UNIVERSITÀ DI ROMA TRE
Dipartimento di Economia

Tesi di dottorato in
Metodi Statistici per l'Economia e l'Impresa

Ciclo XXVI

Assessing Sapienza University alumni job careers: Enhanced Partial Least Squares latent variable path models for the analysis of the UNI.CO administrative archive

Candidato
Francesca Petrarca

Supervisori
Prof.ssa Silvia Terzi
Prof. Giorgio Russolillo

Coordinatore: Prof.ssa Julia Mortera

Anno 2014

Acknowledgements

Dalla scrittura della tesi specialistica sono passati un pò di anni nei quali ho dovuto affrontare situazioni che mi hanno messo a dura prova e che mi hanno aiutato a far "crescere le spalle" e ad andare avanti; quindi come prima cosa ringrazio tutti quelli che mi hanno aiutato a superare questi momenti e anche quelli che hanno reso questi momenti più difficili.

È difficile in poche righe ricordare tutte le persone che, a vario titolo, hanno contribuito a rendere "migliori" questi anni. In questi i tre anni del Corso di Dottorato, ho avuto la fortuna di conoscere e collaborare con persone speciali, che desidero ringraziare per tutto il loro sostegno e per il contributo alla realizzazione di questo lavoro.

Grazie innanzitutto alla prof.ssa Silvia Terzi, tutor interno della tesi, per la sua disponibilità e per i suoi consigli. Un affettuoso ringraziamento va poi alla prof.ssa Julia Mortera, coordinatrice del corso di dottorato, per aver creduto in me.
Un ringraziamento ai colleghi di dottorato presenti e passati per il costante confronto e per i suggerimenti ricevuti. Desidero ringraziare specialmente la dott.ssa Vincenzina Vitale per suoi preziosi consigli e per la sua pazienza e disponibilità e la dott.ssa Flaminia Musella per il sostegno scientifico e amicale. Spero che queste amicizie restino vive anche al di fuori dell'ambito universitario...
Desidero poi ringraziare immensamente il prof. Giorgio Alleva, che ha sempre creduto in me e che con disponibilità e generosità è sempre stato pronto ad aiutarmi con suggerimenti, spunti e interessanti chiacchierate. È stato un piacere lavorare con lei.

Desidero ringraziare il dott. Giorgio Russolillo, per avere accolto la mia disperata richiesta di aiuto e per essere entrato a far parte nel mio progetto di tesi. Lui e la sua famiglia sono e resteranno (lo spero) un punto di riferimento morale e accademico...Ed ancora, desidero ringraziare il prof. Gilbert Saporta per avermi invitato e ospitato a Parigi per lavorare insieme a Giorgio presso il *CNAM- Conservatoire National des Arts and Metiers*.
Desidero ringraziare anche il dott. Gaston Sanchez per la sua disponibilità per l'utilizzo del nuovo package di R.
Ed ancora... Ringrazio il gruppo UNI.CO per avermi coinvolto in questo progetto..in particolare la dott.ssa Eleonora Renda, compagna di avventura (o di sventura in alcuni momenti)..
Un ringraziamento va anche a i miei nonni che, anche se non ci sono più, il solo pensiero di poterli far felici mi ha dato la forza di andare avanti per farli essere fieri di me... ma anche alla nonna, che e' ancora tra di noi, e che ogni giorno lotta per restarci..
Ringrazio anche le "nuove" persone che sono entrate a far parte della mia famiglia...

Un ringraziamento particolare va a David per essermi sempre vicino e per dividere con me questa vita, per aver sempre sopportato sia bei momenti ma anche quelli difficili. Non ho mai incontrato nella mia vita una persona così dolce, sensibile, affettuosa e sempre presente con me. È bello svegliarsi ogni giorno con te accanto....

Ed infine....un grazie di cuore anche ai miei genitori e mio fratello che, hanno sempre mostrato, sia per me ma anche per quello che faccio una fiducia cieca e priva di incertezze, spronandomi sempre ad andare avanti per la mia strada.

....Grazie mille...

Francesca

Contents

List of Figures

List of Tables

Introduction

A model is a simplified representation of reality derived by experimental observations as well as by logical deductions. It expresses a complex reality in a parsimonious way, making efficient use of the available information.
Simple models are not suitable to represent complex human behaviors because their intrinsic limitation in taking into account the relationships among the variables. This is the reason, along with computer-science development, for the increase in the use of Structural Equation Models (SEM), Bollen (1989)[7], Kaplan (2000)[41]. As a matter of fact, in Structural Equation Models the real word complexity can be studied taking into account a whole number of causal relationships among latent concepts (i.e. the Latent Variables (LVs)), each measured by several observed indicators usually defined as Manifest Variables (MVs).
The job quality is a popular topic of discussion in any country and the subject of the research for many international organizations. The European Commission, in the final reports of its Stockholm and Nice Meetings held in 2001, advocates that the quality of jobs should rise continuously in order to enhance the competitiveness and productivity of the European economic system. This purpose was originally stated in Lisbon in the years 2000 (and 2010), when the hope was expressed that work should be developed not just quantitatively but also qualitatively. Job quality is a complex and multidimensional concept that involves the characteristics both of the job itself and of the worker, Commission of the European Communities (2008) [24]. In the seventies a good job could be defined as *well-paid, secure and connected to paths of upward mobility* and on the contrary a bad job as *low-paid, unstable and lead to an end*, Tilly (1996)[67]. Those definitions may still be valid today at least partially. As a matter of fact, an increasing importance is also given to other job

dimensions, such as the quality of the workplace, the perspective of the fulfillment and the social role achievable through the job.
Anyway it is generally recognized that measuring the job quality is not easy either in absolute or in relative terms. Jencks et al. (1988) [39] suggested an *index of job desirability*, which merged both monetary and non-monetary indicators, among them the risk of being fired, job features and the required level of education are considered.
Clark (1998) [17] enumerated seven dimensions of a good job: earnings, working hours, career chances, environmental safety, practical difficulties, job satisfaction and interpersonal relations. The author, merging seven aspects as if they were independent, defined two indices of job quality, the overall satisfaction and the job quality index. More recently, The European Trade Union Institute, Leschke and Watt (2008) [46] and Leschke et al. (2008) [47] proposed a *Job quality index* based on six sub-dimensions, namely earnings, non-standard work contracts, possibility to conciliate private and working lives, working environments and safety, career development and competence improvement and collective interest representation, see European Commission (2008)[24].
ISTAT, the Italian Statistical Institute, examines the quality of the initial jobs by collecting data every three years on Italian high school graduates and on university graduates. Quality may be defined according to both the objective feature of a job (as type of work contract, safety of the workplace, professional position and income) and to the individual satisfaction offered by various aspects of the job.
The search of a possible correspondence between education and labour implies the ascription of a social value, as opposed to just individual value, to the university education. Employability and successful career attainments should be in the background of the university aims together with that of developing culture and civic spirit of students.
For several years, the Italian universities have been urged to better understand the characteristics of labour demand for their alumni. For this purpose periodic reports of the results of sample surveys are produced by institutions (ISTAT) or private entities. These studies are detected directly from the alumni and their employment status after graduation (the most common sample surveys coming from

Inter-University Consortium *AlmaLaurea*). These results constitute important information for students, families and universities. The sample surveys are based on studies that take place at a distance of six months and of two years from the date of graduation. The alumni are interviewed regarding their satisfaction as well as to obtain information on their employment situation. It must be emphasized that spontaneous opinions can express interesting points of view after some work experiences (even unsuccessful ones). Obviously, all ex post assessments provide us with a lot of information to define overall judgements regarding the educational career but also their employment status.

Of course, the definition of a graduate's employment status is not the unique. With regards to Italy, the ISTAT records the status of employment when the individual performs, in a reference week, at least one hour of paid work (regulated by a registered contract) regardless of the type of contract and the regularity of employment. Timing is certainly a critical element: studies too close to the end of the educational experience do not sufficiently reflect the students' rethinking of their educational lives. However, if the surveys are too far from this experience, the rethinking can itself be strongly influenced by exogenous factors. The outcome of the sample survey depends on the technique used (CATI, CAWI etc.). For example, the Computer Assisted Telephone Interviewer (CATI) technique indicates a method of direct detection of statistical units carried out through telephone interviews. The interviewer reads the questions and records the answers on a computer, using a special software. It is known that, the CATI sample surveys are usually biased, subjective and over dependent on the memory of the interviewee.

Linkage of databases could certainly be an interesting approach to overcome most of the above-mentioned problems. At a small cost we could compare the performance of persons that received a university degree with those without it (or who have had a different degree) either by simply observing the social security agency status (INPS in Italy) after a certain period or the Italian Ministry of Labour and Social Policy archive. Apart from the methodological problem of linkage, the issue here is often a "political" one, involving the willingness of different institutions. The data exist, but what is often missing is the willingness to link the data for comparative purposes. A non-trivial problem is data privacy, although it is frequently used as

an excuse.
The labour market demand for the employability skills and knowledge of Sapienza graduates can now be analysed through the new administrative archive called UNI.CO, which was built through a record-linkage between the Sapienza University of Rome archive with those of the Italian Ministry of Labour and Social Policy. The UNI.CO archive has allowed us to evaluate all the signed contracts obtained by Sapienza graduates in addition to full information on all the enterprises and institutions that offered these contracts. This archive allows us to perform a statistical analysis on more sophisticated and robust (and more reliable) details than the standard analysis based on sample surveys. In our study we have used objective real data unaffected by any emotional influence. Another important feature is that the data from the UNI.CO archive allow us to study the evolution of alumni throughout the observation period and therefore to analyse changes in job qualifications (differently from the common sample survey supply information about the status of alumni at some fixed point in time (like a photograph)).
In this thesis we adopted the UNI.CO administrative archive in order to study the integration of Sapienza University of Rome alumni into the current labour market. The general framework is the study of the subordinate and para-subordinate employment opportunities offered to the Sapienza University of Rome alumni by the Italian labour market. Different types of graduate integration into the subordinate and para-subordinate labour market are identified, and variables for the integration into job positions are defined in coherence with the university career path. In this study, we propose and discuss some variables which can influence the probability of "job success" in terms of Sapienza graduates best employment status. The evolution of the job contract type, actual job duration and professional qualifications, is studied through a longitudinal analysis based on new specific synthesis indicators. In the following we propose three models which have different purposes. The first one concerns the quantitative study of the job success of Sapienza University alumni in terms of quality of work. In order to study the job career in terms of the work contract type and quality evolution in the three years after graduation, we propose two different models one for the alumni who at the time of the master's degree had an active contract and a different one for the other alumni's category.

The scheme of the thesis is the following. In Chapter 1 a detail of the presentation of the UNI.CO archive is reported. In Section 1.1 we present a brief description of the merging of the two administrative archives. More specifically, in Section 1.1.1 we describe the information contained in the Sapienza archive (Infostud), in Section 1.1.2 the information contained in the Italian Ministry of Labour archive (SISCO) and in Section 1.1.3 we describe the UNI.CO archive obtained by merging the two sources. In Section 1.2 we propose the use of synthesis indicators. Thus, in Section 1.2.1 we describe the variables that can predict the integration of the alumni into job positions that are coherent with the educational curriculum; in Section 1.2.2 we illustrate a series of indicators on the quality of the contracts and the evolution of the type of contracts in the three years of observation. These variables concern the contract type, its actual duration and the professional qualification. Finally, in Section 1.3 we discuss the panorama that results from a general analysis of the UNI.CO archive.

In Chapter 2 latent variable path models are reviewed. In Section 2.1 the basis of the Structural Equation Models are introduced, then the focus moves to PLS-PM, Section 2.2. These models are composed by two sub-models: the structural model, described in Section 2.2.1, and the measurement model, Section 2.2.2. In Section 2.2.3 the technical phases of PLS-PM algorithm are briefly illustrated. The Chapter 2 ends up with the list of hypotheses concerning the model validation, Section 2.2.4, the discussion of the indexes of the assessment model, Section 2.2.5 and the some remarks on the optimization criteria, Section 2.2.6.

In Chapter 3 the extension of PLS-PM to the case of non-metric variables proposed by Russolillo (2012) [56] is described. This recent method, called Non-Metric PLSPM is based on the implementation of the optimal scaling approach to this field.

A brief discussion about different measurement scales is presented in Section 3.1 as a necessary preliminary issue to the NM-PLSPM. Then, in Section 3.1.1 the NM-PLSPM algorithm is discussed and in Section 3.1.2 the main characteristics of optimization criteria are stressed. In the second part of this Chapter, Section 3.2, the extension of the NM-PLSPM to binary endogenous latent variables is presented. In this thesis the NM-PLSPM has been modified to implement the logistic

regression for the computation of the path coefficients concerning latent variables measured though only one manifest variable which is binary.
A review of the logistic regression is reported in Section 3.2.1 together with the descriptive measures of the fit quality Section 3.2.2 and a presentation of the model validation techniques with a special attention on the ROC curve Section 3.2.3.
Finally, the modification applied to the NM-PLSPM necessary in order to encompass the logistic regression is described in Section 3.2.4.
Chapter 4 is devoted to the application of NM-PLSPM methods with the logistic extension to two particular analyses of the UNI.CO archive data. The first analysis concern the possibility offered by the UNI.CO archive to model the job quality reached by Sapienza alumni, Section 4.1, the discussion of the results in Section 4.1.1 and some remarks in Section 4.1.2.
The other analysis concerns the study of the job career of Sapienza alumni in the three years after graduation, Section 4.2. The career of Sapienza alumni have been studied by two different models: one for the alumni who do not have at the moment of the master's degree an active job contract, Section 4.2.1 and the discussion of the results in Section 4.2.2, and another for those with an active contract at the moment of the master's degree Section 4.2.3 and the discussion of the results in Section 4.2.4.
In Section 4.2.5 some remarks about the last two models are reported.
In Conclusions our overall remarks are discussed.
In the Appendixes, for completeness, two important issues related with the binary variables are briefly reviewed: the exponential families in Appendix A.1 and the Generalized Linear Models in Appendix A.2, the parameters estimation in Appendix A.2.1 and the numerical methods in Appendix A.2.2.

Chapter 1

The UNI.CO Archive

In the last few years Italian universities have shown an increasing interest in studying the characteristics of the labour market demand for their alumni as generated by the Italian Industry (Carpita 2011[12]). The information collected from graduate surveys has so far been of relevance to students, families and universities (Fabbris 2012 [25]). In addition periodic reports have been produced by public institutions, for example, the Italian National Institute of Statistics (ISTAT) 2012 [38] or private entities, such as CENSIS 2012 [14] and AlmaLaurea (2013) [5]. However, since March 2008, the Italian government has issued regulations for the development of administrative archives for the labour market. The trial investigation approved in 2011, establishes the integration of the Sapienza archive (Infostud) with the Italian Ministry of Labour archive, (CO, i.e. Compulsory Communications). The integration of these archives has generated a new archive called UNI.CO. The first statistical

analysis, carried out by Alleva et al. (2012) [4] [1], based on UNI.CO data, revealed a significant improvement in the retrieval of knowledge and information about the careers of the alumni, especially the years after graduation. The UNI.CO archive has allowed us to evaluate all the signed contracts obtained by Sapienza alumni in addition to full information on all the enterprises and institutions that offered these contracts. Moreover, the information on all the contracts obtained by each Sapienza graduate retrieved in the integrated database also includes data about the actual duration, the type of contract and a description of the professional qualification required for the job. The same information can be related to the sectors of economic activity, the location of businesses and, of course, information on the educational curriculum of the alumni from upper secondary school diploma to university graduation. Furthermore, this information contains details of the courses taken during the student's study program and the grades obtained in the examinations. The UNI.CO archive is therefore a powerful tool for the university and for single departments, who can use it not only to evaluate the position of alumni "at a certain date", but also to access information on the various job paths during the three year period of observation; for example on the quality of the job, the coherence with the academic studies, and on the number of permanent job positions obtained in the labour market. The UNI.CO archive also includes all the activations, terminations[2], extensions and transformations of subordinate and para-subordinate job relationships. The UNI.CO archive contains a large and complete amount of administrative data

[1] It was possible to merge these administrative archives thanks to the Convention of the 14 June 2011 between the Directorate-General for innovation, technology and communication of the Ministry of Labour and Social Policy (Dr. Maria Grazia Strano) and "Sapienza" University of Rome. The first report on Compulsory Communications archive for the study of labour demand for Sapienza alumni has been realized by the working group UNI.CO under the supervision of Giorgio Alleva and presented for the first time during the workshop "Monitoring of the dynamics of the professional alumni" of the 24th September 2012.This group is formed by researchers from Sapienza University of Rome, Italian Ministry of Labour and ItaliaLavoro. For Sapienza University: Pietro Lucisano, Carlo Magni, Silvia Massimi, Francesca Petrarca, Alessandro Sanzo, Bruno Sciarretta and Eleonora Renda. For the Italian Ministry of Labour: Daniele Lunetta and Maurizio Sorcioni and for ItaliaLavoro Giuseppe De Blasio. The workgroup was supported by a Scientific and Technical Committee of Sapienza University formed by: Giorgio Alleva, Tiziana Catarci, Rosalba Natale and Cristiano Violani. [2] The term "terminations", here, means the end of a job contract

which can allow us to build models of data interpretation through the identification of:

- categories of Sapienza alumni in relation to integration into subordinate and para-subordinate work;
- indicators that can predict the integration of alumni in job positions that are coherent with university studies.

From the point of view of official statistics, the experiment with the UNI.CO archive highlights the notable improvement provided by the integration of archives from different sources. This new information is entirely additional with respect to that provided by each of them singularly.
In this thesis work we adopted the UNI.CO administrative archive in order to study the integration of Sapienza University of Rome alumni into the subordinate and para-subordinate labour market. (For a preliminary statistical analysis based on a subset of UNI.CO see Petrarca (2013) [53] and Petrarca (2014) [54]). This archive allows us to perform a statistical analysis on more sophisticated and robust (and more reliable) details than the standard analysis based on sample surveys (Ciriaci et al. (2011) [16]; Capecchi et al. (2012) [11]). For example ISTAT provides a sample survey on the education-to-work transition that is an important source of data for a comparative analysis of the introduction into the world of work regarding different qualifications; on the whole, it provides a useful tool for evaluating the effectiveness of the higher education system. The sample surveys occur every three years; they are conducted on individual student groups about three years after graduation and they are carried out by using the Computer Assisted Telephone Interviewing (CATI) technique. This survey technique indicates a method of direct detection of statistical units carried out through telephone interviews. The interviewer reads the questions and records the answers on a computer, using a special software. Our choice of using the UNI.CO archive allowed us to avoid the usual problems that arise during the CATI sample surveys, which are usually biased, subjective and over dependent on the memory of the interviewee. In our study we have used objective real data unaffected by any emotional influence. Other sample surveys like the Labor Force (ISTAT) and AlmaLaurea, supply information about the status of alumni

at some fixed point in time (like a photograph), whereas the data from the UNI.CO archive allow us to study the evolution of alumni throughout the observation period and therefore to analyse changes in job qualifications. A drawback found in the use of the UNI.CO archive is that, for alumni who have obtained a permanent contract, the changes in job qualifications are not recorded in the CO. This fact can be overcome by integrating UNI.CO with the INPS archive, which contains information such as annual wages. This further integration would allow us to complete the position of alumni in order to investigate the integration of Sapienza alumni into the labour market. This would be a fundamental improvement for making the transition from a interpretation in terms of graduate's demand in companies and institutions (subordinate and para-subordinate), towards reaching an overall assessment of a graduate's professional careers. This makes it possible to study not only the flow of activations and terminations over time (using the CO information), but also the stock of alumni in different working positions at one or more successive circumstances after graduation. In order to expand the information contained in the UNI.CO archive, it would be interesting to integrate the ASIA (ISTAT) and "Studies Sector" archives. Another desirable improvement would be the integration of data from alumni of other universities.

