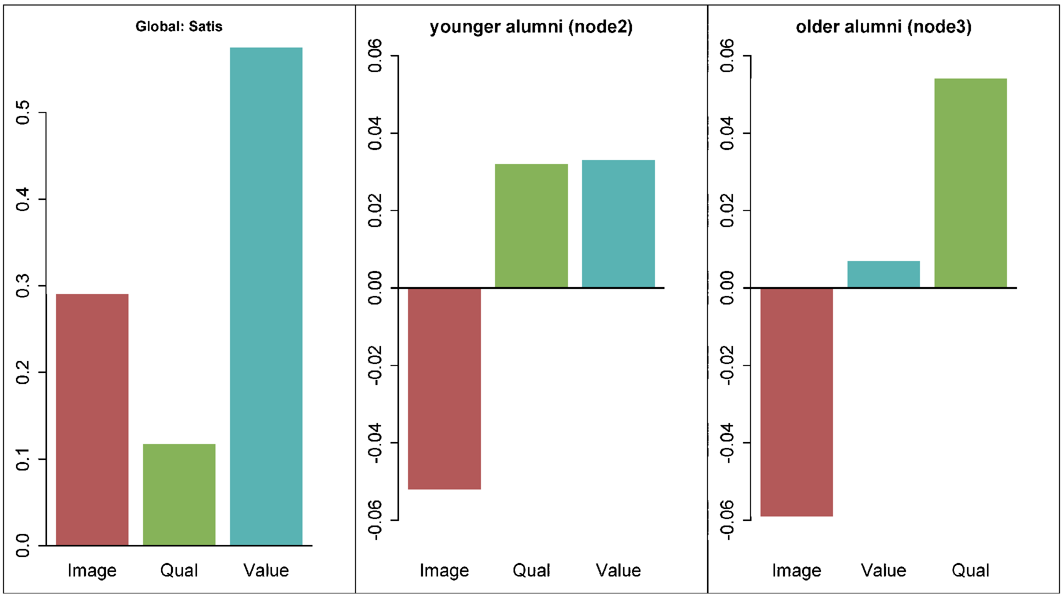


Figure 7.5: Output function `plot()` for the `fib.treemodel` object. Effects on Satisfaction: comparison of the three segments with respect to the global model

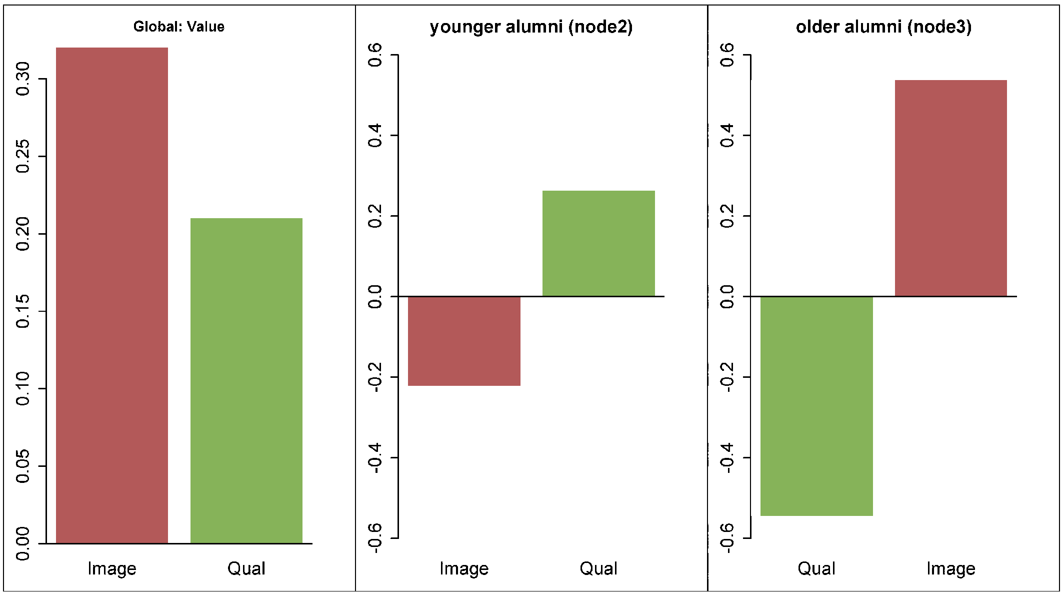


Figure 7.6: Output function `plot()` for the `fib.treemodel` object. Effects on Vale: comparison of the three segments with respect to the global model