The structure of this chapter is as follows. In Section 1.1 we present a brief description of the merging of the two administrative archives. More specifically, in Section 1.1.1 we describe the information contained in the Sapienza archive (Infostud), in Section 1.1.2 the information contained in the Italian Ministry of Labour archive (SISCO) and in Section 1.1.3 we describe the UNI.CO archive obtained by merging the two sources. In Section 1.2 we propose the use of synthesis indicators. Thus, in Section 1.2.1 we describe the variables that can predict the integration of the alumni into job positions that are coherent with the educational curriculum; in Section 1.2.2 we illustrate a series of indicators on the quality of the contracts and the evolution of the type of contracts in the three years of observation. These variables concern the contract type, its actual duration and professional qualification. The synthesis variables chosen allow us to quantitatively follow up the career of the alumni in terms of improvement, worsening or stability of their job position for the three years after graduation. Finally, in Section 1.3 we discuss the (positive) panorama that

results from a general analysis of the UNI.CO archive.

1.1 Data

1.1.1 The Infostud Archive

The Sapienza University administrative archive (called Infostud) has provided data on Sapienza alumni for the following calendar years 2008, 2009, 2010 and 2011, on the basis of a well-structured path divided into five sections: the personal data of the, including his/her secondary school diploma, university degree, enrollment in another degree course in Sapienza University and his/her declared income. The first section (personal data of the alumni), is made up of five areas, and contains the following information: tax code, gender, date of birth, place of birth and residence. The second section (secondary school diploma) consists of three areas: the subjects taken, the grades and the year the diploma was obtained. The third section (university degree) is divided into twelve areas concerning the university career of the alumni from enrollment to graduation (e. g. year of enrollment , college, degree, etc.). The fourth section provides information on Sapienza alumni that enrolled in other courses at Sapienza university after graduating (for example Master degree, Phd), with course specifications, the date of enrollment and the duration. The fifth and final section deals with income declared at the time of enrollment (ISEE), specifying the level and amount of the income. The scale of the information exported from the Infostud archive is heterogeneous: alphanumeric, categorical (nominal and ordinal), date-time, alphabetic and numeric. For this reason new data encoding has been carried out following the official ISTAT classification. Moreover, the early analysis based on Infostud data has highlighted the need for some "cleaning" and "normalization" of the data, inducing a series of activities aimed at:

- reducing the duplication of errors (by verifying the repetition of tax codes in the same year);

- reclassifying some variables according to the code system adopted by the Italian Ministry of Labour (in particular regarding place of birth and residence).

1.1.2 The SISCO Archive

Compulsory Communications (CO) is an archive containing the information that public and private employers are required to provide for the Italian Ministry of Labour at the start of employment or at the moment of the extension, transformation and termination of employment. In agreement with the Ministerial Decree of the 30th October, 2007, on compulsory communication from employers ("Comunicazione obbligatorie telematiche dovute dai datori di lavoro pubblici e privati ai servizi competenti"), since March 2008, all employers (individuals, businesses and public entities) are required to communicate to the appropriate departments, facts regarding the start, extension, transformation and termination of employment relationships, through a series of communication modules. The CO information system of the Ministry of Labour and Social Affairs manages the information flow of administrative data so that the data can reach the Central National Coordination office in real time. The information flow goes through specific peripheral systems that are delegated to data collection. The system is able to trace the flows linked to the regular employment relationships of subordinate and para-subordinate jobs. These flows are regulated within the current legislation concerning Italian citizens, and also include foreigners who have valid residence permits, even if in the country only temporarily. The system does not record, following current regulations, particular employment relationships, such as: judges, lawyers, state prosecutors, military personnel, the police force, the diplomatic corps, the mayor, and people in high administrative positions in public and private companies (e. g. presidents, managing directors). The transformation of the administrative information contained in the CO system, into statistical information is a complex task that has been entrusted to a special technical working group made up of experts from the Italian Ministry of Labour, ISTAT, ItaliaLavoro and ISFOL. The normalization of the archive, which is still in progress, has produced a preliminary version of the new archive called the Statistical Information System of Compulsory Communications (SISCO). It needs to be pointed out here, that all the supply job contracts communicated by Italian employment agencies and all employment relationships involving individuals registered in the 'Boat People' lists have so far been excluded from this archive. To better understand the path career of the alumni, it is worth investigating the infor-

mation content of the SISCO archive related to subordinate and para-subordinate job contracts. Taking advantage of the information contained in SISCO, we can:

a. identify the job relationships concerning Sapienza alumni for a certain time interval. In this way we are able to extract the nature, duration and professional skills involved in the contracts in order to correlate the information with academic and socio-professional profiles;

b. identify and analyse the sectors, the production units and territorial basins in which the job relationships take place;

c. monitor the effects of incentive policies (apprenticeships, internships) for the transition into the job market.

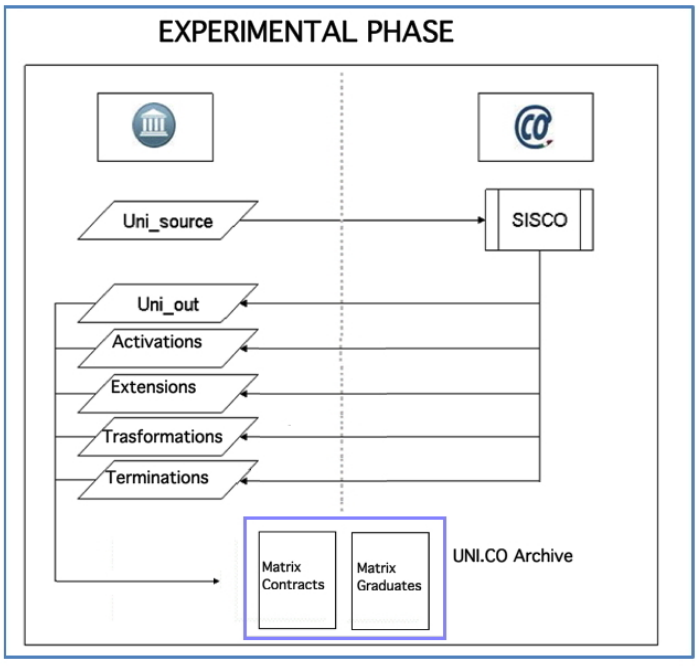
Under this perspective, the main contribution that can be guaranteed by SISCO data is a very detailed representation of the status for subordinate and para- subordinate demand offered by enterprises to Sapienza alumni. Moreover, the analysis concerning the transition processes of Sapienza alumni and post alumni takes on new connotations, which are completely original compared to other surveys. The SISCO system, (see Figure 1.1), allows the CO to combine with the archive of the alumni by merging the tax codes from each archive. The SISCO experimental phase provided Sapienza University with five unidentified files :

- a file on Sapienza University alumni (files with unidentified personal data called Uni_out);
- an activations file, i. e. a file containing subordinate and para-subordinate contracts signed by three-year degree alumni and five-year master degree alumni of Sapienza University for a certain time interval;
- extensions file, i. e. a file containing the prolongations of the subordinate and para-subordinate contracts signed by the bachelor degree and master degree alumni of Sapienza University for a certain time interval.
- transformations file, i. e. a file containing the changes of subordinate and para-subordinate contracts signed by the bachelor and master degree alumni of Sapienza University for a certain time interval;

- terminations file, i. e. a file containing the terminations of subordinate and para-subordinate contracts signed by the three year degree and master degree alumni of Sapienza University for a certain time interval.

Fig. 1.1 describes the general process of development and connection between the two archives highlighting the flows of information. As can be seen, the Sapienza University archive (Uni_source) once connected with SISCO, has been reconstructed and transferred to Sapienza University in the form of an unidentified file containing data of the alumni (Uni_out) and the corresponding Compulsory Communications data (Activations, Extensions, Transformations and Terminations) for a chosen time interval. The data obtained by this process allows us to reconstruct for all alumni the inflows and outflows from the subordinate and para-subordinate labour market. This process has generated the UNI.CO archive made up of two different matrices

Figure 1.1: Flows of information between the Sapienza alumni archive (Infostud) and the System of Compulsory Communications archive (SISCO)



according to the experimental phase established:

- the matrix (or archive) of contracts that shows all the information related to labour signed by the alumni for a specific time interval;
- the matrix (or archive) of alumni that traces, from a longitudinal point of view, all the activations, all the extensions, all the transformations and all the terminations of employment over time.

The activation file contains all the activations recorded in SISCO for each tax code in the file of the alumni provided by the University. There is no checking of the activations before or after the graduation date. In this file, the ID of the worker is repeated as many times as the activations recorded by SISCO from March 2008 to the time of the file extraction. The data file Uni_out is therefore the result of the merging between the Uni_source archive and the SISCO archive, where the tax codes have been replaced by unidentified identifiers and the dates of birth with the years of birth. For every item in the file UNI_OUT described above we have:

- no activation if the graduate has never had a communication recruitment from March 2008 to the extraction date.
- one or more activations when the graduate has had at least one communication recruitment from March 2008 to the date of extraction and therefore one or more lines of information as a function of the number of recruitments for the period from March 2008 to the date of extraction.

The activations file provided by SISCO is made up of 3 sections (see Tab. 1.1):
The first section (worker information), made up of six areas, contains information about the worker: a single identifier, gender, country of birth, citizenship code, code of residence permit, age at recruitment time. The second section (enterprise information), made up of three areas, consists of: information about the unidentified enterprise identification, work location and the economic sector code (ATECO classification). The third and final section provides information on the employment relationships, e. g. the starting date, the expected ending date and the legal duration at the time of recruitment (according to Tab. 1.1). The terminations file contains all the employment relationships included in the SISCO system records of job termination dates for each tax code contained in the Sapienza alumni file.

Table 1.1: Description of the Activations file (activated labour relationships)

Section	Description
Worker Information	Single identifier Genger Country of birth Citizenship number Code of residence permit Age (at recruitment)
Enterprise Information	Single identifier Work location Code of the economic sector (ATECO)
Information on the employment relationship	Starting date Expected ending date (at recruitment) Legal duration (at recruitment) Professional qualification (ISTAT) Contract type code Code working hours Internal code of application form type Internal code of action type Benefits code (at recruitment) Code of collective contract The social security code

In fact, the termination date of a contract may be updated by communicating an extension of the job (contained in the extension file). For example, an extension moves the initially planned ending date forward, as well as a transformation of a temporary contract to a permanent contract, which makes the termination date undetermined (contained in the transformations file), and finally a job termination (such as dismissal, resignation, retirement) sets a definite ending date (contained in the terminations file). The file thus contains the actual ending date of employment at the time of data extraction. The ending date of an employment relationship is a matter of great importance for the analysis of job transitions in order to calculate the job duration, which can be quantified as the number of working days in a worker's employment history. It is important to take into account the basic distinction between expected and actual duration of an employment relationship. The expected duration is the duration declared at the beginning of the signing of the relationship. Whereas, the actual duration takes into account the actual ending date of the employment relationship, which is the result of the following algorithm:

- the date of job termination (if this information is contained in the file of terminations);

- Alternatively, no date is given if the last available job change is a transformation to a permanent contract;

- Alternatively, the maximum ending date of the last extension (if at least one extension is recorded);

- Alternatively, the expected ending date if the job employment is temporary and there are no further job changes.

The difference in the information contained in the data set extracted from SISCO represent job relationships in their temporal evolution allowing us to reconstruct the true stories of the employments of the alumni. Moreover, this dataset contains the actual duration of work experience gained over time.

1.1.3 The UNI.CO archive

The new UNI.CO archive, born from the joint intersection of the information system of Sapienza University of Rome and the Italian Ministry of Labour archive, is adopted from now on, in order to analyse the demand for the skills and knowledge of Sapienza alumni.
In order to determine the temporal period and populations of interest, we have defined the design of the analysis in the following way:

- The "field of observation" is the temporal interval adopted to study the professional experiences of Sapienza alumni. In this study we have considered the three years following the date of university graduation, so as to consider alumni at the same time interval ("ceteris paribus").

- The "alumni set", is the set composed of all the Sapienza alumni for whom we have complete information in terms of contracts and job relationships in the field of observation. In this thesis we consider all the alumni who got their university degree in the classes between 01/03/2008 and 28/02/2009, a total, of 21,782 alumni.

- The "contracts set" is the set of working relationships (paid contracts) and professional experience (unpaid contracts) activated during the field of observation (44,804 contracts) or previously activated, but existing at the time of graduation (1,695 contracts). We have a total of 46,499 contracts.

The UNI.CO archive is made up of two different matrices:

1. The matrix of contracts: 44,499 statistical units. This matrix consists of all the Sapienza alumni' contracts which represent the population of interest for the analysis. In this matrix, for each contract, all the information in the CO archive has been associated with all the features of the Sapienza alumni' archive. This matrix allows us to study the complete set of features regarding the Sapienza alumni' contracts.

2. The matrix of alumni: 12,351 statistical units. Among the 21,782 alumni of the population potentially subject to analysis, we considered 12,351 alumni (56.7%

of the total) who had at least one occurrence recorded in the CO in the field of observation (we called them "matching alumni"). An occurrence is defined either by a contract which started in the three years following the graduation date, or as a contract which already existed at the time of graduation. The 9,431 alumni who do not match the SISCO archive are all alumni who have not had a job experience in the form of subordinate or para-subordinate work (e. g. alumni who went abroad, self-employed alumni with a VAT number, unemployed alumni). This matrix allows us to perform longitudinal statistical studies on the job relationships of all Sapienza alumni after graduation. The 9,431 alumni who do not match the SISCO archive are all alumni who have not had a job experience in the form of subordinate or para-subordinate work (e. g. alumni who went abroad, self-employed alumni with a VAT number, unemployed alumni.). This matrix allows us to perform statistical studies of longitudinal types about the work relationships of all Sapienza alumni after graduation.

Because of the fragmentation of the subordinate and para-subordinate work compared to all other possible forms of employment, it is not possible to interpret the matching in SISCO as the only positive result for job search during the three years after graduation. Rather, it is possible to consider this matching as an important indication of the demand characteristics for subordinate and para-subordinate work available for Sapienza alumni in Italy. In Tab. 1.2 we show the percentage of matching alumni considered in the matrix of alumni divided according to graduation level and gender. This percentage can be interpreted as the 'probability' of a Sapienza graduate signing up for at least one contract of subordinate and para-subordinate work in the observation field . This table allows us to make some preliminary observations on the different job positions of Sapienza alumni. The number of female alumni (13,591) is higher than the number of male alumni (8,191). In relative terms, among the three year degree alumni, there is a greater number of matching alumni in the female component (58%) than in the male component (52.3%). This is also observed for the master degree alumni. This observation agree with national data on the labour market according to which the female employment rate slightly exceeds that of male employment probably due to a progressive outsourcing of the produc-

Table 1.2: Number of Sapienza alumni and percentage of matching alumni according to the graduation level and gender

		Gender		
		Female	**Male**	**Total**
All graduates	Num. of Sapienza alumni	13591	8191	21782
	% matching alumni	58.7	53.4	56.7
Degree	Numb. of Sapienza alumni	7170	4361	11531
	% matching alumni	58	52.3	55.9
Master's Degree	Numb. of Sapienza alumni	6421	3830	10251
	% matching alumni	59.5	54.6	57.7

tion system of goods and services. In Table 3 we show the results of the Sapienza matching alumni for disciplinary sectors. The disciplinary sectors with higher proportions of matching alumni are: education (81%), chemistry and pharmacy (76%). The disciplinary sectors with a percentage higher than the overall average (57%) are: political science and social science (67%), economics and statistics (65%), languages (60%). The disciplinary sectors with a lower percentage than the overall average are: law (31%), architecture (44%) and the sciences (50%). All the other sectors have values around the overall average with a smaller positive and negative variation.

Table 1.3: Results of the Sapienza matching alumni for disciplinary sectors

Disciplinary Sector	Matching		Non-Matching		Total	
	Num. alumni	%	Num. alumni	%	Num. alumni	%
Architecture	889	44	1141	56	2030	100
Chemical & Pharmaceutical	439	76	142	24	581	100
Economics & Statistics	1332	65	726	35	2058	100
Geography & Biology	483	52	448	48	931	100
Law	301	31	658	69	959	100
Engineering	1215	59	829	41	2044	100
Education	142	81	33	19	175	100
Literature	1390	53	1236	47	2626	100
Linguistical	748	60	507	40	1255	100
Medical	1519	53	1323	47	2842	100
Political & Social	2248	67	1119	33	3367	100
Psychological	1355	59	961	41	2316	100
Scientific	290	50	292	50	582	100
ND	-	-	16	100	16	100

1.2 Methodology

The employment problem for alumni lies not only in the difficulty of getting a job, as shown by a reduced number of days per year of the contract, but also in the quality of demand expressed by the system of production of goods and services. An additional extended problem in this period is the nature of short-term job contracts, with no professional qualifications and which are not coherent with the educational curriculum. Taking into account data concerning the history of the integration of Sapienza alumni into the Italian subordinate and para-subordinate market, a very complex situation emerges which can only be analysed and correctly interpreted through comparing the different forms of contract and professional qualifications specified in the contracts. With these aims in mind, we suggest, new indicators that can be of help in studying this complex situation. Job quality has been the focus of both conceptual and political theories. In the neo-classical model of a perfectly competitive labour market, wages fully capture the job quality aspects. However, the economic literature suggests that, in practice, wage differentials do not fully

compensate for all job differences, mainly due to a number of market failures, such as incomplete information matching costs, monopsony power, and human capital. Hence, wage alone does not capture all aspects of the quality of work. In addition, other characteristics of the job including human capital, working conditions, health risks and job satisfaction, need also to be considered to create an overall picture, (European Commission [EES] (2008) [24]). Quality in work is a multidimensional phenomenon, which in principle may touch a broad set of individual jobs and worker characteristics ranging from wages, work organization, reconciliation between work and family life. As the socio-economic literature does not appear to have reached a clear consensus on a definition of variables which attempt to analyse and monitor job quality, this factor needs careful consideration . Additionally, some of the relevant aspects are more of a qualitative and subjective nature, thus raising obvious measurement problems. The transitional labour markets school (known as TLM theory, Schmid and Gazier (2002) [60]), highlights the alleged 'erosion of standard employment', stressing the importance of studying labour market transitions, and the distinction between good and bad ones. Reforms of employment protection legislation that have loosened regulations in fixed-term and other non-standard contracts, while maintaining existing legislation on permanent contracts, may be counterproductive. In particular, they may help create segmented labour markets, where workers under non-standard contracts face poorer working conditions and less favorable career prospects. All these new theoretical policy considerations provide an opportunity for revisiting the EU concept of job quality and reignite a discussion on the main empirical determinants of job quality indicators. According to the economics of happiness literature, Frey and Stutzer (2002) [30], the absolute level of wages is weakly correlated with subjective levels of job satisfaction. Ranking and habit formation effects seem to dominate when compared with wage-level effects; furthermore workers are unhappy if they are paid less than their colleagues, while a wage rise tends to have only a transitory effect. A 'good' job quality concept should be multidimensional, including both objective and subjective indicators. Green (2006) [32] adopts a broad definition focusing exclusively on job characteristics, and job quality is evaluated looking at the range of capabilities and rewards granted to workers to achieve their own well-being and fulfill the company's goals,

including wages, skills used in the job and the intensity of work. In the context of Green's framework, TLM theory, Schmid and Gazier (2002) [60], is highly relevant for giving a dynamic or lifecycle perspective to employment quality issues, and for highlighting the interactions between employment and other life spheres. This school stresses the change in paradigm from standard employment to differentiated employment careers, with a variety of working-time and contractual arrangements, and more frequent changes of statuses between employment, unemployment, inactivity, education, family care and non-paid activities.

1.2.1 The chance of getting a "good job"

In this section we focus on the assessment of the significance that contracts with optimal characteristics take on in relation to the number of Sapienza alumni who are able to obtain contracts in this field of observation. We study individual characteristics (gender and age), the choice of the disciplinary sectors, the performance during the student's university studies, the student's diploma, the number of contracts or employment relationships after graduation. All these variables lead to the achievement of "optimal positions". In this way, the main aim of this study is to verify the achievement of permanent contracts with coherent qualifications regarding the university career path. According to the "Eurostat model", alumni are considered properly placed in the labour market when they occupy professional qualifications identified with ISCO1 (managers) and ISCO2 (intellectual and scientific professions) according to the ISCO classification. We have defined the concept of "optimal contract" as a contract that presents the following characteristics simultaneously:

- permanent position;
- highly qualified position, identified with the ISCO 1 and ISCO 2 classification;
- actual duration more than or equal to 8 months[3].

[3] The choice of this threshold (at least 8 months) comes from D.lgs.181/2000, which considers those who work for less than eight months as having unemployment status. This legislation was recently amended by Law 28, June 2012, n. 92 "Provisions for the reform of the labor market with a view to growth."

Furthermore we have defined the concept of "quasi-optimal contract" as a contract that presents the following characteristics simultaneously:

- highly qualified position, identified with the ISCO 1 and ISCO 2 classification
- actual duration more than or equal to 8 months.

Recently, many papers have also considered the ISCO 3 classification in the context of highly qualified positions according to ISFOL. In this study we decided to consider only highly qualified contracts as defined by ISCO 1 and ISCO 2. Tab. 1.4 shows the percentage of Sapienza alumni who have signed at least one subordinate or para-subordinate contract with optimal characteristics and the average number of working days of such contracts in the field of observation. As we can see:

- at least one permanent contract has been signed by 22.9% of Sapienza students and by 24.1% master degree alumni. The number of working days with a permanent position is very low (116.9 for the 3-year degree alumni and 132.3 for master degree alumni) with respect to the total number of working days in the field of observation. This number may depend either on the fact that many contracts are signed at the end of this period, or the contracts are short-term due to unforeseen termination.
- the percentage of alumni with at least one highly qualified professional contract is 27.8% for the 3-year degree alumni, and 48% for master degree alumni in the field of observation. The number of work days with a high professional qualification contract is low (66.4 for the 3-year degree alumni and 186.2 for the master degree alumni). Contracts of more than or equal to 8 months duration, have been signed by 60.5% (with 331.8 days worked) of the 3-year degree Sapienza alumni, and by 69.9% (with 418.1 work days) of those with a master degree.
- an optimal contract has been signed only by 2.2% of the 3-year degree alumni and 6.2% of the master degree alumni. In the field of observation, there is only an average of 11.4 and 35.3 observed days worked for the 3-year degree and master degree alumni respectively.

Table 1.4: Percentage of alumni who signed at least one contract with optimal characteristics and the average number of worked days of these contracts during the field of observation

Optimal Contract features	**% of Sapienza Alumni**		
	Degree	**Master Degree**	**Total**
Permanent Position	22.9	24.4	23.6
Highly qualified position	27.8	48	37.5
Actual duration $\geq$ 8 months	60.3	69.9	64.9
Optimal contract	2.2	6.2	4.1
Quasi-Optimal contract	10.7	25.9	18
	Numbers of worked days		
Permanent Position	116.9	132.3	124.3
Highly qualified position	66.4	186.2	123.7
Actual duration $\geq$ 8 months	331.8	418.1	373.1
Optimal contract	11.4	35.3	22.8
Quasi-Optimal contract	44.6	140.1	90.3

- an optimal contract has been signed only by 2.2% of 3-year degree alumni and 6.2% of the master degree alumni. In the field of observation, there is only an average of 11.4 and 35.3 observed work days for the 3-year degree and master degree alumni respectively.

- 10.7% of the 3-year degree Sapienza alumni and 25.9% of the master degree alumni have a quasi-optimal contract. There is a marked difference between the percentages: even though it is reassuring that at least one quarter of the master degree alumni obtained a quasi-optimal position. Note that in this case the work day average is not high: 44.6 for the 3 -year degree alumni and 140.1 for the master degree alumni.

We have considered different definitions of optimal and quasi-optimal contract in order to choose the definition to be adopted for this thesis work. We can see in Tab. 1.5:

- In "Case a" we have the co-presence of the three best characteristics: permanent contract, with high professional qualifications (ISCO1 and ISCO2) with an actual duration of at least 8 months. If we remove the condition of at least 8 months duration (case b), the percentage of alumni who signed at least one contract with these characteristics changes moderately from 4.1% to 4.4% (the difference is based only on 43 contracts).

- "Case c" is defined by substituting the actual duration with days worked in "Case a" condition. The result is that the number of alumni with at least one optimal contract is reduced from 4.1% to 3.3%, because of signed contracts for at least 8 months, which are only partly carried out during the field of observation.

- For the quasi-optimal contracts, in "Case d", we have the co-presence of high professional qualifications with an actual duration of at least 8 months.

- "Case f" is obtained by substituting the request of actual duration with the one of days worked. The result is that the number of alumni with at least one quasi-optimal contract is reduced from 18.1% to 14.6%.

- "Case e" is obtained by removing, in the adopted quasi-optimal contract definition, the request of counting the unpaid contracts. The number of alumni with at least one quasi-optimal contract is reduced from 18.1% to 16.9%..

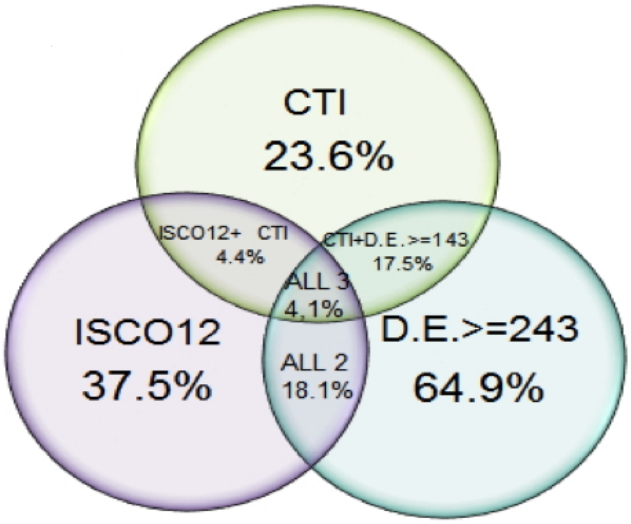
- "Case g" is the same as "Case e" with actual duration substituted by the days worked. The percentage of alumni is reduced from 14.6% to 13.7%.

Fig. 1.2 shows a graphic representation of the distribution sets of Sapienza alumni according to the different definitions of optimal and quasi-optimal contracts. We conducted a preliminary statistical study on the relationship between the integration of the subordinate and para-subordinate market in coherence with their educational curriculum and a few variables representing a university career, job path and the personal data. In particular, we have built a predictive model of the probability of getting an "optimal contract" and a "quasi-optimal contract" in the field of observation through a series of logistic regression models.

Table 1.5: Alternative definitions in optimal and quasi-optimal contracts. Number and percentage of alumni who signed at least one contract with optimality characteristics during the field of observation

Type contracts	**Case**	**Optimal features**	**Num. of alumni with (A) or without (B) optimal features**		**Sapienza alumni**	%
			A	**B**	**(TOT)**	$(\frac{\mathbf{A}}{\mathbf{TOT}})$
Optimal	a	CTI+ISCO12+D.E≥ 243	506	11845	12351	4,1
	b	CTI+ISCO12	549	11802	12351	4,4
	c	CTI+ISCO12+W.D≥ 243	407	11944	12351	3,3
Quasi Optimal	d	CTI+ISCO12+D.E≥ 243	2231	10120	12351	18,1
	e	CTI+ISCO12	2092	10259	12351	16,9
	f	CTI+ISCO12+W.D≥ 243	1808	10543	12351	14,6
	g	CTI+ISCO12+W.D≥ 243 without unpaid contract	1692	10659	12351	13,7

Figure 1.2: Graphic representation of the definitions adopted.



CTI: alumni with a permanent contract; ISCO12: alumni with a high professional qualification contract . (made up of manager, intellectual and scientific professions); D.E. ≥243: alumni with a contract with actual duration more than or equal to 8 months, in days.

1.2.2 Indicators for a longitudinal analysis: worsening, improvement and stability in the field of observation

To study the evolution of the type of contract, the actual duration, and the professional qualification in the three years after graduation, we considered:

- the first and the last contract in the field of observation (indicated by I and Z)
- the two most important contracts in the field of observation in terms of maximum actual duration (indicated, in chronological order, by K1 and K2).

The comparison between the features of these pairs of contracts allows us to monitor the evolution of the type of contract, the actual duration, and professional qualification, in order to classify Sapienza alumni in terms of improvement, worsening or stability of their job position. Note that the double comparison was designed precisely because of the fact that the contracts have very fragmented durations and therefore to consider only the first and the last contract signed in three years may not be sufficient for a dynamic assessment. The comparison of contracts can be performed only for 7,910 alumni. For the remaining 4,441 alumni the comparison cannot be made, mainly because they did not sign more than one contract during the three years after graduation. It is worth noting that for alumni who sign only two contracts in three years, the two comparisons coincide, and the first and last contract (I and Z) correspond to K1 and K2. We constructed and analysed a series of indicators associated with the type of work, the type of contract, the dynamics of the duration and the professional qualification of the contracts. These indicators have been designed with the aim of identifying improvements, a worsening or stability in the job experiences of alumni following the three years of graduation, and these represent one of the major elements made available for reading the results of this experimentation. The indicators designed for the comparison between the contracts and the evaluation of their evolution during the field of observation, can be divided into three different sessions:

1. Indicators of the type of contracts.
2. Indicators of the actual duration of contracts.

3. Indicators of the professional qualification of contracts.

The first indicator allows us to study the evolution of the type of contract. We defined two new variables, one for the comparison between K1 and K2 and one for the comparison between I and Z, with the following four modalities:

- Improvement: if the graduate changes from an internship to a job contract;
- High stability: if the evolution is from a job contract to another;
- Low stability: if the evolution is from an internship to another;
- Worsening: if the graduate changes from a job contract to an internship contract.

A first assessment of labour demand for Sapienza alumni was conducted with the identification of prominent features of the contracts (for example type, duration, title), based on the information and classifications available at different levels of details. The variable type of work, which is relevant for the interpretation of the "quality" of the signed contracts by alumni in the three years after graduation, was defined by classifying the contracts in the following five modalities:

a1) Standard (full time permanent work);

a2) Partially standard (part-time permanent work);

b1) Atypical (e.g. fixed-term employment, employment with project contract);

b2) Mixed cause contracts (e.g. contracts that provide a training as a component of Apprenticeship, insert, CFL in public administration);

c) Professional experience (not properly work relationships, such as an Internship).

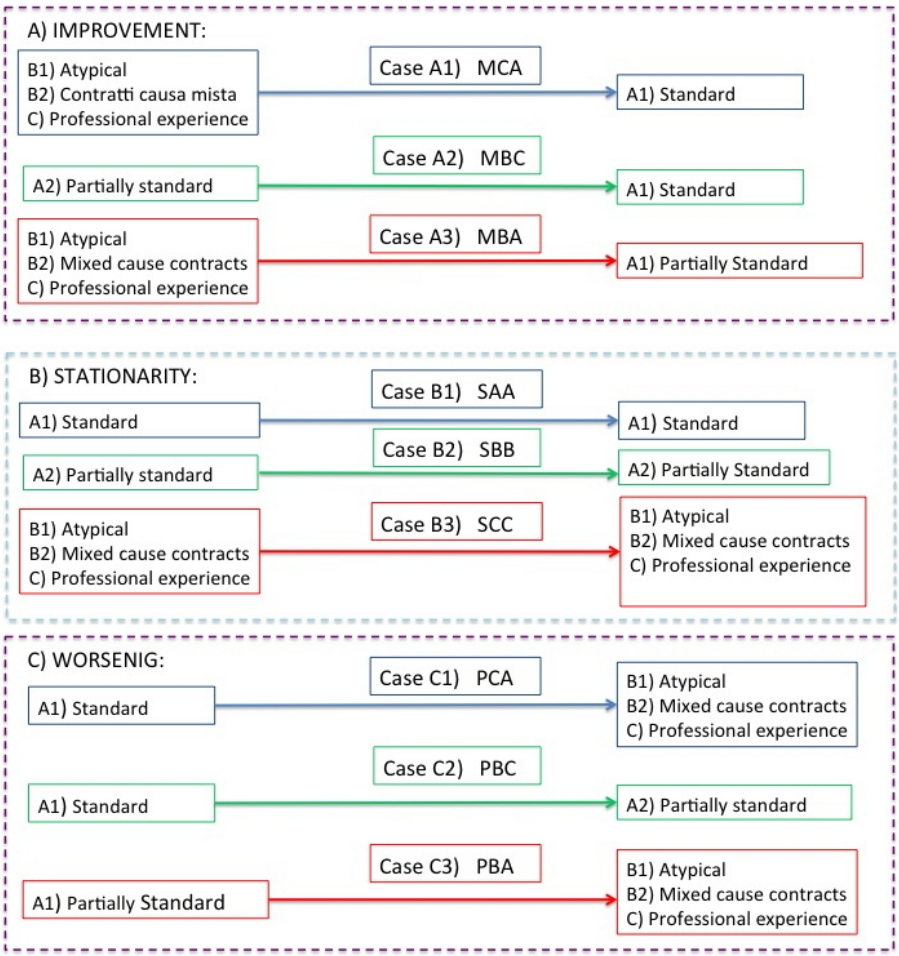
In order to study the evolution of the type of work, starting from the variables type of work, we defined two new variables, one for the comparison between K1 and K2, and one for the comparison between I and Z, with the following modalities. See Fig. 1.3:

- MCA: improvement of type 1, the graduate changes from a non standard to a standard contract, i. e. from an atypical or a mixed cause or internship job contract to a standard job contract (case A1);
- MBC: improvement of type 2, the graduate changes from a partially standard to a standard contract (case A2);
- MBA: improvement of type 3 the graduate changes from a non standard to a partially standard contract, i. e. from an atypical or a mixed cause or internship contract to a partially standard job contract (case A3);
- SAA: stability of type 1, the graduate changes from a standard contract to a new standard contract (case B1);
- SBB: stability of type 2, the graduate changes from a partially standard contract to a new partially standard contract (case B2);
- SCC: stability of type 3, the graduate changes from a non standard contract to a new non standard contract (case B3);
- PAC: worsening of type 1, the graduate changes from a standard to a non standard contract, i. e. from a standard job contract to an atypical or a mixed cause or internship job contract (case C1);
- PBC: worsening of type 2, the graduate changes from a partially standard to a non standard contract, i. e. from a partially standard job contract to an atypical or a mixed cause or internship job contract (case C2);
- PAB: worsening of type 3, the graduate changes from a standard to a partially standard contract (case C3).

In order to study the evolution of the actual duration of contracts we have introduced two indicators: one for the comparison between K1 and K2 and one for the comparison between I and Z, with the following four modalities. See Fig. 1.4.

- M8: improvement, the graduate changes from a job contract with an actual duration of less than 8 months to a job contract with a duration of at least 8 months (case a);

Figure 1.3: Graphic representation of the type of work evolution



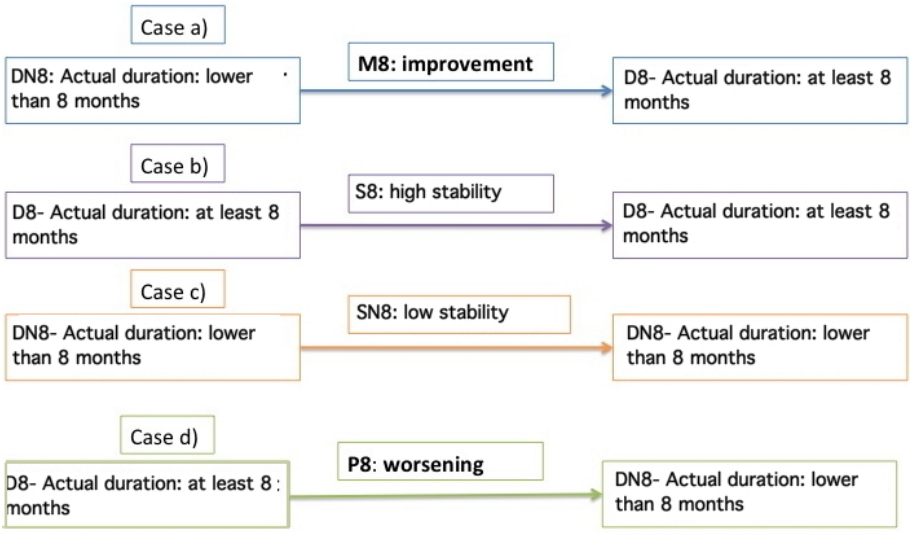
- S8: high stability, the graduate changes from a job contract with actual duration of at least 8 months to a new job contract with the same duration (case b);

- SN8: low stability, the graduate changes from a job contract with actual duration of less than 8 months to a new job contract with the same duration (case c);

- P8: worsening, the graduate changes from a job contract with actual duration of at least 8 months to a job contract with a duration of lower than 8 months

(case d).

Figure 1.4: Graphic representation of the evolution of the actual duration



As far as the evolution of the actual duration, two variables have been defined, each of them based on the ratio between the lengths of the two contracts which are compared: the duration of the second contract is set in the numerator in chronological order; whereas the determinator consists of the duration of the first contract. The first variable compares K1 to K2, the second one compares I to Z. We have established the following three modalities:

- actual duration increased by at least 50%, i. e. the ratio is equal or greater than 1.5;
- actual duration varied less than 50%, i. e. the ratio is between 0.5 and 1.5;
- actual duration decreased by at least 50%, i.e. the ratio is equal or less than 0.5.

In order to study the evolution of the professional qualification at different levels of disaggregation, we decided to adopt the ISCO classification that is used at European levels. Regarding the professional qualifications we aggregated the ISCO classifications in three hierarchical levels according to the ISTAT standard:

1. high skilled professions (ISCO 1 and ISCO 2) made up of managers and intellectual and scientific professions;

2. medium skilled professions (from ISCO 3 to ISCO 6) made up of technical professions, skilled employees and professionals in commercial activities and services, specialized personnel assigned to agriculture, forestry and fisheries;

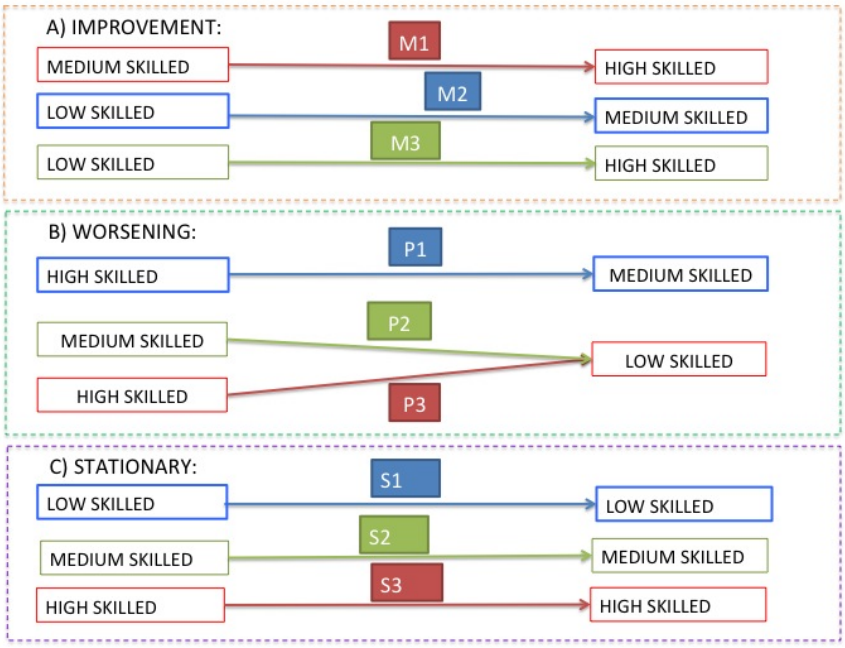
3. low skilled professions (from ISCO 7 to ISCO 9) made up of craftsmen and skilled workers, workers of industrial machinery and assembly lines, and unqualified workers.