Starting by the satisfaction we can see that there is no difference between the node 2 (*"younger alumni"*) and the node 3 (*"older alumni"*). The models are similar each others and both similar to the global model. Furthermore, the scale of the bar-chart vary from -0.04 to +0.06 confirming that there is no difference respect the global model and F-coefficients p-values of the satisfaction drivers, analyzed before, remark the absence of difference. In the case of Value, confirming the F-coefficients results, we can note two important difference: starting by the node two, we can observe that the most important effect is the Quality, whereas, in the case of the alumni with more than 30 years (node 3) the most important effect on Value is the Image. Looking the t-test (`fib.pls.treemodel`*sign*`Value`) we appreciate that both path coefficients are significant.

Finally, once we have established the existence of model heterogeneity at structural level, we consider whether the measurement model is invariant across the detected terminal nodes or they are specific for every segment by the code `fib.pls.treemodel$invariance.test`. The obtained result is a Chi-square statistic of 14.4950, giving a p-value of 0.9118, which is largely non significant, hence we can admit a common measurement model in the two leaves, as the invariant weights obtained by fitting the corresponding null hypothesis. When a common measurement model is admitted, the function `pls.treemodel()` also provides the common weights:

```
> fib.pls.treemodel$weights[,4]
```

```
 ima1  ima2  ima3  ima4  ima5  ima6  ima7  ima8 quaf1 quaf2 quaf3  val1  val2
0.199 0.140 0.187 0.154 0.152 0.134 0.139 0.162 0.432 0.396 0.413 0.390 0.327
 val3  val4  sat1  sat2  sat3  sat4
0.271 0.290 0.513 0.346 0.287 0.166
```

Bibliography

[1] Akaike, H. (1974) A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6): 716-723;

[2] Akaike, H. (1978a) A new look at the Bayes procedure. Biometrika 65, 53-59;

[3] Akaike, H. (1978b) A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics 30, 9-14;

[4] Alexopoulos, G.S., Abrams, R.C., Young, R.C., *et al.* (1988) Cornell scale for depression in dementia. Biological Psychiatry 23(3), pp. 271-284;

[5] Aluja, T., and Morineau, A. (1999) Aprender de los Datos: El Análisis de Componentes Principales - Una Aproximación desde el Data Mining. Barcelona: EUB;

[6] Aluja, T., Lamberti, G., Sánchez, G. (2013) Extending the PATHMOX app.roach to detect which constructs differentiate segments. In Abdi, H., Chin, W. W., Esposito Vinzi, V., Russolillo, G., and Trinchera, L. (Eds.). New Perspectives in Partial Least Squares and Related Methods, Springer;

[7] Aluja, T., Lamberti, G., Sánchez, G. (2013) Modelling with heterogeneity. In: Proceedings in SIS 2013;

[8] Apté, C., and Weiss, S. (1997) Data mining with decision trees and decision rules. Future Generation Computer Systems, 13(2-3): 197-210;

[9] Brandmaier, A. M., Oertzen, T. V., McArdle, J. J., & Lindenberger, U. (2013) Structural equation model trees. Psychological Methods 18, 71-86;

[10] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984) Classification and Regression Trees. California: Chapman & Hall;

[11] Brown, R. M., Mazzarol, T. W. (2009) The importance of institutional image to student satisfaction and loyalty within higher education. Higher education. Springer Netherlands. 58(1): pp. 81-95;

[12] Birkes, D. and Dodge, Y. (1993) Alternative Methods of Regression, Wiley;