We defined two new variables, one for the comparison between K1 and K2, and one for the comparison between I and Z, with the following modalities:

- M1: improvement, the graduate's profession changes from a medium skilled to a highly qualified profession;
- M2: improvement, the graduate's profession changes from a low skilled to a highly qualified profession;
- M3: improvement, the graduate's profession changes from a low skilled to a medium skilled profession;
- P1: worsening, the graduate's profession changes from a highly qualified to a medium skilled profession;
- P2: worsening, the graduate's profession changes from a medium skilled to a low skilled profession;
- P3: worsening, the graduate's profession changes from a highly qualified to a low skilled profession;
- S1: stability, the graduate's profession changes from a highly qualified profession to a new one with the same qualification;
- S2: stability, the graduate's profession changes from a medium skilled profession to a new one with the same qualification;

- S3: stability, the graduate's profession changes from a low skilled profession to a new one with the same qualification.

Figure 1.5: Graphic representation of the job professional qualification



1.3 Discussion

A detailed analysis of the based on the UNI.CO archive concerning the type of work evolution in the field of observation reveals a positive panorama. In Tab. 1.6 shows that 80.7% of the alumni have paid contracts as a first and last job experience in an employment relationship in the field of observation. A percentage of 14.7% have an improvement in the type of work; while for 2.9% of the alumni a worsening is recorded (this percentage is 3.7% for the 3-year degree alumni and 2% for the master degree alumni). Finally, only 1.7%. of the alumni have low stability contracts. The evolution of the type of work in the course of three years can be described in more detail taking into consideration the type of contract.

Table 1.6: Evolution of the type of work during the field of observation

Evolution of the work type	**Comparison K1-K2**			**Comparison I-Z**		
	Degree	**Master Degree**	**Total**	**Degree**	**Master Degree**	**Total**
IMPROVEMENT internship→ job contract	10.7	17.7	14.3	10.2	19.2	14.7
WORSENING job contract→internship	3.7	2.6	3.2	3.7	2	2.9
HIGH STABILITY job contract	83.3	77.4	80.3	84.2	77.2	80.7
LOW STABILITY internship	2.2	2.2	2.2	1.9	1.5	1.7
Total	**100**	**100**	**100**	**100**	**100**	**100**

In the Tab. 1.7, the study the evolution of the contract type is reported. We observe that:

- the vast majority of alumni do not substantially change their own job position: this stability concerns 80.8% of alumni when we consider the comparison between K1 and K2, and 79.4% in the comparison between I and Z. The difference between the 3-year degree alumni and the master degree alumni is very low.

- the main component of the stability type of contract is represented by the set of alumni maintaining a non-standard contract (equivalent to more than 3/4 of the total number of alumni of both degree levels); the fraction of alumni maintaining a standard contract in the two compared situations is only slightly more than 2%. This percentage rises to 3% when partially standard contracts are also considered;

- analysing the comparison between K1 and K2, 13.6% of the alumni obtained an improvement, this percentage changes to 14.8% in the case of the comparison between I and Z. As we can see, there is no significant difference between the 3-year degree and master degree alumni. The main component of the improvement of the type of contract is represented by the set of alumni changing

from a non-standard to a standard contract (9.4% and 10.4% of the total, for the two comparisons respectively);

- alumni who reported a worsening in the type of contract are fewer than those who reported an improvement: this set is composed of 5.6% of the alumni in the comparison between K1 and K2, the percentage rises to 5.9% when the comparison between the first and last contract is considered. The number of alumni who worsen their condition is higher for students with the 3-year degree (6.6% versus 4.6% in the comparison between K1 and K2, and 6.4% against 5.3% for I and Z). It is important to observe finally that the most relevant worsening recorded is from partially standard contracts to non-standard contracts (2.9% and 3.2% of the total number of alumni in the two comparisons), and the second most important one from the standard contract to non-standard ones (2.3% in both comparisons). The worsening in relation to the passage from standard to partially standard contracts is less relevant (only 0.4% and 0.3% of the total number of alumni in both comparisons).

Concerning the study of the actual duration of the contracts, first of all we have compared contracts with an actual duration of greater or less than 8 months. This threshold is chosen to divide the work experience as sufficiently important or of little importance from a duration point of view.
In Tab. 1.8, the evolution of the actual duration of the contract is described:

- a significant percentage of alumni improve their job position, from a contract of less than 8 months to one of at least 8 months. In the case of the comparison between K1 and K2 the percentage is 31.6% of the alumni, whereas it changes to 38.3% when considering the comparison between the I and Z. In both the comparisons the improvement is slightly higher for the master degree in terms of type of degree.

- In the two comparisons the majority of alumni is in a situation of stability with more or less the same actual job duration (59% and 54% respectively). A negative element is the fact that the main component of the stability comes from the alumni with contracts with an actual duration of less than 8 months in both situations (31.9% and 34.2% respectively).

- The alumni who hold contracts of more than 8 months (27.1% and 19.8% in the two comparisons) together with the alumni that improve their position by changing from shorter-term contracts to contracts of at least 8 months, represent more than half of the total number of alumni (58.7% in the comparison between K1 and K2, and 58.1% in the comparison between I and Z).

Another assessment of the dynamics of the actual duration of the contracts in the field of observation can be carried out by classifying alumni depending on whether the two contracts have registered or not changes in actual duration greater than 50%. In Tab. 1.9, the variation of the contract actual duration data is reported; we observe that:

- more than half of the alumni increase the duration of the contract by more than 50% (53% of the alumni in the comparison between K1 and K2, and 62.4% of alumni in the comparison between I and Z). In both the comparisons this effect is slightly higher for master degree as regards graduation degree.

- only 12.3% and 14.3% of the alumni, respectively in the two comparisons, have decreased by 50% the actual duration of the contract.

- the actual duration of the contract does not change more than 50% in 34% and in 23.3% of the cases in the two comparisons.

Finally, the evolution in the field of observation of professional qualifications that emerges from the job contracts of the alumni is of particular interest. With reference to the classification described above (high, medium and low levels) the picture that arises from Tab. 1.10 is:

- about 3/4 of the alumni do not substantially change their level of professional qualification;

- in particular, 63.9% of 3-year degree alumni and 47.6% of master degree alumni appear to maintain a medium professional qualification in the comparison between K1 and K2; these percentages are 62.6% and 46.2% when comparing the first and last contract. The second important element with regard to stability is represented by the number of alumni who maintain high level qualifications:

this component represents 12.9% of 3-year degree alumni and 24.4% of master degree alumni in the comparison between K1 and K2, and 12.9% and 22.9% in the other case. The stability of the low professional qualifications is marginal, 1.5% and 1.4% of the total number of alumni in both comparisons;

- 13.3% and 14.7% of alumni (respectively in the two comparisons) improve their level of professional qualifications; the improvement is more frequent for the master degree alumni 15.1% (17%) with respect to 11.5% (12.3%) of the 3-year degree alumni (in the comparison between K1 and K2 (I and Z)). The most frequent improvement is represented by the transition from a medium qualification to a high qualification (10.9% and 11.7% of the total number of alumni in the two comparisons);

- a worsening of the professional qualification status is less common and represents 10.8% and 11.6% of alumni in the two comparisons. The worsening is slightly higher for the master degree alumni 11.7% (12.9%) regarding 9.8% (10.3%) of 3-year degree (in the comparison between K1 and K2 (I and Z)). Finally, a worsening is the most common situation for the transition from a high to a medium professional qualification.

Table 1.7: Evolution of the type of contract in the field of observation (percentage values)

Evolution of the contract type	**Comparison K1-K2**			**Comparison I-Z**		
	Degree	**Master Degree**	**Total**	**Degree**	**Master Degree**	**Total**
IMPROVEMENT	**13.9**	**13.2**	**13.6**	**14.5**	**15.0**	**14.8**
Partially standard ⇒ Standard (MBA)	0.8	0.7	0.7	0.8	0.7	0.7
Non standard ⇒ Standard (MCA)	9.4	9.4	9.4	9.8	11.0	10.4
Non standard ⇒ Partially Standard (MCB)	3.8	3.2	3.5	3.9	3.3	3.6
STATIONARY	**79.4**	**82.2**	**80.8**	**79.1**	**79.6**	**79.4**
Standard contract (SAA)	2.2	2.2	2.2	2.1	1.9	2.0
Partially standard contract (SBB)	1.1	1.3	1.2	1.0	1.3	1.2
Non standard contract (SCC)	76.2	78.7	77.4	76.0	76.5	76.2
WORSENIG	**6.6**	**4.6**	**5.6**	**6.4**	**5.3**	**5.9**
Standard ⇒ Partially standard (PAB)	0.4	0.5	0.4	0.3	0.4	0.3
Standard ⇒ Non standard (PAC)	2.6	2.0	2.3	2.3	2.4	2.3
Partially standard ⇒ Non standard (PBC)	3.6	2.1	2.9	3.8	2.6	3.2
TOTAL	**100**	**100**	**100**	**100**	**100**	**100**

Table 1.8: Evolution of the contract duration in the field of observation (percentage values)

Evolution of the contract actual duration	**Comparison K1-K2**			**Comparison I-Z**		
	Degree	**Master Degree**	**Total**	**Degree**	**Master Degree**	**Total**
Less of 8 month ⇒ at least 8 month (M8)	30.65	32.48	31.58	36.80	39.77	38.29
At least 8 month ⇒ less of 8 month (P8)	9.28	9.46	9.37	7.85	7.55	7.70
Stationary at least 8 month (S8)	24.27	29.96	27.13	18.02	21.58	19.81
Stationary less of 8 month (SN8)	35.81	28.10	31.93	37.33	31.09	34.20
TOTAL	**100**	**100**	**100**	**100**	**100**	**100**

Table 1.9: Variation of the contract duration in the field of observation (percentage values)

Variation	**Comparison K1-K2**			**Comparison I-Z**		
	Degree	**Master Degree**	**Total**	**Degree**	**Master Degree**	**Total**
Increased by at least 50%	52.50	53.96	53.24	60.81	64.05	62.44
Decreased by al least 50%	12.99	11.70	12.34	14.76	13.79	14.27
Unchanged (less than 50%)	34.51	34.34	34.42	24.42	22.16	23.29
TOTAL	**100**	**100**	**100**	**100**	**100**	**100**

Table 1.10: Evolution of professional qualification in the field of observation (percentage values)

Evolution of professional qualification	**Comparison K1-K2**			**Comparison I-Z**		
	Degree	**Master Degree**	**Total**	**Degree**	**Master Degree**	**Total**
IMPROVEMENT	**11.54**	**15.09**	**13.35**	**12.34**	**16.97**	**14.70**
Medium⇒ high (M1)	8.38	13.34	10.91	8.76	14.59	11.73
Low ⇒ medium (M2)	2.63	1.20	1.90	2.97	1.65	2.30
Low ⇒ High (M3)	0.53	0.56	0.54	0.61	0.73	0.67
STATIONARY	**78.70**	**73.17**	**75.88**	**77.36**	**70.15**	**73.69**
On high (S1)	2.2	2.2	2.2	2.1	1.9	2.0
On medium (S2)	63.90	47.56	55.58	62.64	46.23	54.28
On low (S3)	1.85	1.20	1.52	1.86	0.98	1.41
WORSENIG	**9.76**	**11.74**	**10.77**	**10.30**	**12.88**	**11.61**
High ⇒ medium (P1)	7.88	10.75	0.35	8.38	11.56	10.00
Medium ⇒ low (P2)	1.52	0.72	1.11	1.54	1.01	1.27
High ⇒ low (P3)	0.36	0.27	0.31	0.38	0.31	0.34
TOTAL	**100**	**100**	**100**	**100**	**100**	**100**

Chapter 2

Latent Variable Path Models

In the 1920s the geneticist Sewall Wright developed the path analysis to study phylogenetic models. Wright was interested in understanding the causal relationships underlying social systems by solving a system of equation based on the correlations among the variables that are supposed to influence the outcome. Several decades later, path analysis was introduced into the social scientific research thanks to the work of Blalock, Ducan and others. During the 1970s, path analysis was introduced also in sociology, psychology, economics, political science ecology and other fields and became more popular.

The path analysis method can be considered an extension of multiple regression analysis because it estimates many regression equations at the same time according to the theoretical relationships among the variables. In the 1980s, the path analysis evolved into a variety of causal or Structural Equation Models (SEM) that can be viewed as a set of methods for analysing multiple relationships among blocks of variables. The very important feature of these methods is the possibility to represent quantitatively concepts that cannot be measured directly and that are called **latent variables (LVs)** or constructs, factors etc.. These concepts encompass and summarize a set of information that in some way reflect the meaning of the concept. The latent variables are indirectly measured by means of variables that are called **manifest variables (MVs)**. The feature of a manifest variable is that it contains information that reflects (partially) some aspects of the construct, in such a way that the information contained in the indicators are elaborated to obtain a repre-

sentation as faithful as possible of the latent variable. These methods can also be used for exploration, visualization, explanation, prediction, classification and study of the structural systems.

2.1 Structural Equation Models: the bases

In this Section we will introduce the notations, the common drawing conventions and the specification of the model. The SEM, Bollen (1989) [7]; Kaplan (2000) [41], include a number of statistical methodologies with the aim to estimate a network of causal relationships, based on a theoretical model, linking two or more latent concepts, each measured through a number of observable indicators. The study of the complexity inside a system is performed through a causality network among the latent concepts. The standard graphical representation of these is depicted in Fig. 2.1. The theoretical causal relationships in the model are represented through a *path diagram* where the latent variables are represented by the ellipses or circles, the manifest variables by the rectangles or squares and the dependence relationships among the variables by the oriented arrows (either latent or manifest). A variable is called *exogenous* if it helps to explain other variables while a variable is called *endogenous* if it is caused by one or more variables within the model, therefore endogenous variables have the graphical representation of incoming arrows.

The SEM model consists of two sub-models: the structural (or inner) model and the measurement (or outer) model. The measurement model describes the relations among the LV and its MVs, see Fig. 2.2. The common term to indicate this part of the graph is *block*. There are three measurement types: the *reflective scheme or Mode A*, the *formative scheme or Mode B* and the *MIMIC mode*. A block is conceptually defined as *reflective* (case (a) in Fig. 2.2) or, more generally, as outwards directed, if the LV is assumed to be a common factor that describes its own MVs. In this case, MVs should be highly correlated, as they are caused by the same common factor. In other words, the block is expected to be unidimensional and internally consistent. Hence, the set of manifest variables are assumed to measure the same singol underlying concept. [1]

[1] In this case, the LV is considered the cause of the MVs.

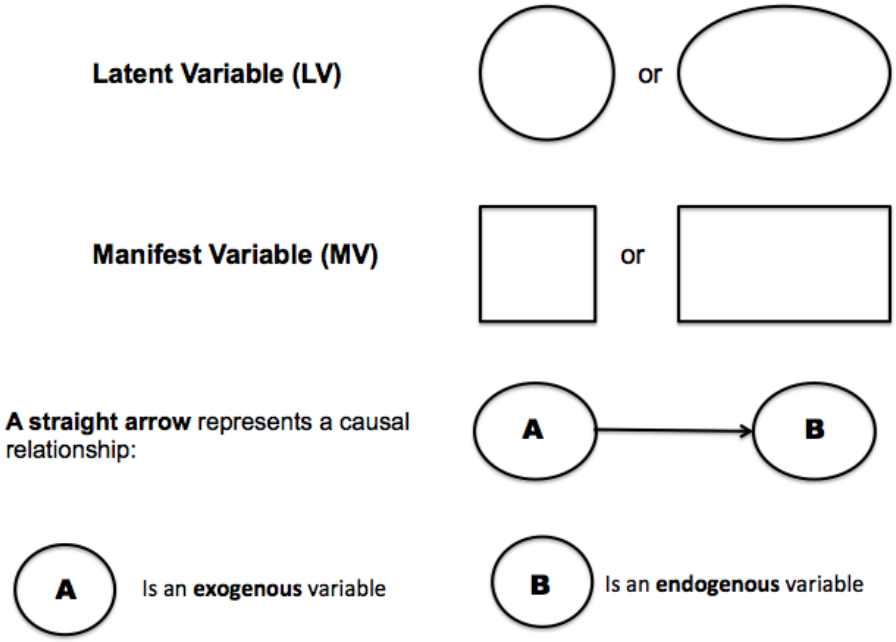


Figure 2.1: The most common symbols in the Structural Equation Models

Differently, when each MV (or sub-blocks of MV) represent different ingredients of the underlying concept, the block is defined as *formative* (cases (b) and (c) in Fig. 2.2) or inwards directed.
In this case the manifest variables within a block are weakly correlated among them. Finally, a block can be composed of both reflective and formative MVs: this is the *MIMIC* (multiple-indicator multiple indicator causes) case. Regardless of the measurement model used, the parameters to be estimated are the so-called outer weights ω_{pq} and loadings λ_{pq}.
The structural (or inner) model specifies the relationships among the LVs, see Fig. 2.3. Using this model, it is possible to estimate the regression coefficients connecting the latent variables among them. The regression coefficients are the path coefficients (indicated by β_{mj}).
There exist two different approaches to Structural Equation Models estimation:

Covariance-based : the goal is to reproduce the sample covariance matrix of the manifest variables by means of the model parameters. It is a confirmatory

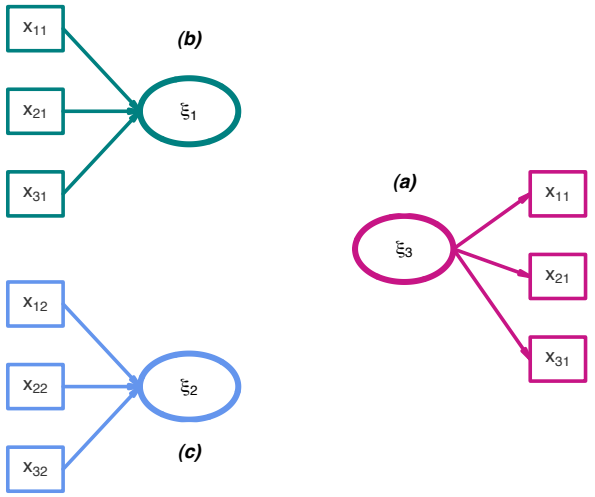


Figure 2.2: Path Diagram depicting outer model. The two measurement options are showed: the reflective way case (a) and the formative way case (b) and (c).

approach aimed at validating a model. This techniques can be considered as a generalization of the Confirmatory Factor Analysis to the case of the multi-tables data linked to one another. Several estimation techniques exist. Jöreskog (1970) [40] proposed one of the first method which is based on estimating the Structural Equation Models using the maximum likelihood (ML) method (SEM-ML). In literature, all these techniques are called *LInear Structural RElations LISREL*-type techniques.

Component-based : the goal is to provide an estimation of the latent variables in such a way they are the most correlated with one another (according to the path diagram structure) and the most representative of each corresponding block of manifest variables (as happens in the principal component analysis). It is to be considered more an exploratory approach than a confirmatory one.