[13] Chow, G.C. (1960) Test of equality between sets of coefficients in two linear regressions. Econometrica, 28(3): 591-605;

[14] Dastoor, M. C. D. (1975) Mini-mental state, a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 12(3), pp. 189-198;

[15] Dastoor, D., Cole, M. (1988) Age related patterns of decline in dementia as measured by the hierarchic dementia scale (HDS). American Journal of Alzheimer's Disease and Other Dementias 3(6), pp. 29-35;

[16] Diamantopoulos, A., and Winklhofer, H. (2001) Index Construction with Formative Indicators: An Alternative to scale development. Journal of Marketing Research, 38(2): 269-278;

[17] Dobson, A. (1990) An Introduction to Generalized Linear Models. CRC Press;

[18] Cassel, C., Hackl, P., and Westlund, A.H. (1999) Robustness of partial least squares method for estimating latent variable quality structures. Journal of Applied Statistics, 26(4): 435-446;

[19] Cassel, C.M., Hackl, P., and Westlund, A.H. (2000) On measurement of intangible assets: a study of robustness of partial least squares. Total Quality Management, 11(7): 897-907;

[20] Cattell, R. B. (1978). The scientific use of factor analysis in behavioral and life sciences. New York: Plenum;

[21] Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, 829-836;

[22] Chin, W.W., Marcolin, B.L., and Newsted, P.R. (1996) A Partial Least Squares Latent Variable App.roach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and Voice Mail Emotion/Adoption Study. In: Proceedings of the Seventeenth International Conference on Information Systems, 21-41. J.I. DeGross, S. Jarvenpaa, A. Srinivasan (Eds);

[23] Chin, W.W. (1998) The partial least squares app.roach to structural equation modeling. In: Modern Methods for Business Research, 295-336. Marcoulides, G.A. (Ed). London: Lawrence Erlbaum Associates;

[24] Chin, W.W., and Newsted, P.R. (1999) Structural Equation Modeling Analysis with Small Samples using Partial Least Squares. In: Statistical Strategies for Small Sample Research, 307-341. Hoyle R. (Ed). London: Sage Publications;

[25] Chin, W.W. (2000) Frequently Asked Questions - Partial Least squares PLS-Graph. Available from http://disc-nt.cba.uh.edu/chin/plsfaq/plsfaq.htm;

[26] Chin, W.W. (2003) A Permutation Based Procedure for Multi-Group Comparison of PLS Models. In: Proceedings of the PLS'03 International Symposium;

[27] Chin, W.W., Marcolin, B.L., and Newsted, P.R. (2003) A Partial Least Squares Latent Variable App.roach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and Voice Mail Emotion/Adoption Study. Information Systems Research, 14(2): 189-217;

[28] Duarte, P. O., Raposo., M. B. and Alves, H. B. (2012) Using a Satisfaction Index to Compare StudentsÕ Satisfaction During and After Higher Education Service Consumption, Tertiary Education and Management, 18(1): 17-40;

[29] Escofier, B., and Pagés, J. (1998) Analyses Factorielles Simples et Multiples: Objectifs, Methods et Interpretation. Paris: Dunod;

[30] Eskildsen, J., Martensen, A., Gronholdt, L., and Kristensen, K. (1999) Benchmarking student satisfaction in higher education based on the ECSI methodology. In: Proceedings of the TQM for Higher Education Institutions Conference: Higher Education Institutions and the Issue of Total Quality, Verona, 30-31 August, pp. 385-402;

[31] Fornell, C., and Bookstein, F.L. (1982) Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. Journal of Marketing Research XIX: 440-452;

[32] Ledoit, O. and M. Wolf. (2004) Honey, I Shrunk the Sample Covariance Matrix. Journal of Portfolio Management, 30(4): 110-119;

[33] Ghilagaber, G. (2004) Another Look at Chow's Test for the Equality of Two Heteroscedastic Regression Models. Quality & Quantity, 38: 81-93;

[34] Goodhue, D., Lewis, W. and Thompson, R. (2006) PLS, Small Sample Size, and Statistical Power in MIS Research. In: Proceedings of the 39th Hawaii International Conference on System Sciences;

[35] Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338;

[36] Grace, D., Weaven, S., Bodey. K., Ross, M., Weaven, K. (2012) Putting student evaluations into perspective: The Course Experience Quality and Satisfaction Model (CEQS). Studies in Educational Evaluation 38 35-43;

[37] Hahn, C., Johnson, M.D., Herrmann, A., and Huber, A. (2002) Capturing Customer Heterogeneity Using a Finite Mixture PLS Approach. Schmalenbach Business Review, 54, 243-269;

[38] Hanafi, M. (2007) PLS Path modelling: computation of latent variables with the estimation mode B. Computational Statistics, 22: 275?292;

[39] Henseler, J. and Fassot, G. (2005) Testing Moderating Effects in PLS Path Models. In: Proceedings of the PLSÕ05 International Symposium, T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&Go, 371-377;

[40] Henseler, J. (2007) A New and Simple App.roach to Multi-Group Analysis in Partial Least Squares Path Modeling. In: H. Martens and T. Naes. (Eds.), In: Proceedings of the PLSŠ07 International Symposium, Matforsk, As, Norway, 104-107;

[41] Hotelling, H. (1933) Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24: 417-441;

[42] Jackson, J.E. (1991) A User's Guide to Principal Components. New York: John-Wiley & Sons;

[43] Jeffers, J. N. R. (1967) Two Case Studies in the Application of Principal Component Analysis. Applied Statistics, 16, 225-236;

[44] Jolliffe, I.T. (2002) Principal Component Analysis. Second Edition. New York: Springer;

[45] J*ö*reskog, K.G., and Wold, H. (1982) The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects. In: Systems under indirect observation: Causality, structure, prediction. Part I, 263-270. K.G. J*ö*reskog & H. Wold (Eds). Amsterdam: North Holland;

[46] Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 129(2): 119-127;

[47] Kaiser, H. F. (1958) The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187-200;

[48] Kerlin, K. (2000) Measuring student satisfaction with the service process of select student educational supp.ort services at Everett Community College (PhD dissertation). Oregon State University;

[49] Koenker, R. and Bassett, G. (1978) Regression quantiles, Econometrica, 46, 33-50;

[50] Koenker R. and Bassett, G. (1982) Robust Tests for Heteroscedasticity Based on Regression Quantiles. In: Econometrica Vol. 50, No. 1, pp. 43-61;

[51] Koenker, R.W., and d'Orey (1987, 1994) Computing regression quantiles. Applied Statistics, 36, 383-393, and 43, 410-414;

[52] Koenker, R., and Hallock, K. (2001) Quantile Regression. Journal of Economic Perspectives-Volume, 15(4),mpp.143-156;

[53] Koenker, R., (2005) Quantile Regression. Cambridge U. Press;

[54] Jöreskog, K. G. (1971) Simultaneous factor analysis in several populations. Psychometrika, 36(4), 408-426;

[55] Lebart, L., Morineau, A., and Féenelon, J.P. (1979) Traitement des donneés statistiques. Paris. Dunod;

[56] Lebart, L., Morineau, A., and Fénelon J.P. (1985) Tratamiento estadístico de datos. Barcelona: Marcombo;

[57] Lebart, L., Morineau, A., and Piron, M. (2004) Statistique Exploratoire Multidimensionnelle. Paris: Dunod;

[58] Lubke, G.H., and Mutheń, B. (2005) Investigating Population Heterogeneity with Factor Mixture Models. Psychological Methods, 19(1): 21-39;

[59] Lohmöller, J. B. (1989) Latent Variable Path Modeling with Partial Least Squares. Heidelberg: Physica-Verlag;

[60] McCullagh, P., and Nelder, J. A. (1989) Generalized Linear Models, 2nd ed. London: Chapman and Hall;

[61] Moreno, E., Torres, F., and Casella, G. (2005) Testing equality of regression coefficients in heteroscedastic normal regression models. Journal of Statistical Planning and Inference, 131: 117-134;