In the following explanation we will focus our discussion on the Partial Least Squares Path Modeling (PLS-PM), Tenenhaus (2008) [62], that is the principal estimation technique among the component-based methods.
It is an iterative algorithm that separately estimates the several blocks of the mea-

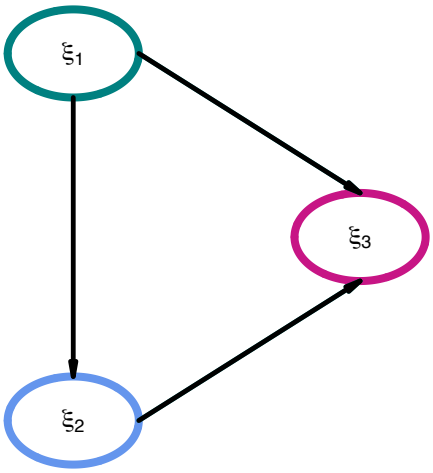


Figure 2.3: Path Diagram depicting inner model

surement model and then, in a second step, estimates the structural model coefficients. The feature of the PLS-PM is to explain at best the residual variance of the latent variables and also of the manifest variables in any regression of the model, Fornell and Bookstein (1982) [29].

Since the PLS-PM does not aim at reproducing the sample covariance matrix, it is considered more an explorative approach than a conservative one. The PLS-PM does not require strong assumption with respect to the distributions, the sample size and the measurement scale and therefore it is to be considered a completely free method and therefore can be contextualized as *soft-modeling* approach. For this reason the classical parametric inferential framework cannot be used and it is replaced by empirical confidence intervals and hypothesis testing procedures based on resampling methods, Chin (1998) [15]; Tenenhaus et al. (2005) [66], such as jacknife and bootstrap. Finally, PLS-PM is more oriented to optimizing predictions (explained variances) than statistical accuracy of the estimates.

2.2 The PLS Path Modeling

The family of PLS methods have been initiated by the work of Swedish statistician Herman Wold and his research group in the period from 1960s to early 1980s. They developed a set of approaches based on iterative algorithms applied on Ordinary Least Squares regressions. The set of PLS methods was developed to tackle several problems of data analysis by applying a number of adapted least squares procedures. Different names for these methods were proposed over the years such as *Nonlinear Iterative Least Squares* (NILES), *Nonlinear Iterative Partial Squares* (NIPALS), *Partial Least Squares basic design*, and PLS soft modeling, among others.
In the first period of evolution, several versions of least-squares-based iterative algorithms were applied to a variety of analysis and modeling problems related to data coming from social sciences (e.g. economics, psychology, sociology). In the early 1980s, further developments of the PLS principles and the applications based on regression problems to chemistry and the food industry have contributed to the success of the PLS techniques.
PLS methods are mainly composed of two schools: the PLS regression models (PLS-R) and the PLS path modeling (PLS-PM). In the first period only the PLS-R technique had a great success. In the late 1990s, thanks to the work of Michel Tenenhaus and his group, the general framework of the PLS methods has been repositioned on the context of data analysis. In the last decade (2000-2010), the PLS methods were subjected to a great effort of international diffusion which continues today.
PLS-PM estimates the network of linear relations among the MVs and their own LVs, and among the LVs inside the model, through a system of inter-dependent equations based on simple and multiple regressions.
Let us call $\boldsymbol{X}$ the dataset containing N standardized units observed on P variables, $\boldsymbol{X}$ can be seen as a NxP matrix. The variables are divided into Q mutually and exclusive blocks $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q$ where the generic q-th block $\boldsymbol{X}_q$ has P_q variables with $\sum_q^Q P_q = P$. The PLS-PM studies the relationships among Q blocks $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q$ of manifest variables MVs, which are the expressions of Q latent variables $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_Q$.
This partition of data reflects the PLS-PM objective which is that of charactering the unobserved latent variables that are supposed to represent the underline struc-

ture of the data which we are interested in highlighting. Therefore each group of variables influences a unique LV exclusively.
As said earlier, the corresponding conceptual model can be represented by path diagrams and, also, the PLS path model consists of two sub-models: the structural (or inner) model and the measurement (or outer) model.

2.2.1 The Structural Model

The structural model specifies the relationships among the LVs, see Fig. 2.3. In general it is possible to classify the LVs in two types according to the path diagram. The LV is endogenous if it is supposed to depend on other LVs and it is exogenous otherwise. Structural relationships can be taken in account by means of a lower triangular matrix $\mathbf{L}$ of order Q according to the mutual relationships among the LVs. The generic element $l_{qq'}$ is equal to 1 if $\boldsymbol{\xi}_q$ depends on $\boldsymbol{\xi}_{q'}$ and 0 otherwise ($l_{qq} = 0$). In the structural model each endogenous LV ξ_q, $(q = 1, \ldots, Q)$ is linked to other LVs by the following multiple regression model:

$$\boldsymbol{\xi}_q = \beta_{q0} + \sum_{q'=1}^{Q} l_{qq'}\beta_{qq'}\boldsymbol{\xi}_{q'} + \boldsymbol{\nu}_q \tag{2.1}$$

where $\boldsymbol{\xi}_q$ $(q = 1, \ldots, Q)$ is the generic endogenous LV, β_{q0} is the intercept term, $\beta_{qq'}$ is the path coefficient interrelating the q-th exogenous latent variable to the q'-th endogenous one, and the $\boldsymbol{\nu}_q$ is the error of the inner relation. The subscript q' of $\boldsymbol{\xi}_{q'}$ refers to all the latent variables that are supposed to predict $\boldsymbol{\xi}_q$. The path coefficients represent the "strength and direction" of the relations between the response $\boldsymbol{\xi}_q$ and the predictors $\boldsymbol{\xi}_{q'}$. This model is recursive, therefore the paths formed by the arrows of the inner model cannot form a loop.
The only hypothesis of this model is what Wold named *predictor specification hypothesis*, Wold (1982) [70]:

$$E(\boldsymbol{\xi}_q|\boldsymbol{\xi}_{q'}) = \beta_{q0} + \sum_{q'=1}^{Q} l_{qq'}\beta_{qq'}\boldsymbol{\xi}_{q'} \tag{2.2}$$

which implies that the latent variables $\boldsymbol{\xi}_q$ is uncorrelated with the residual $\boldsymbol{\nu}_q$ and the residual $\boldsymbol{\nu}_q$ has zero mean. The idea behind this specification is that the linear relationships are conceived from a standard regression perspective: we want to understand as far as possible the conditional expected value of the response ξ_j determined by its predictors ξ_q. It is important to note that the only request is the existence of the first and second order moments in the variables.

2.2.2 The Measurement Model

The measurement model describes the relations among the LV and its MVs, see Fig. 2.2. When a block is defined as *reflective* (case (a) in Fig. 2.2) the MVs should be highly correlated, as they are caused by the same common factor. In other words, the block is expected to be unidimensional and internally consistent. Hence, the set of manifest variables are assumed to measure the same unique underlying concept. In this case, the relation between each MV $\boldsymbol{x}_{pq}$, $(p = 1, \ldots, P_q)$ and the corresponding LV is considered to be linear and it is generally modeled as

$$\boldsymbol{x}_{pq} = \lambda_{pq0} + \boldsymbol{\lambda}_{pq}\boldsymbol{\xi}_q + \boldsymbol{\epsilon}_{pq} \tag{2.3}$$

where λ_{pq0} is a location parameter, $\boldsymbol{\lambda}_{pq}$ is a loading term stemming from simple regression model and the imprecision in the measurement process is represented by the error term $\boldsymbol{\epsilon}_{pq}$. Also in this model the predictor specification hypothesis $E(x_{pq}|\xi_q) = \lambda_{pq0} + \lambda_{pq}\xi_q$ is required.
To check the unidimensionality of a block several tools exist.

Cronbach's alpha represents a measure of internal consistency and it is defined as:

$$\alpha = \frac{\sum_{p \neq p'} cor(\boldsymbol{x}_{pq}, \boldsymbol{x}_{p'q})}{P_q + \sum_{p \neq p'} cor(\boldsymbol{x}_{pq}, \boldsymbol{x}_{p'q})} \text{ x } \frac{P_q}{P_q - 1} \tag{2.4}$$

where P_q is the number of manifest variables in the q-th bolck and $\boldsymbol{x}_{pq}$ and $\boldsymbol{x}_{p'q}$ are two MVs of the q-th block. A block is considered unidimensional if this index is larger than 0.7 for confirmatory studies.

Dillon-Goldstein's rho (or Jöreskog's) measures the composite reliability of the

block and for this reason it is called also composite reliability index. It is defined as:

$$\rho = \frac{\left(\sum_{p=1}^{P_q} \lambda_{pq}\right)^2}{\left(\sum_{p=1}^{P_q} \lambda_{pq}\right)^2 + \left(\sum_{p=1}^{P_q} 1 - \lambda_{pq}\right)^2} \tag{2.5}$$

A block is considered unidimensional if this index is larger than 0.7.

Principal component analysis of a block a block may be considered unidimensional if the first eigenvalue of its correlation matrix is higher than 1, while the others are smaller.

Confirmatory TETRAD Analysis [Bollen & Ting (1993) [8], Gudergan, Ringle, Wende & Will (2008) [33]] A TETRAD is defined as the difference of the products of two pairs of covariances between MVs of the same block. Using a bootstrap-based test, all non redundant TETRADs are tested to be different from zero. A block is assumed to be reflective if all null hypotheses are accepted, in others cases, a block is considered formative.

According to Chin (1998) [15] the Dillon-Goldstein's rho is considered to be a better indicator of the unidimensionality of a block than the Cronbach's alpha. Cronbach's alpha actually provides a lower bound estimate of reliability.
When a block is defined as *formative* (cases (b) and (c) in Fig. 2.2), the measurement model can be expressed as:

$$\boldsymbol{\xi}_q = \sum_{p=1}^{P_q} \mathrm{w}_{pq} \boldsymbol{x}_{pq} + \boldsymbol{\delta}_q \tag{2.6}$$

where w_{pq} is the coefficient linking each manifest variable to the corresponding latent variable and the fraction of the corresponding latent variable not accounted for by block of manifest variables is represented by the error term $\boldsymbol{\delta}_q$. The assumption behind this model is the *predictor specification*:

$$E(\boldsymbol{\xi}_q | \boldsymbol{x}_{pq}) = \sum_{p=1}^{P_q} \mathrm{w}_{pq} \boldsymbol{x}_{pq} \tag{2.7}$$

In such case the LV is better understood as an emergent construct that summarizes its own MVs. In other words the MVs are considered the cause of the LV.
Whatever scheme is used to built the measurement model, the parameters to be estimated are the so-called outer weights ω_{pq} and loadings λ_{pq}.

2.2.3 The PLS-PM Algorithm

The PLS-PM consists of an iterative algorithm composed by two steps: in the first step it separately solves out the blocks of the measurement model and then, in the second step, estimates the path coefficients in the structural model, see Tab. 2.1. In the PLS-PM the outer weights ω_{pq}, linking each MV to corresponding LV, are estimated by an iterative algorithm in which the latent variable scores are obtained through the alternation of the outer ($\boldsymbol{\nu}_q$) and inner estimations ($\boldsymbol{z}_q$) of the LVs. No formal proof of convergence of the general algorithm has been provided until now, but in some cases the PLS-PM loop has been proved to monotonically convergence versus a criterion. However, convergence is always assured in practice.

Step 1, see Fig. 2.4

The procedure starts by choosing arbitrary outer weights vectors (generally $\boldsymbol{w}_q = \mathbf{1}$). These weights are then standardized in order to obtain LVs with unitary variance. Once the outer weights have been initialized, the first stage of the iterative procedure consist in obtaining the outer estimation in which each LV ($\boldsymbol{\nu}_q$) is calculated as a linear combination of its own centered MVs:

$$\boldsymbol{\nu}_q \propto \sum_{p=1}^{P_q} w_{pq}\boldsymbol{x}_{pq} = \mathbf{X}_q\boldsymbol{\omega}_q \tag{2.8}$$

where $\boldsymbol{\nu}_q$ is the standardized (zero mean and unitary standard deviation) outer estimate of the $q-$th latent variable $\boldsymbol{\xi}_q$. The symbol $\propto$ means that the left side of the equation corresponds to the normalized right side ($\boldsymbol{\nu'}_q\boldsymbol{\nu}_q = N$).
Then, the inner estimation is calculated in which each LV (z_q) is obtained as a

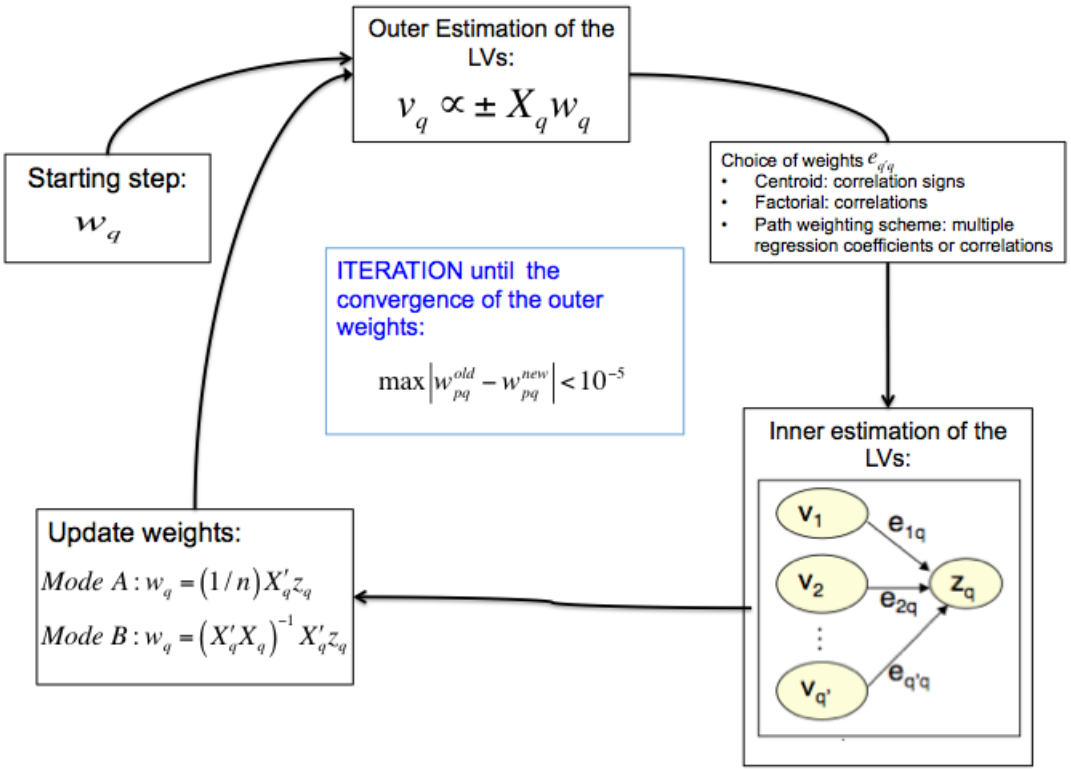


Figure 2.4: Scheme of the PLS-PM iterative procedure

normalized linear combination of the outer estimates of the connected LVs:

$$\boldsymbol{z}_q \propto \sum_{q'=1}^{Q} c_{qq'} e_{qq'} \boldsymbol{\nu}_q \tag{2.9}$$

where $\boldsymbol{\nu}_q$ is the standardized inner estimate of the q-th latent variable $\boldsymbol{\xi}_q$, each inner weight $(e_{qq'})$ is the weight of this linear combination and the $c_{qq'}$ is a generic element of the square matrix $\boldsymbol{C}$ of order Q, where $c_{qq'} = 1$ if $\boldsymbol{\xi}_q$ is connected to $\boldsymbol{\xi}_{q'}$ in the path diagram and $c_{qq'} = 0$ otherwise.
There are three different methods for evaluating the inner weights:

1. the Centroid scheme (Wold's original scheme) in which each inner weight $e_{qq'}$ is equal to the sign of the correlation between the outer estimate $\boldsymbol{\nu}_q$ of the q-th latent variable and the outer estimate $\boldsymbol{\nu}_{q'}$ connected with $\boldsymbol{\nu}_q$.

2. the Factorial scheme, proposed by Lohmöller (1989), [48], in which each inner weight $e_{qq'}$ is equal to the correlation between the outer estimate $\boldsymbol{\nu}_q$ of the q-th latent variable and the outer estimate $\boldsymbol{\nu}_{q'}$ connected with $\boldsymbol{\nu}_q$.

3. the Structural or path weighting scheme in which each inner weight $e_{qq'}$ is equal to

   - $\boldsymbol{\nu}_{q'}$ coefficient in the multiple regression of $\boldsymbol{\nu}_q$ if $\boldsymbol{\xi}_{q'}$ is a latent predictor of $\boldsymbol{\xi}_q$;
   - correlation between $\boldsymbol{\nu}_q$ and $\boldsymbol{\nu}_q$ if $\boldsymbol{\xi}_{q'}$ is a latent response of $\boldsymbol{\xi}_q$.

Generally the most common method is the centroid one because it is well-adapted to the situation in which the manifest variables in a block are strongly correlated to each other. In situations in which the correlation among the manifest variables inside a block is weaker, it is more appropriate to use the factorial scheme. The procedure goes on by updating the outer weights ω_{pq} using the first inner estimation ($\boldsymbol{z}_q$) of the latent variables obtained. The *Mode A* or the *Mode B* mode is used to update the outer weights.

Mode A (Reflective): each outer weight ω_{pq} is update as regression coefficient in the simple regression of the p-th manifest variable of the q-th block ($\boldsymbol{x}_{pq}$) on the inner estimation $\boldsymbol{z}_q$ of the q-th latent variable. So, since $\boldsymbol{z}_q$ is standardized, the generic outer weight ω_{pq} is obtained by:

$$\omega_{pq} = \text{cov}^2(\boldsymbol{x}_{pq}, \boldsymbol{z}_q) \tag{2.10}$$

Mode B (Formative): the vector $\boldsymbol{\omega}_q$ of the weights ω_{pq} associated to the manifest variables of the q-th block is updated as vector of the regression coefficients in the multiple regression of the inner estimate of the q-th latent variable $\boldsymbol{z}_q$ on its manifest variables $\boldsymbol{X}_q$:

$$\boldsymbol{\omega}_q = (\boldsymbol{X'}_q \boldsymbol{X}_q)\, \boldsymbol{X'}_q \boldsymbol{z}_q \tag{2.11}$$

[2] One characteristic of the regression coefficient is to reduce the variance between each manifest variable and the corresponding inner estimate of the latent variable, if the manifest variables have been also standardized, the covariance becomes a correlation.

where $\boldsymbol{X}_q$ is built by the P_q manifest variables $\boldsymbol{x}_{pq}$ previously centered and scaled by $\sqrt{1/N}$.

The choice of the method for updating the outer weight is related to the nature of the measurement model. Generally, *Mode A* is suggested for endogenous latent variables while *Mode B* for the exogenous ones. Inner and outer estimation stages are alternated until the convergence on the outer weights reached.

Step 2

Once the convergence of the outer weights ω_{pq} is obtained, the procedure continues computing the LVs score, the path coefficients and the loadings.

The LVs scores are evaluated as linear combination of its own block of manifest variables (called *weight relation*):

$$\hat{\boldsymbol{\xi}}_q = \sum_{p=1}^{P_q} \omega_{pq} \boldsymbol{x}_{pq} \propto \boldsymbol{X}_q \boldsymbol{\omega}_q \tag{2.12}$$

where $\hat{\boldsymbol{\xi}}_q$ is the standardize latent variable scores associated to the q-th latent variable $\boldsymbol{\xi}_q$, the variables $\boldsymbol{x}_{pq}$ are centered and ω_{pq} are the outer weights obtained once the algorithm has reached convergence.
It is important not to avoid the mistake of exchange the *weight relation* defined in Eq. (2.12) with the expression of the measurement model in the case of formative way defined in Eq. (2.7). The weight relation only establishes that any LV is defined as a weighted sum of its own MVs. It does not affect the direction of the relationship between the LV and its MVs in the outer model. Such direction (inwards or outwards) determines how the weights in Eq. (2.7) are estimated.