[62] Martensen, A., Gronholdt L., Eskildsen J.K., Kristensen, K. (2000) Measuring student oriented quality in higher education: application of the ECSI methodology. Sinergie Rapp.orti di Ricerca 9: 372-383;

[63] Nelder, J.A. and Wedderburn R.W.M. (1972) Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370-384;

[64] Palumbo, F., and Romano, R. (2008) Possibilistic PLS Path Modeling: A New App.roach to the Multigroup Comparison. In: Proceedings in Computational Statistics, 303-314. Paula Brito (Ed), Heidelberg: Physica-Verlag;

[65] Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2: 559-572;

[66] Pompei,P., Foreman, M., Cassel, C. K. , Alessi, C., Cox,D. (1995) Detecting delirium among hospitalized older patients. JAMA Intrnal medicine 155(3), pp. 1301-307;

[67] Quinlan, J.R. (1986) Induction of Decision Trees. Machine Learning, 1: 81-106;

[68] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. California: Morgan Kauffman;

[69] Quinlan, J.R. (1998) C5/See5 Software;

[70] Rao, C.R. (1964) The Use and Interpretation of Principal Component Analysis in App.lied Research, Sankhya A, 26, 329-358;

[71] Ringle C.M., Wende S., and Will, A. (2005) Customer Segmentation with FIMIX-PLS. In: Proceedings of the PLS'05 International Symposium, T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&Go, 507-514;

[72] Ringle, C.M., and Schlittgen, R. (2007) A Genetic Algorithm Segmentation App.roach for Uncovering and Separating Groups of Data in PLS Path Modeling. In: H. Martens and T. Naes. (Eds.), Proceedings of the PLS'07 International Symposium, Matforsk, As, Norway, 75-78;

[73] Reinartz, W.J., Echambadi, R., and Chin, W.W. (2002) Generating Non-normal Data for Simulation of Structural Equation Models Using Mattson's Method. Multivariate Behavioral Research, 37(2): 227-244;

[74] Sánchez, G., and Aluja, T. (2006) PATHMOX: A PLS-PM Segmentation Algorithm. In: Electronic Proceedings of the Workshop on Knowledge Extraction and Modeling (KNEMO), V. Esposito Vinzi, C. Lauro, A. Braverman, H.A.L. Kiers, M.G. Schimek (Eds);

[75] Sánchez, T. (2007) A Simulation Study of PATHMOX (PLS Path Modeling Segmentation Tree) Sensitivity. In: H. Martens and T. Naes. (Eds.), Proceedings of the PLS'07 International Symposium, Matforsk, As, Norway, 33-36;

[76] Sánchez, G.,(2009) PATHMOX Approach: Segmentation Trees in Partial Least Squares Path Modeling. Doctoral Dissertation. http://gastonsanchez.com/thesis/pathmox-approach-thesis-gaston-sanchez.pdf

[77] Sonquist, J.A., and Morgan, J.N. (1964) The Detection of Interaction Effects. Institute for Social Research, University of Michigan;

[78] Sonquist, J.A., Baker, E.L., and Morgan, J.N. (1971) Searching for Structure. Institute for Social Research, University of Michigan;

[79] Saporta, G. (2006) Probabilités, analyse de données et statistique. Paris: Editions Technip;

[80] S*ö*rbom, D. (1974) A general method for studying differences in factor means and factor structures between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239;

[81] Schwarz, Gideon, E. (1978) Estimating the dimension of a model. Annals of Statistics 6 (2): 461-464;

[82] Serenko, A. (2011) Student satisfaction with Canadian music programmes: the app.lication of the American Customer Satisfaction Model in higher education, Assessment & Evaluation in Higher Education, 36:3, 281-299

[83] Temizer, L., Turkyilmaz, A. (2012) Implementation of Student Satisfaction Index Model in Higher Education Institutions, Procedia - Social and Behavioral Sciences,46, pp. 3802-3806;

[84] Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., and Lauro, C. (2005) PLS path modeling. Computational Statistics and Data Analysis, 48, 159-205;