The path coefficients are obtained through an OLS multiple regression among the estimated latent variable scores, according to path diagram structure. Let us assume that $\boldsymbol{\xi}_j \, (j = 1, \ldots, J)$ is the generic endogenous latent variable and $\hat{\boldsymbol{\Xi}}_{\to j}$ the matrix of the corresponding latent predictors, so the path coefficient

vector for each $\boldsymbol{\xi}_j$ is:

$$\hat{\boldsymbol{\beta}}_j = \left(\hat{\boldsymbol{\Xi}}'_{\to j}\,\hat{\boldsymbol{\Xi}}_{\to j}\right)^{-1}\hat{\boldsymbol{\Xi}}'_{\to j}\,\hat{\boldsymbol{\xi}}_j \tag{2.13}$$

where $\hat{\boldsymbol{\Xi}}$ includes the scores of the latent variables that explain the j-th endogenous latent variable $\boldsymbol{\xi}_j$, and $\hat{\boldsymbol{\xi}}_j$ is the latent variable score of the j-th endogenous latent variable.

The loadings are evaluated as correlations between a latent variable scores and its manifest variables:

$$\hat{\boldsymbol{\lambda}}_{pq} = \text{cor}\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q\right). \tag{2.14}$$

for each block q with $P_q > 1$.

2.2.4 Model validation

The PLS-PM is a *soft modeling* approach, for this reason the validation of the model regards only the way in which relationships are modeled, in the two models that make up this method: the structural and measurement model. The following null hypotheses should be rejected:

- $\lambda_{pq} = 0$ because each MV should be correlated to its corresponding LV;
- $\omega_{pq} = 0$ because each LV should be affected by all MVs of its block;
- $\beta_{qq'} = 0$ because each latent predictor should be causative with respect to its latent response;
- $R^2_j = 0$ because each endogenous $\boldsymbol{\xi}_j$ should be explained by its latent predictors;
- $\text{cor}(\boldsymbol{\xi}_q, \boldsymbol{\xi}_{q'}) = 0$ because the LVs are assumed to be connected by a statistically significant correlation. Rejecting this hypothesis means assessing the *Nomological Validity* of the PLS Path Models;
- $\text{cor}(\boldsymbol{\xi}_q, \boldsymbol{\xi}_{q'}) = 1$ because the LVs are assumed to measure concepts that are different among them. Rejecting this hypothesis means to assess the *Discriminant Validity* of the PLS Path Models;

Table 2.1: PLS Path Modeling algorithm

Input: $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q]$, i. e. Q blocks of centred manifest variables
Initialize: the outer weights, i. e. $\boldsymbol{\omega}_q = \mathbf{1}$ and $s = 0$

Step 1: Iterative algorithm

1: s=s+1
2: **for all** $q = 1, \ldots, Q$ **do**
3: Calculate the outer estimation $\boldsymbol{\nu}_q$ of the LVs: $\boldsymbol{\nu}_q^{(s)} \propto \sum_{p=1}^{P_q} \omega_{pq}^{(s)} \boldsymbol{x}_{pq}$
4: Updating of the inner weights $e_{q'q}^{(s)}$ choosing among these schemes:

Centroid: $e_{qq'}^{(s)} = sign[cor(\boldsymbol{\nu}_q^{(s)}, \boldsymbol{\nu}_{q'}^{(s)})]$
Factorial: $e_{qq'}^{(s)} = cor(\boldsymbol{\nu}_q^{(s)}, \boldsymbol{\nu}_{q'}^{(s)})$
Path weighting: multiple regression coefficients or correlations.

5: Calculate the inner estimation $\boldsymbol{z}_q^{(s)}$ of the LVs: $\boldsymbol{z}_q \propto \sum_{q'=1}^{Q'} c_{qq'} e_{qq'}^{(s)} \boldsymbol{\nu}_{q'}^{(s)}$
6: Updating of the outer weights $\boldsymbol{\omega}_q$ choosing between:

Mode A: $\boldsymbol{\omega}_q^{(s+1)} = (1/N) \boldsymbol{X'}_q \boldsymbol{z}_q^{(s)}$
Mode B: $\boldsymbol{\omega}_q^{(s+1)} = (\boldsymbol{X'}_q \boldsymbol{X}_q)^{-1} \boldsymbol{X'}_q \boldsymbol{z}_q^{(s)}$

7: **end for**
8: **Stages 1-7 are repeated until convergence** of the outer weights is achieved

i. e. until: $\max |\omega_{pq}^{(s+1)} - \omega_{pq}^{(s)}| < 10^{-5}$

Output Step 1: the outer weights $\boldsymbol{\omega}_q$

Step 2: Computation

9: Computation of the LV scores: $\hat{\boldsymbol{\xi}}_q = \sum_{p=1}^{P_q} \omega_{pq} \boldsymbol{x}_{pq} \propto \boldsymbol{X}_q \boldsymbol{\omega}_q$

10: Computation of the Path Coefficients: $\boldsymbol{\beta}_j = \left(\hat{\boldsymbol{\Xi}}'_{\rightarrow j} \hat{\boldsymbol{\Xi}}_{\rightarrow j} \right)^{-1} \hat{\boldsymbol{\Xi}}'_{\rightarrow j} \hat{\boldsymbol{\xi}}_j$

11: Computation of the loadings: $\hat{\boldsymbol{\lambda}}_{pq} = \text{cor}\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q \right)$.

Output Step 2: the LV scores $\hat{\boldsymbol{\xi}}_q$, the path coefficients $\boldsymbol{\beta}_j$ and the loadings $\hat{\boldsymbol{\lambda}}_{pq}$.

- AVE_q and $AVE_{q'}$ smaller than $\text{cor}(\boldsymbol{\xi}_q, \boldsymbol{\xi}_{q'})$ because the LV should be related more strongly with its block of indicators than with another LV representing a different block of indicators.

If some of these hypotheses are not rejected, somethings was wrong in choosing variables or in model specification.

2.2.5 Model assessment

In the PLS-PM frame, due to the fact that the model does not require distributional assumptions, the estimates of the parameter variability are obtained empirically by means of a bootstrap procedure. The validation of the quality of the model can also be studied by the evaluation of a few indicators that we briefly discuss in the following.
Taking into account the PLS-PM structure (according to the path diagram), each part of the model needs to be validated: the measurement model, the structural model and the total model. For this reason there exist some different measures.
To verify the quality of the measurement model, there exist the following measures:

Communality index measures how much of the MV variability in the q-th block is explained by its own LV score. This quantity represents the average of the squared correlation between each MV of the q-th block and the corresponding latent variable score $\hat{\boldsymbol{\xi}}_q$. It is defined as

$$\text{Com}_q = \frac{1}{P_q}\sum_{p=1}^{P_q} \text{cor}^2(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q) = \frac{1}{P_q}\sum_{p=1}^{P_q} \hat{\boldsymbol{\lambda}}_{pq}^2, \qquad \forall q = 1, \ldots, Q. \tag{2.15}$$

So, a measurement model is good if each MV is well summarized by its own LV score.

Average Variance Extracted [Fornell & Larcker (1981) [28]] represents the part of variance of the block explained by the latent variable score $\hat{\boldsymbol{\xi}}_q$. It is defined as

$$\text{AVE}_q = \frac{\sum_{p=1}^{P_q} \hat{\lambda}_{pq}^2}{\sum_{p=1}^{P_q} \text{var}(\boldsymbol{x}_{pq})}, \qquad q = 1, \ldots, Q. \tag{2.16}$$

It is important to note that in the case where the manifest variables are standardized the Communality index coincide with the Average Variance Extracted for less than the constant $1/P_q$.

Average Communality index defined as weighted average of all the Q blocks specific Communality index, where the weights are equal to the number of MVs in each block. It is expressed as

$$\overline{\text{Com}} = \frac{\sum_{q:P_q>1} P_q \text{Com}_q}{\sum_{q:P_q>1} P_q} = \tag{2.17}$$

$$= \frac{\sum_{q:P_q>1} \sum_{p=1}^{P_q} \text{cor}^2(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q)}{\sum_{q:P_q>1} P_q} \tag{2.18}$$

The sum considers only the blocks with a number of manifest variables greater than 1. This index measures the goodness of the whole measurement model.

The quality of each structural equation can be measured by the evaluation of the R^2 of the fit but this is not sufficient to evaluate the whole structural model since R^2 values only take into account the fit of each regression in the structural model. New indexes are computed for each endogenous block in addition to the R^2 value in order to evaluate also the measurement model. So, to verify the quality of the structural model we have:

Redundancy index that measures, for each endogenous LV, the portion of variability of MVs related to an endogenous LV $\boldsymbol{\xi}_j$ explained by its latent predictors. It is expressed as

$$\text{Red}_j = \text{Com}_j \text{ x } R_j^2. \tag{2.19}$$

Average Redundancy represents a global quality measure of the structural model. It is defined as the average of the redundancies in the model:

$$\overline{\text{Red}} = \frac{1}{J} \sum_{j=1}^{J} \text{Red}_q \tag{2.20}$$

where J is the total number of the endogenous LVs in the model.

The GoF index is a commonly used because it takes into account the model performance in both the measurement and structural model and thus provide a single measurement for the overall prediction performance of the model.

Goodness of fit index: proposed by Tenenhaus et al.(2004)[65] represents a global criterion of goodness of fit. It is defined as the geometric mean of the average communality index and the average R^2 value. It is expressed as:

$$\text{GoF} = \sqrt{\overline{\text{Com}} \text{ x } \overline{R^2}} \tag{2.21}$$

where $\overline{R^2}$ is the R^2 average:

$$\overline{R^2} = \frac{1}{J} R^2(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\boldsymbol{\xi}_q \to \boldsymbol{\xi}_j}). \tag{2.22}$$

The use GoF index is conceptually more appropriate in the case of reflective measurement models. In the formative measurement models, the communality can be calculated and interpreted taking into account that we expect lower communalities but higher R^2 values as compared to reflective models. According to the Eq. (2.18) and Eq. (2.22) the GoF index con be re-written as:

$$GoF = \sqrt{\frac{\sum_{q:P_q>1} \sum_{p=1}^{P_q} \text{cor}^2(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q)}{\sum_{q:P_q>1} P_q} \text{ x } \frac{\sum_{j=1}^{J} R^2(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\boldsymbol{\xi}_q \to \boldsymbol{\xi}_j})}{J}}. \tag{2.23}$$

It is possible to define the relative GoF index (GoF_{rel}) proposed by Tenenhaus et al.(2004)[65] introducing the normalized version of the average communality index (T_1) and the normalized version of the average $\overline{R^2}$ index (T_2). T_1 is obtained as a sum of the communalities of each block divided by the first eigenvalue of the block. It is expressed as:

$$T_1 = \frac{1}{P} \sum_{q=1}^{Q} \frac{\sum_{p=1}^{P_q} \text{cor}^2(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q)}{\lambda_q^{(1)}} \tag{2.24}$$

because, when the data are mean centered and with unit variance, as in our

case, $\sum_{p=1}^{P_q} \text{cor}^2(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q) \leq \lambda_q^{(1)}$, where $\lambda_q^{(1)}$ is the first eigenvalue obtained by performing a Principal Component Analysis on the q-th block of MVs.
T_2 can be expressed as:

$$T_2 = \frac{1}{J} \sum_{j=1}^{J} \frac{R^2(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\boldsymbol{\xi}_q \to \boldsymbol{\xi}_j})}{\rho_j^2} \tag{2.25}$$

where ρ_j^2 is the first canonical correlation of the canonical analysis of the matrix $\boldsymbol{X}_j$ containing the MVs associated to the j-th endogenous LV, and $\boldsymbol{X}_q$ containing the MVs associated to the exogenous LVs explaining $\boldsymbol{\xi}_q$.

Relative Goodness of fit index: according to Eq.s (2.21), (2.24) and (2.25), the relative version of the GoF index can be expressed as:

$$\begin{aligned} \text{GoF}_{\text{rel}} &= \sqrt{T_1 \text{ x } T_2} = \\ &= \sqrt{\frac{1}{P} \sum_{q=1}^{Q} \frac{\sum_{p=1}^{P_q} \text{cor}^2(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q)}{\lambda_q^{(1)}} \text{ x } \frac{1}{J} \frac{\sum_{j=1}^{J} R^2(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\boldsymbol{\xi}_q \to \boldsymbol{\xi}_j})}{\rho_j^2}} \end{aligned} \tag{2.26}$$

This index is bounded between 0 and 1.

The relative GoF, like the GoF, is a descriptive index, therefore there is no inference-based threshold to judge the statistical significance of its values. Nevertheless a value of the relative GoF $\geq$0.90 is generally considered an excellent value for the model.

2.2.6 Optimizing criteria

PLS-PM is flexible technique that can be applied to many models with a number of LVs, different path linking them, and different ways assumed to compute both inner and outer weights. For these reasons, it is impossible to discuss the optimization of an overall single scalar function suitable for any model. Recently, a stationary equation for most of the models has been identified which shows that the approach PLS-PM can be rethought as a generalization of the multivariate analysis. In following we report a brief outline of the PLS-PM optimization criteria.

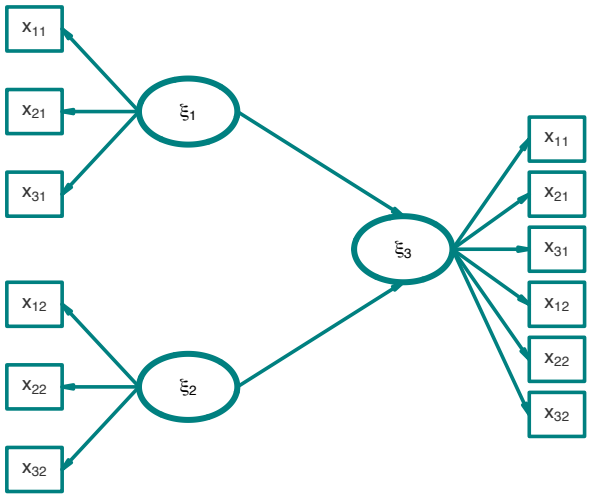


Figure 2.5: An example of hierarchical path model with three reflective blocks.

In the case of only two blocks $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the PLS-PM algorithm converges to three different stationary equations Lyttkens, Areskoug & Wold (1975) [49] depending on the way the outer weights are computed.

- Both the outer weights $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are calculated with Mode A, therefore the covariance between the LVs is maximized: PLS-PM converges to the first component of Inter-battery Analysis, Tucker (1958) [68].
- Both the outer weights $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are calculated with Mode B, therefore the correlation between the LVs is maximized: PLS-PM converges to the first component of Canonical Correlation Analysis (CCA), Hotelling (1936) [37].
- When $\boldsymbol{\omega}_1$ is estimated with Mode A and $\boldsymbol{\omega}_2$ is calculated with Mode B, the redundancy of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$ is maximized: PLS-PM converges to the first component of the corresponding Redundancy Analysis, Van de Wollemberg (1977) [69], D'Ambra & Lauro (1982) [19].

In the Multi-Block case, the following situations can be contemplated:

Hierarchical model : each block $\boldsymbol{X}_q$ is connected to a super-block $\boldsymbol{X}_{q+1}$ obtained by juxtaposing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ see Fig. 2.5 for three reflective blocks. When

Mode B is adopted for all the blocks, the PLS-PM algorithm, depending on the inner estimation scheme, converges to the stationary equations of Horst's or Carroll's generalized CCA, Horst (1961) [35], Carroll (1968) [13]. On the other hand, using Mode A and the path weighting scheme may lead to different techniques (e.g. the multiple factor analysis). For a complete review on the multi-block hierarchical case see Tenenhaus et al. (2005) [66].

Confirmatory model : each LV is related to a single block of MVs, and it is connected to all the LVs linked to the other blocks, see Fig. 2.6. This path model leads to the stationary equations of Kettenring's generalized CCA type, Kettenring (1971) [42], For a complete review on the confirmatory model refer to Tenenhaus & Hanafi (2009) [63].

Mode B model . When Mode B is used for all the blocks and the centroid scheme is used for the inner estimation of the LVs, the stationary equation of the PLS-PM algorithm is given by the Lagrange equation associated with the optimization criterion, following Glang (1988) [31] and Mathes (1993) [50],

$$\sum_{q \neq q'} c_{qq'} |\mathrm{cor}(\boldsymbol{X}_q \boldsymbol{\omega}_q, \boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'})| \tag{2.27}$$

with respect to $||\boldsymbol{\omega}_q|| = 1$.
Moreover, they showed also that, when Mode B is used for all the blocks and the factorial scheme is used for the inner estimation of the LVs, the stationary equation of the PLS-PM algorithm is given by the Lagrange equation associated with the optimization of the criterion

$$\sum_{q \neq q'} c_{qq'} \mathrm{cor}^2(\boldsymbol{X}_q \boldsymbol{\omega}_q, \boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'}) \tag{2.28}$$

with respect to $||\boldsymbol{\omega}_q|| = 1$. Hanafi, Hanafi (2007) [34], proved that Wold's iterative procedure is monotonically convergent to these criteria.

New Mode A model : all the outer weights are computed using the new Mode A estimation process that has been proposed by Tenenhaus and Tenenhaus (2011) [64] in order to overcame some difficulties connected with the Mode

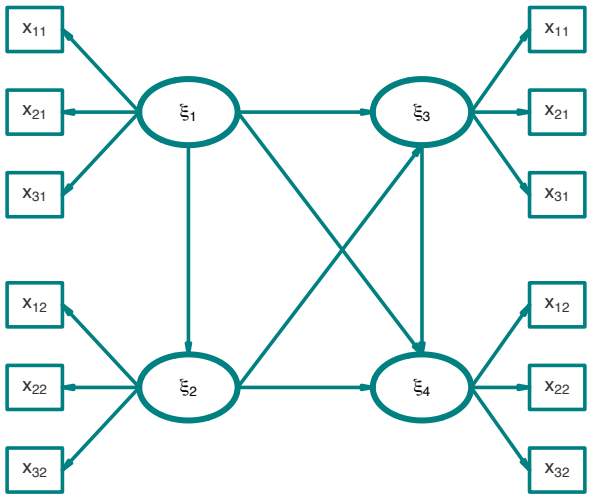


Figure 2.6: An example of confirmatory model with four reflective blocks.

A. As a matter of fact, when Mode A is applied to all the blocks, the PLS-PM algorithm does not seem to optimize any criterion as Krämer (2007) [43] showed that Wold's Mode A algorithm is not based on a stationary equation related to the optimization of a twice differentiable function. However, Tenenhaus and Tenenhaus (2011) [64], have recently extended the results of Hanafi, Hanafi (2007) [34], to a slightly adjusted Mode A in which a normalization constraint is put on outer weights rather than on LV scores. They showed that, when new Mode A is used in all the blocks and the centroid scheme for the inner estimation of the LVs, Wold's procedure monotonically converges to the criterion

$$\arg \max_{||\boldsymbol{\omega}_q||=1} \sum_{q \neq q'} c_{qq'} |\text{cov}(\boldsymbol{X}_q \boldsymbol{\omega}_q, \boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'})|. \tag{2.29}$$

Otherwise, when new Mode A is used for all the blocks and the factorial scheme

for the inner estimation of the LVs, it converges to the criterion

$$\arg\max_{||\boldsymbol{\omega}_q||=1} \sum_{q \neq q'} c_{qq'} \mathrm{cov}^2(\boldsymbol{X}_q \boldsymbol{\omega}_q, \boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'}). \tag{2.30}$$

General model : when both new Mode A and Mode B are used in the same model and the centroid scheme is chosen, Wold's procedure is shown to converge to the criterion

$$\arg\max_{||\boldsymbol{\omega}_q||=1} \sum_{q \neq q'} c_{qq'} |\mathrm{cor}(\boldsymbol{X}_q \boldsymbol{\omega}_q, \boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'}) \sqrt{\mathrm{var}(\boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'})^{\tau_{q'}}} \sqrt{\mathrm{var}(\boldsymbol{X}_q \boldsymbol{\omega}_q)^{\tau_q}}|. \tag{2.31}$$

When the factorial scheme is used, it converges to the criterion

$$\arg\max_{||\boldsymbol{\omega}_q||=1} \sum_{q \neq q'} c_{qq'} \mathrm{cor}^2(\boldsymbol{X}_q \boldsymbol{\omega}_q, \boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'}) \, \mathrm{var}(\boldsymbol{X}_{q'} \boldsymbol{\omega}_{q'})^{\tau_{q'}} \mathrm{var}(\boldsymbol{X}_q \boldsymbol{\omega}_q)^{\tau_q} \tag{2.32}$$

In Eq.s (2.31)-(2.32), $\tau_q = 1$ when the block q is estimated by new Mode A and $\tau_q = 0$ when the block q is estimated by Mode B.

Chapter 3

Non-Metric PLSPM

Partial Least Squares methods are able to handle concepts that cannot be directly measured. These methods, as we showed in Section 2.2.3, are component-based techniques where the components are calculated as linear combination of the corresponding block of the manifest variables. The two assumptions underlying PLS models are:

- Each variable is measured on an interval (or ratio) scale;
- Relationships between variables and latent constructs are linear and monotonic.

In many fields (for example marketing applications, social and health sciences, etc..) the use of categorical data such as person's gender (male/female), different means of transport (automobile, bicycle and bus), etc., is widespread.
These types of data could not be directly handled by PLSPM because they are not based on a metric. Sometimes, a simple way to overcome this problem, implemented by many software, is to replace each non-metric variable by the corresponding indicator matrix, i. e. using a complete disjunctive coding with the idea of considering the categories as if they were variables in themselves. This method is not a good solution because:

- in a PLS analysis the weights, quantities that measures the intensity of the relationship between the original variables and the latent variable, are calculated

for each variable. So in this case the weights are obtained for each dummy variable without having a global weight value for the entire variable. This means that the weight of each variable measure the impact of each individual category on the latent variable.

- the binary coding increases the size of the data matrix and this fact could generate sparse matrix.
- the weight of a dummy variable representing a category mainly associated with central values of the corresponding LC scores distribution is systematically underestimated.

In order to take into account these kind of data, often the categories of non-metric variables are arbitrarily quantified and then used as numerical indicators.
An interesting possibility to address this issue has been recently offered by a new procedure called Non-Metric Partial Least Squares Path Modeling (NM-PLSPM), Russolillo (2012) [56], which is based on the implementation of the *optimal scaling method* applied to PLS algorithms. The NM-PLSPM is able to quantify non-metric variables to allow us the analysis of non-metric variables together with variables measured at a higher scale level and also to take consideration of standard statistical indexes. The optimal scaling is a scaling technique which can be formulated according to the following criterion informed by Young (1981) [71]: "*Optimal scaling is a data analysis technique which assigns numerical values to observation categories in a way which maximizes the relation between the observations and the data analysis model while respecting the measurement character of the data*". The resulting scaling obtained adopting the optimal scaling criterion must be *suitable* because it must respect the constraint that has to be preserved among the properties of the original measurement scale and it is *optimal* as it optimizes the analysis in which it is involved.

NM-PLS extends the applicability of PLS methods to data measured on different measurement scales, as well as to variables linked by non-linear relationships. A distinctive feature of these algorithms is that they provide a new metric both to non-metric and to metric variables.

3.1 Measurement scales

One of the main features of the available data is the scale of measurement; that is how variables are measured. According to Stevens (1946) [61] there are four different scales of measurement: *interval, ratio, nominal and ordinal* scales. Following Russolillo (2012) [56], the four different measurement scales are grouped in two classes: *metric* and non-metric.

Metric variables : variables observed on interval or ratio scales. These variables have a unit of measurement, they have metric structures and then it is possible to calculate the distance among elements.

- **Interval scale**: Contains categories in which the actual distances, or intervals, between categories can be compared. Differences between numbers anywhere on the scale are the same. For example, we can say that the difference between ages 20 and 25 is the same as the difference between ages 50 and 55.
- **Ratio scale**: Like the interval scale variable, however it has a non-arbitrary zero value. The zero point represents the absence of the property being measured. This property implies that equalities between ratios can be assessed. An example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money.

Non-Metric variables : variables observed on nominal or ordinal scales.

- **Nominal scale**: each number (or names) defines a particular group of units where the categories cannot be ranked. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories; the hair color is also a categorical

variable having a number of categories (blonde, brown, brunette, red, etc.) and again, there is no agreed way to order these from highest to lowest.

- **Ordinal scale**: numbers (or names) can be ranked, such as *low*, *medium* and *high*. Examples are favorite type of music (classical, country, folk, jazz, rock), and favorite place to shop (local mall, local downtown, Internet, other).

3.1.1 The NM-PLSPM algorithm

The name Non-Metric comes from the capability of these models to provide optimally scaled data ($\hat{\boldsymbol{x}}$) with a new metric structure, which does not depend on the metric properties of the raw data ($\boldsymbol{x}^*$). In other words, NM-PLS methods yield a metric to non-metric data, and a new metric to metric data, linearizing the relations between variables and latent constructs, as required by the hypotheses of standard PLS models. The NM-PLS algorithms optimize criteria under two sets of parameters: the model parameters and the scaling parameters constrained to the restrictions due to the scaling level chosen for each raw variable $\mathbf{x}^*$.
In the NM-PLS framework the quantifications are not determined by an external criterion but are obtained by the optimal quantifications method with respect to a latent construct called *Latent Criterion (LC)* which is represented by an unknown vector (centered by construction), for which we use the generic symbol $\boldsymbol{\gamma}_{x^*}$. For the NM-PLS, three levels of scaling are adopted according to measurement scale of the variables: nominal, ordinal and polynomial (or functional).
A scaling (numeric) value, Russolillo (2012) [56], is assigned to each of the K categories (or distinct values) $\phi_k\,(k = 1, \ldots, K)$ of $\mathbf{x}^*$, such that:

- it is coherent with the chosen scaling level;
- it optimizes the model criterion.

In this way, each raw variable $\mathbf{x}^*$ is transformed as $\hat{\mathbf{x}} \propto \tilde{\mathbf{X}}\boldsymbol{\phi}$ where $\boldsymbol{\phi}' = (\phi_1, \ldots, \phi_K)$ is the vector of optimal scaling parameters. The matrices $\tilde{\mathbf{X}}$ are the indicator matrices of the different categories of variables and they define a space in which the

constraints imposed by the scaling level are respected. For example, at nominal scale level grouping property is preserved while ordinal scale level preserves grouping and order properties. The symbol $\propto$ means that the left side of the equation corresponds to the right side normalized to unitary variance. The raw data $\mathbf{x}^*$ are transformed by different real functions (scaling functions) $\mathcal{Q}(\mathbf{x}^*\boldsymbol{\phi}, \gamma_{x^*})$, one for each scaling level, which generate the optimal scaled value $\hat{\mathbf{x}}$ for each observation. The scaling functions $\mathcal{Q}$ optimize the criterion

$$\arg\max_{\phi} \text{cor}^2 \left(\tilde{\mathbf{X}}\phi, \boldsymbol{\gamma}_{\boldsymbol{x}^*}\right) \tag{3.1}$$

under the constraints chosen for the $\mathbf{x}^*$.
The geometric representation of the scaled variable $\hat{\mathbf{x}}$, normalized to unitary variance, can be obtained projecting $\gamma_{\mathbf{x}^*}$ on the space defined by the columns of $\tilde{\mathbf{X}}$.
In *(new) Mode A* PLS-PM, there is one relevant LC for each block of manifest variables $\boldsymbol{X}_q$, i.e. the corresponding inner estimate $\boldsymbol{z}_q = \sum_{q'} c_{qq'} e_{qq'} \tilde{\mathbf{X}}_{q'} \boldsymbol{\omega}_{q'}$. All the LCs are then expressed using the generic notation $\gamma = f(\boldsymbol{\omega}_q)$. In Sec.2.2.3 we discussed the standard PLSPM algorithm. In this section we will discuss the Non-Metric PLSPM algorithm which is an extension of PLSPM to non metric data. Non-Metric PLS-PM loop differs from the standard PLS-PM loop because the iterative procedure contains a new stage called "quantification" phase. Another difference is that it starts by initializing the inner estimate of each LV to obtain a first scaling of the MVs. Then the algorithm go on like the standard PLSPM, see Fig. 3.1 and Tab. 3.1.
In the quantification stage each raw MV $\boldsymbol{x}^*_{pq}$ is maximally correlated to the corresponding LV inner estimate $\boldsymbol{\nu}_q$. Each level of scaling has a corresponding *ad hoc* scaling function $\mathcal{Q}$, which is the projection operator of the LC in a suitable geometrical space spanned by $\tilde{\mathbf{X}}$-columns. While nominal and ordinal scaling involve the quantification of numerals (i.e. numeric labels with no quantitative meaning), polynomial scaling exclusively addresses non-linearity, as it involves the transformation of a metric raw variable.

Nominal scaling if a raw MV is analysing at a nominal scale level, a variable is

quantified as the orthogonal projection of the LC $\gamma_{\mathbf{x}^*}$ linked to $\mathbf{x}^*$ on the space spanned by the columns of the indicator matrix $\tilde{\mathbf{X}}^n$("n" is for nominal scaling) generated by the K categories of $\mathbf{x}^*$. The quantification function is:

$$\mathcal{Q}(\tilde{\mathbf{X}}^n, \boldsymbol{\gamma}_{\mathbf{x}^*}) = \tilde{\mathbf{X}}^n(\tilde{\mathbf{X}}^{n\prime}\tilde{\mathbf{X}}^n)^{-1}\tilde{\mathbf{X}}^{n\prime}\boldsymbol{\gamma}_{\mathbf{x}^*} \tag{3.2}$$

In according to the grouping constraint, for each pair of observations i and i',

$$(x_i^* \sim x_{i'}^*) \Rightarrow (\hat{x}_i = \hat{x}_{i'}) \tag{3.3}$$

where the symbol $\sim$ indicates membership in the same category. The scaling function $\mathcal{Q}(\tilde{\mathbf{X}}^n, \boldsymbol{\gamma}_{\mathbf{x}^*})$ maximizes the Eq. (3.1). The resulting scaling values for the different $\mathbf{x}^*$ are the least square regression coefficients of $\tilde{\mathbf{X}}^n$ on $\boldsymbol{\gamma}_{\mathbf{x}^*}$ which correspond to the average of $\gamma_{\mathbf{x}^*}$ conditioned to $\mathbf{x}^*$ categories. The scaled variable contains the LC values predicted by the regression of $\boldsymbol{\gamma}_{\mathbf{x}^*}$on $\tilde{\mathbf{X}}^n$.
Note that the outer weights of a variable is a function of its correlation with the corresponding LC:

$$\boldsymbol{\omega}_q \propto \mathrm{cor}(\boldsymbol{\gamma}_{\boldsymbol{x}*}, \hat{\boldsymbol{x}}) = \eta_{\boldsymbol{\gamma}_{\boldsymbol{x}*}|\boldsymbol{x}*} \tag{3.4}$$

where $\eta_{\boldsymbol{\gamma}_{\boldsymbol{x}*}|\boldsymbol{x}*}$ is the Pearson's correlation ratio which represent, through a linear correlation, the relationship between $\boldsymbol{\gamma}_{\boldsymbol{x}}*$ and $\boldsymbol{x}*$.

Ordinal scaling when a raw MV is analysing at a ordinal scale level, the quantification function is

$$\mathcal{Q}(\tilde{\mathbf{X}}^o, \boldsymbol{\gamma}_{\mathbf{x}^*}) = \tilde{\mathbf{X}}^o(\tilde{\mathbf{X}}^{o\prime}\tilde{\mathbf{X}}^o)^{-1}\tilde{\mathbf{X}}^{o\prime}\boldsymbol{\gamma}_{\mathbf{x}^*} \tag{3.5}$$

where $\tilde{\mathbf{X}}^o$ ("o" is for ordinal scaling) is constructed according to the Kruskals secondary least squares monotonic transformation of $\mathbf{x}^*$, Kruskal (1964) [45]. The ordering group constraint, for each pair of observations i and i',

$$(x_i^* \sim x_{i'}^*) \Rightarrow (\hat{x}_i = \hat{x}_{i'}) \text{ and } (x_i^* \prec x_{i'}^*) \Rightarrow (\hat{x}_i \leq \hat{x}_{i'}) \tag{3.6}$$

where the symbol $\prec$ indicates empirical order, is preserved in the optimal

scaling values which are contained in the vector of the regression coefficient $(\tilde{\mathbf{X}}^{o\prime}\tilde{\mathbf{X}}^{o})^{-1}\tilde{\mathbf{X}}^{o\prime}\boldsymbol{\gamma}_{\mathbf{x}^*}$. In this case the interpretation of the outer weights is the following:

$$\boldsymbol{\omega}_q \propto \text{cor}(\boldsymbol{\gamma}^*_{\mathbf{x}}, \hat{x}) = \begin{cases} \sqrt{1 - \text{STRESS}^2_{(\boldsymbol{\gamma}^*_{\mathbf{x}}, \mathbf{x}^*)}} & \text{if } \text{cor}(\boldsymbol{\gamma}^*_{\mathbf{x}}, \hat{\mathbf{x}}) \geq 0 \\ -\sqrt{1 - \text{STRESS}^2_{(\boldsymbol{\gamma}^*_{\mathbf{x}}, \mathbf{x}^*)}} & \text{if } \text{cor}(\boldsymbol{\gamma}^*_{\mathbf{x}}, \hat{\mathbf{x}}) < 0 \end{cases} \tag{3.7}$$

The $STRESS$ index is bounded between 0 and 1 and it measures the deviation of the relationship between $\mathbf{x}^*$ and $\boldsymbol{\gamma}^*_{\mathbf{x}}$ from the assumption of monotonicity. The type of monotonic transformation influences the sign of the $\text{cor}(\boldsymbol{\gamma}^*_{\mathbf{x}}, \hat{\mathbf{x}})$:

- if $\text{cor}(\boldsymbol{\gamma}^*_{\mathbf{x}}, \hat{\mathbf{x}}) = 1$ there is a perfect increasing monotonic relationship between $\boldsymbol{\gamma}^*_{\mathbf{x}}$ and $\mathbf{x}^*$;
- if $\text{cor}(\boldsymbol{\gamma}^*_{\mathbf{x}}, \hat{\mathbf{x}}) = -1$ there is a perfect decreasing monotonic relationship between $\boldsymbol{\gamma}^*_{\mathbf{x}}$ and $\mathbf{x}^*$.

Polynomial scaling can be adopted when there is an advance knowledge of the degree D of the polynomial relationship between a raw numerical variable and the LC. The quantification matrix is

$$\mathcal{Q}(\tilde{\mathbf{X}}^p, \boldsymbol{\gamma}^*_{\boldsymbol{x}}) = \tilde{\mathbf{X}}^p(\tilde{\mathbf{X}}^{p\prime}\tilde{\mathbf{X}}^{o})^{-1}\tilde{\mathbf{X}}^{p\prime}\boldsymbol{\gamma}^*_{\mathbf{x}} \tag{3.8}$$

where, according to Young (1981) [71], optimal parameters for the polynomial transformation are found by projecting $\boldsymbol{\gamma}^*_{\mathbf{x}}$ on the space spanned by the columns of the matrix $\tilde{\mathbf{X}}^p$ ("p" is for polynomial) which is built with one row for each observation and with $D + 1$ columns, each column being an integer power of the vector $\mathbf{x}^*$. The special case in which the results obtained with the NM-PLS methods are the same as obtained with the standard PLS methods (applied to standardized data) is when we assume that $D = 1$, i. e. the raw variable and the LC are linked by a linear relationship.

It is important to note that the quantification function (3.2), 3.5 and (3.8) cannot be directly applied to raw variables because the LC is unknown by definition. This is the reason why model and scaling parameters are alternately estimate in a modified PLS

loop with an added quantification phase. Non-Metric PLSPM algorithm maximizes the criterion alternating:

- a quantification phase: scaling parameters are estimated for given model parameters and raw variables, in other words the criterion is optimized with respect to $\hat{X}_q$ keeping $\boldsymbol{\omega}_q$ fixed.
- classical PLS-PM loops: the model parameters are estimated for given scaling parameters, that is the criterion is optimized with respect to $\boldsymbol{\omega}_q$ keeping $\hat{X}_q$ fixed.

Once we get the quantified variables, the standard PLS loop starts: the LVs are first estimated in the outer estimation process, and then re-estimated in the inner estimation process. After obtaining new inner estimates of the LVs, another iteration starts with a new quantification of the MVs, and the algorithm goes on until convergence.

3.1.2 Optimizing Criteria

In the case of Non-Metric PLS-PM framework as well as in the PLS-PM as seen in Section 2.2.6, it is impossible to discuss the optimization of an overall single scalar function.
The NM-PLSPM approach is based on the concept of optimal scaling which considers all the observations as categorical representing them by a scaling parameter $\boldsymbol{\phi}_{pq}$ which is subject to constraints induced by the characteristics of the original variables. As shown by Russolillo (2012) [56], the NM-PLSPM approach, when *New Mode A* is used, optimizes the following criterion:

$$\arg \max_{\forall \boldsymbol{\omega}_q, \boldsymbol{\phi}_{pq}} \sum_q c_{qq'} g\left(\text{cov}(\hat{\boldsymbol{X}}_q \boldsymbol{\omega}_q, \hat{\boldsymbol{X}}_{q'} \boldsymbol{\omega}_{q'})\right) \quad s.t \quad ||\boldsymbol{\omega}_q|| = \sqrt{n},\ ||\hat{\boldsymbol{x}}_{pq}|| = \sqrt{n} \quad (3.9)$$

where $g(\cdot)$ is the square function if the factorial scheme is used, and the absolute value function if the centroid scheme is used. This criterion requires two sets of parameters to be optimized: the model parameters and the scaling parameters. The algorithm alternately optimizes the criterion (3.9) with respect to each subset,

Table 3.1: Non-Metric PLS Path Modeling algorithm

Input: $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q]$, i. e. Q blocks of centred manifest variables
Initialize: the inner weights, i. e. $\boldsymbol{z}_q = \mathbf{1}$ and $s = 0$

Step 1: Iterative algorithm

1: s=s+1
2: **for all** $q = 1, \ldots, Q$ **do**
3: **Quantification phase:**
3.1: the raw MV $\hat{\mathbf{x}}$ is calculated: $\hat{\mathbf{x}} = \mathcal{Q}(\boldsymbol{z}_q^{(s)}, \boldsymbol{x}_{pq}^{*(s)})$
3.2: the matrix $\hat{\mathbf{X}}$ of the scaled MV is built: $\hat{\mathbf{X}}^{(s)} = [\hat{\mathbf{x}}_1^{(s)}, \ldots\ldots, \hat{\mathbf{x}}_{P_q}^{(s)}]$
4: Calculate the outer estimation $\boldsymbol{\nu}_q$ of the LVs: $\boldsymbol{\nu}_q^{(s)} \propto \sum_{p=1}^{P_q} \omega_{pq}^{(s)} \hat{\boldsymbol{x}}_{pq}$
5: Updating of the inner weights $e_{q'q}^{(s)}$ choosing among these schemes:

Centroid: $e_{qq'}^{(s)} = sign[cor(\boldsymbol{\nu}_q^{(s)}, \boldsymbol{\nu}_{q'}^{(s)})]$
Factorial: $e_{qq'}^{(s)} = cor(\boldsymbol{\nu}_q^{(s)}, \boldsymbol{\nu}_{q'}^{(s)})$
Path weighting: multiple regression coefficients or correlations.