[85] Tenenhaus, M. (2006) Statistique. Paris: Dunod;

[86] Tenenhaus, M., Mauger, E., and Guinot, C. (2006) Test of a group effect in a regression model relating two blocks of binary variables with ULS-SEM and SEM-PLS. In: Electronic Proceedings of the Workshop on Knowledge Extraction and Modeling (KNEMO), V. Esposito Vinzi, C. Lauro, A. Braverman, H.A.L. Kiers, M.G. Schimek (Eds);

[87] Tenenhaus, M., Hanafi, M. (2010) A bridge between PLS path modeling and multi?block data analysis. Handbook of Partial Least Squares. Esposite Vinzi et al (eds);

[88] Tenenhaus, A., Tenenhaus, M. (2011) Regularized Generalized Canonical Correlation Analysis, Psychome- trika, 76 (2), pp. 257-284;

[89] Tenenhaus, A., Tenenhaus, M. (2014) Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. European Journal of Operational Research 238(2): 391-403;

[90] Timm, N.H., (2002) Applied Multivariate Analysis. New York: Springer-Verlag;

[91] Trinchera, L., Squillacciotti, S., Esposito Vinzi, V., and Tenenhaus, M. (2007) PLS path modeling in presence of a group structure: REBUS-PLS, a new response-based approach. In: H. Martens and T. Naes. (Eds.), Proceedings of the PLS'07 International Symposium, Matforsk, As, Norway, 79- 82;

[92] Trinchera, L., Balzano, S. (2011) Structural Equation Models and Student Evaluation of Teaching: A PLS Path Modeling Study. In: M. Attanasio and V. Carputi. (Eds.), Statistical Methods for the Evaluation of University Systems;

[93] Tryon, R.C. (1939) Cluster analysis correlation profile and orthometric analysis for the isolation of unities in mind and personality. Ann Arbor, Mich., Edwards brother, Inc., lithoprinters and publishers;

[94] Thurstone, L. L. (1947) Multiple-factor analysis. Chicago: University of Chicago Press, pp. 535;

[95] Esposito Vinzi, V., Ringle, C.M., Squillacciotti, S., and Trinchera, L. (2007) Capturing and Treating Unobserved Heterogeneity by Response Based Segmentation in PLS Path Modeling: A Comparison of Alternative Method;

[96] Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., and Tenenhaus, M. (2008) REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. Applied Stochastic Models in Business and Industry, 24-5, pp. 439-458. John Wiley & Sons, (Ltd);

[97] Wold, H. (1975a) PLS path models with latent variables: the nipals app.roach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Cappecchi (Eds.), Quantitative sociology: international perspectives on mathematical and statistical modeling. New York: Academic Press;

[98] Wold, H. (1975b) Modelling in complex situations with soft information. Third World Congress of Econometric Society, Toronto, Canada;

[99] Wold, H. (1975c) Soft modeling by latent variables: the nonlinear iterative partial least squares app.roach. In J. Gani (Ed.), Perspectives in probability and statistics, papers in honor of M. S. Bartlett (pp. 117-142). London: Academic Press;

[100] Wold, H. (1980) Model construction and evaluation when theoretical knowledge is scarce. In J. Kmenta, & J. B. Ramsey (Eds.), Evaluation of econometric models, pp. 47-74;

[101] Wold, H. (1982) Soft modeling: the basic design and some extensions. In K. G. J*ö*reskog, and H. Wold, (Eds.), Systems under indirect observation, Part II (pp. 1-54). Amsterdam: North-Holland.

[102] Wold, H. (1985) Partial least squares. In S. Kotz, and N. L. Johnson, (Eds.), Encyclopedia of Statistical Sciences, Vol. 6 (pp. 581-591). New York: Wiley;

[103] Westlund, A.H., Cassel, C.M, Ekl*ö*f, J., and Hackl, P. (2001) Structural analysis and measurement of customer perceptions, assuming measurement and specifications errors. Total Quality Management, 12(7& 8): 873-881;