6: Calculate the inner estimation $\boldsymbol{z}_q^{(s)}$ of the LVs: $\boldsymbol{z}_q \propto \sum_{q'=1}^{Q'} c_{qq'} e_{qq'}^{(s)} \boldsymbol{\nu}_{q'}^{(s)}$
7: Updating of the outer weights $\boldsymbol{\omega}_q$ choosing between:

Mode A: $\boldsymbol{\omega}_q^{(s+1)} = (1/N) \hat{\boldsymbol{X}}'_q \boldsymbol{z}_q^{(s)}$
Mode B: $\boldsymbol{\omega}_q^{(s+1)} = \left(\hat{\boldsymbol{X}}'_q \hat{\boldsymbol{X}}_q\right)^{-1} \hat{\boldsymbol{X}}'_q \boldsymbol{z}_q^{(s)}$

8: **end for**
9: **Stages 1-7 are repeated until convergence** of the outer weights is achieved

i. e. until: $\max |\omega_{pq}^{(s+1)} - \omega_{pq}^{(s)}| < 10^{-5}$

Output Step 1: the outer weights $\boldsymbol{\omega}_q$ e $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \ldots\ldots, \hat{\mathbf{X}}_Q]$

Step 2: Computation

Upon convergence
10: Computation of the LV scores: $\hat{\boldsymbol{\xi}}_q = \sum_{p=1}^{P_q} \omega_{pq} \hat{\boldsymbol{x}}_{pq} \propto \hat{\boldsymbol{X}}_q \boldsymbol{\omega}_q$

11: Computation of the Path Coefficients: $\boldsymbol{\beta}_j = \left(\hat{\boldsymbol{\Xi}}'_{\rightarrow j} \hat{\boldsymbol{\Xi}}_{\rightarrow j}\right)^{-1} \hat{\boldsymbol{\Xi}}'_{\rightarrow j} \hat{\boldsymbol{\xi}}_j$

12: Computation of the loadings: $\hat{\boldsymbol{\lambda}}_{pq} = \text{cor}\left(\hat{\boldsymbol{x}}_{pq}, \hat{\boldsymbol{\xi}}_q\right)$.

Output Step 2: the LV scores $\hat{\boldsymbol{\xi}}_q$, the path coefficients $\boldsymbol{\beta}_j$ and the loadings $\hat{\boldsymbol{\lambda}}_{pq}$.

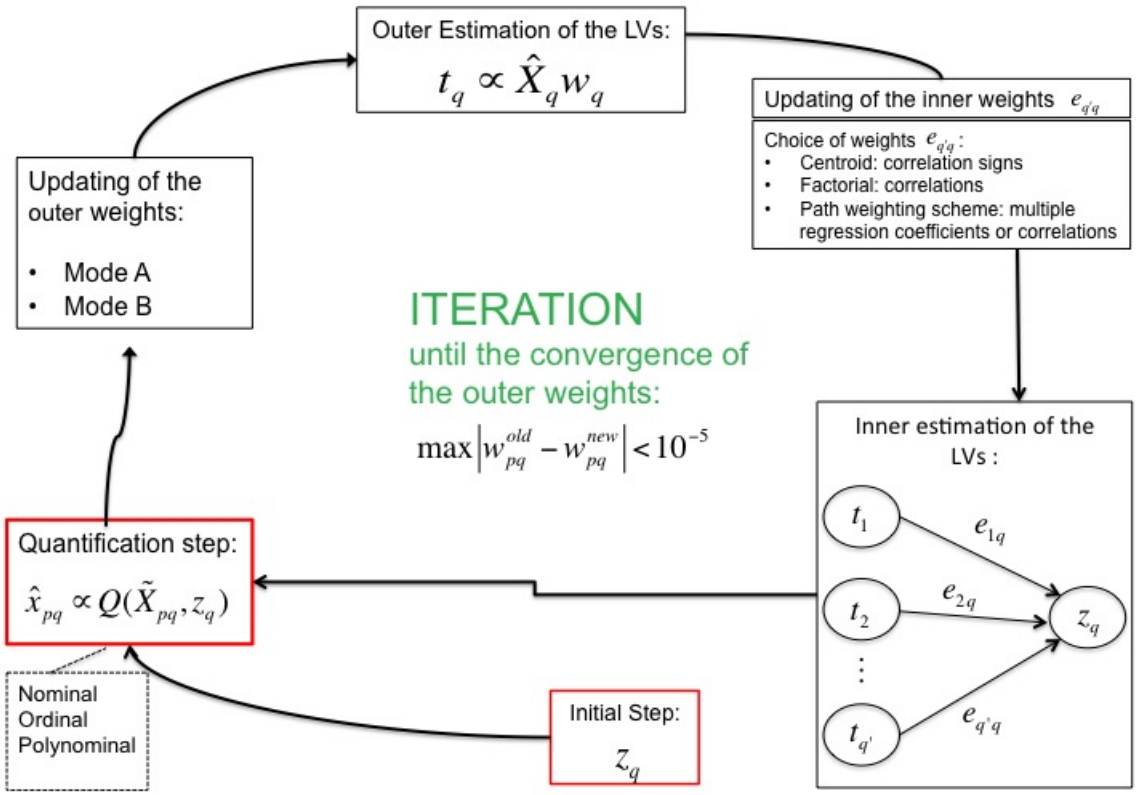


Figure 3.1: Iterative procedure

keeping the other one fixed. In fact, keeping fixed PLS parameters $\boldsymbol{\omega}_q$, the optimal solution for $\boldsymbol{\phi}_{pq}$ was given by the quantification function $\mathcal{Q}(\tilde{\boldsymbol{X}}_{pq}, \boldsymbol{z}_q)$ which orthogonally projects $\boldsymbol{z}_q$ on the space spanned by $\tilde{\boldsymbol{X}}_{pq}$.

In order to extend the results obtained by Russolillo (2012) [56] to *Mode B* and to whatever combination of the modes and schemes, the following optimization problem must be solved:

$$\arg \max_{\forall \boldsymbol{\omega}_q} \sum_q c_{qq'} g\left(\text{cov}(\hat{\boldsymbol{X}}_q \boldsymbol{\omega}_q, \hat{\boldsymbol{X}}_{q'} \boldsymbol{\omega}_{q'})\right) \; s.t \quad ||\hat{\boldsymbol{X}}_q \boldsymbol{\omega}_q|| = \sqrt{n}, \; ||\hat{\boldsymbol{x}}_{pq}|| = \sqrt{n}. \tag{3.10}$$

Tenenhaus and Tenenhaus (2011) [64] showed that this criterion can be written as:

$$\arg \max_{\forall \boldsymbol{\omega}_q, \boldsymbol{\phi}_{pq}} \sum_q \text{cor}(\hat{\boldsymbol{X}}_q \boldsymbol{\omega}_q, \boldsymbol{z}_q). \tag{3.11}$$

Russolillo (2013) [57] has very recently demonstrated that the NM-PLSPM algorithm can be used to maximize this criterion. They showed that for fixed scaling parameters, the optimization problem in Eq. (3.11) can be solved with respect to $\boldsymbol{\omega}_q$ by using the usual PLS-PM iteration; while in order to optimize the problem with respect to $\boldsymbol{\phi}_{pq}$, keeping fixed the PLS parameters $\boldsymbol{\omega}_q$, they propose a *backfitting procedure*, see Breiman and Friedman (1985) [9] and De Leeuw et al. (19875) [20]. The solution of the criterion Eq. (3.11) is then obtained by alternating a standard PLS iterative loop, a back fitting step and a quantification step. At each step the value of the Eq. (3.11) increases, therefore the algorithm is expected to converge to a maximum.
The *plspm* R-package has recently been updated to account for the Non-Metric analysis; the improved version contains also the possibility to choose between Mode A or Mode B.

3.2 Extension to binary endogenous latent variables

The basic objective of a statistical model is to find a mathematical representation of the relationship between the response variable and a set of explanatory variables, along with a measurement of the uncertainty inherent in any relationship.
The use of binary data is necessary when for each unit observed the response variable is dichotomous (when the encoding is present-absent, success-failure, working-broken, dead-alive, and so on). In the case of binary data, it is important to study, as a function of the explanatory variable, both the probability that the response variable assumes a modality rather than another, and the total number of units (proportion, relative frequency) that have a given value of the response variable.
In this thesis the binary model will be adopted in the case of latent variables mea-

sured through only one manifest variable which is of binary type. Therefore, in order to implement the logistic regression for the computation of the path coefficients the inner structure of the Non-Metric PLS-PM will be slightly modified as described in the following Sect. 3.2.4.
In the following, we discuss the logistic regression for binary data and we introduce the ROC curve as a measurement of the quality of the classification obtained by the logistic regression. Together with the ROC curve, we will also discuss the indicator *Area under the curve-AUC* and the Pseudo-R^2's.
The logistic model belongs to the *Generalized Linear Models (GLM)* that constitutes a large class of models including the multiple linear regression model as a special case. In these models, as well as in the classical linear model, the expected value of the response variable depends on a linear combination of the explanatory variables. What distinguishes them from classical linear model is that the response variable may be continuous, discrete quantitative but also qualitative nominal or ordinal.
In this context, it is also possible to drop the assumption about the homoscedasticity of the response variable.
The individual models (linear regression, logit, probit, loglinear, survival, etc..) have been introduced to study real-life situations whenever the classical linear regression model is not adequate. Only in the early 70s, Nelder and Wedderburn (1972) [52], underlining the contact parts among the models, proposed that they could be gathered in a single class. In Appendix A.2 for the sake of completeness, a brief description of the GLM is given.

3.2.1 Logistic Regression

Consider the binary random variables y_i, describing the occurrence of independent events and denote by $y_i = \{0, 1\}$, $i = 1, \ldots, n$ the corresponding realizations; moreover, it is assumed that the vectors $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$ represent the assumed values of $(p + 1)$ random variables. It is supposed in correspondence of the generic i-th statistical unit that y_i is distributed according to a Bernoulli distribution, with parameters p_i:

$$f(y_i; p_i) = {p_i}^{y_i}(1 - p_i)^{1-y_i}, \tag{3.12}$$

where p_i indicates the *success* probability $p_i = \Pr(Y_i = 1) = E(Y_i) = \mu_i$, while the *failure* probability $1 - p_i$ is given by $\Pr(Y_i = 0) = 1 - p_i$.
The mean and variance of the response variable Y are, respectively, equal to:

$$\mathrm{E}(Y_i) = \Pr(Y_i = 1) = p_i, \quad \mathrm{var}(Y_i) = p_i(1 - p_i). \tag{3.13}$$

The covariate vector $\boldsymbol{x}_i$ influences the response variable through a linear predictor which is defined by $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^{p+1}$, $\mathcal{B}$ is the set of admissible parameters.
It is important to note that the function (3.12) represents an element of the simple exponential family, see Appendix A.1 because the density function of the exponential family Eq.(A.4) can be rewritten as $f(y_i; \theta_i)$ when the canonical parameters θ_i are defined as

$$\theta_i = \log\left(\frac{p_i}{1 - p_i}\right) = \log\left(\frac{\mu_i}{1 - \mu_i}\right). \tag{3.14}$$

In fact:

$$\begin{aligned} f(y_i; \theta_i) &= {p_i}^{y_i}(1 - p_i)^{1 - y_i} \\ &= \exp\left\{\log(1 - p_i) + y_i \log\left(\frac{p_i}{1 - p_i}\right)\right\} \\ &= \exp\left\{y_i \log\left(\frac{p_i}{1 - p_i}\right) - \log\left(\frac{1}{1 - p_i}\right)\right\} \\ &= \exp\left\{y_i\theta_i - \log[1 + \exp(\theta_i)]\right\} \end{aligned} \tag{3.15}$$

where the functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are defined through the following relations

$$a(\phi) = \phi = 1, \quad b(\theta_i) = \log[1 + \exp(\theta_i)], \quad c(y_i, \phi) = 0. \tag{3.16}$$

The function that connects θ_i to μ_i in the Eq. (3.14) is also called *link logit function*, that is $\mathrm{logit}(p) \equiv \log\left(\frac{p}{1-p}\right)$ con $p \in (0, 1)$. The *logit model* is defined by taking the canonical parameter θ_i equal to the linear predictor η_i:

$$\mathrm{logit}(p_i) = \mathrm{logit}(\Pr(Y = 1)) = \log\left[\frac{p_i}{1 - p_i}\right] = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i, \tag{3.17}$$

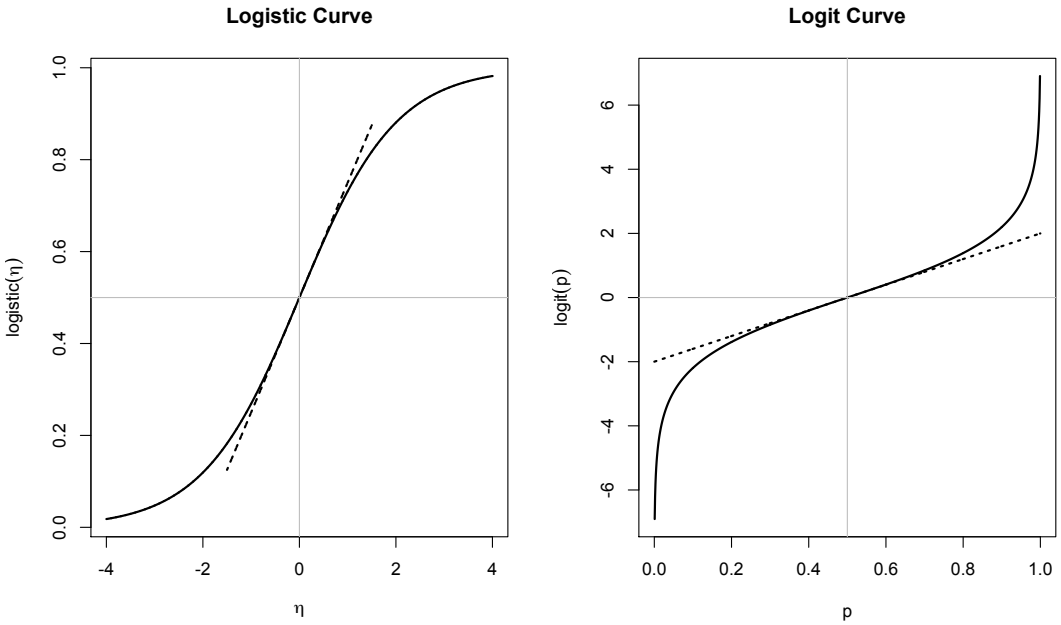


Figure 3.2: The logit and the logistic function. The linearization of the curves is shown.

where the quantity $\frac{p_i}{1-p_i}$ is called *odds* of the event, while the logarithm of the odds of success is called *logit* of the success probability. To every value of p in the interval $(0, 1)$ corresponds a value of $\text{logit}(p)$ in the interval $(-\infty, \infty)$. The $\text{logit}(p)$ is a sigmoid curve that is symmetric with respect to $p = 0.5$ and it is approximately linear for p in the range $(0.2, 0.8)$, while outside this range has a significantly non-linear trend, see Fig.3.2.

The inverse of the logit function Eq. (3.17) is called *logistic function*:

$$\text{logistic}(\eta) = \frac{e^{\eta}}{1 + e^{\eta}} \tag{3.18}$$

and therefore

$$p_i = \Pr(Y_i = 1) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} = \frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}}}. \tag{3.19}$$

Calling π_1 the probability of an event E_1 (e. g. a man chosen at random has a given disease) then the ratio $\frac{\pi_1}{1-\pi_1}$ is the *odds of the occurrence of this event* (e. g.

the ratio between the probability have a disease and the probability to not have a disease). If we consider π_2 be the probability of another event E_2 (e. g. a woman chosen at random has the same disease), then we introduce the *odds ratio* (OR)

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

and the *log-odds ratio* (log-OR)

$$\log - \mathrm{OR} = \log\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}\right) = \mathrm{logit}(\pi_1) - \mathrm{logit}(\pi_2).$$

If the odds ratio is equal to 1, the odds of the event are the same in the two groups (men and women), and then the occurrence of the event is independent of the explanatory variable that distinguishes the two groups. Therefore, the logs-odds ratio is a measure of the difference between the odds, interpretable in terms of comparison between the probability π_1 and π_2. Note that the odds may be greater than 1.
The *log-odds ratio* is generally used instead of the *odds ratio* because the estimator of the OR is strongly asymmetric with a distribution in $[0, \infty)$ while the estimator of the log-OR is in $(-\infty, \infty)$ and its distribution is asymptotically normal, for sample sizes smaller than those required for the asymptotic distribution of the OR.
The logistic regression model, for a binary response variable Y with a single explanatory variable X has the form

$$\mathrm{logit}[\pi(\mathrm{x})] = \log\left(\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})}\right) = \beta_0 + \beta_1 \mathrm{x} \tag{3.20}$$

The interpretation of the logistic regression coefficient β_1 can be given in terms of odds ratio in this way. When x is continuos variable, increasing its value of one-unit $x^* \to x^* + 1$ we get

$$\mathrm{logit}(\tilde{\pi}) = \beta_0 + \beta_1(\mathrm{x}^* + 1) \tag{3.21}$$

that can be written in this way,

$$\frac{\tilde{\pi}}{1-\tilde{\pi}} = e^{\beta_0+\beta_1(x^*+1)} = e^{\beta_0+\beta_1 x^*} e^{\beta_1} = \frac{\pi}{1-\pi} e^{\beta_1} \tag{3.22}$$

therefore e^{β_1} represents the factor that multiplies the odds when the variable x is increased of one-unit. Another way to say the same thing is that e^{β_1} is the odds ratio for one-unit increase for the variable x. In multiple logistic regression analysis is usual to quote the odds ratio for each variable in decreasing order to underline the relative importance of the contribution of each explanatory variables in the model. It is worth to note that this result does not depend on the value of x^*. So for an increase equals to c corresponds to an increase in the logarithm of odds ratio of $\beta_1 c$ or or equivalently, a multiplicative increase in the odds of the probability equal to $\exp(\beta_1 c)$. The coefficients β_0 represents a basic level of the odds of the occurrence of an event, regardless of the values and the modalities of the explanatory variables (called *background odds*).

The logistic regression is widely adopted in many contexts. It is often used to represent those phenomena of growth and development that find in the system a brake to an unlimited growth, due to constraints which are proportional to the same growth. For these reasons, it is adopted for the dynamics of living populations (e.g. demographic development), the effects of a bacterial infection, the penetration of pollutants, the diffusion of new technologies and so on. The models based on the logistic transformation are particularly appropriate for the analysis of data collected in a retrospective manner (as in the epidemiology field in the case-control studies [1] with a fixed total number of observations for both values of the response variable: case ($Y = 1$) or control ($Y = 0$).

Moreover it is extensively adopted in engineering (probability of failure of a process or product, in marketing, customer's propensity to purchase a product), etc. It is worth to note that besides the practical interpretation in terms of the logarithm of the odds of success, the logistic transformation has the theoretical advantage to identify a canonical link and therefore a *sufficient statistic* for $\boldsymbol{\beta}$.

To prove this statement, consider the likelihood function for the logistic model

[1] The terms of case-control, in epidemiology field, have different meanings based on the occurrence of an event or exposure to a risk factor (cohort study).

Eq. (A.9) specified for the case of the Bernoulli distribution:

$$\mathscr{L}(\boldsymbol{y}, \boldsymbol{p}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{3.23}$$

where p_i is obtained using the Eq. (3.17)

$$\mathrm{E}(Y_i) = p_i = \Pr(Y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{\exp({\boldsymbol{x}_i}^T \boldsymbol{\beta})}{1 + \exp({\boldsymbol{x}_i}^T \boldsymbol{\beta})}. \tag{3.24}$$

Substituting the expression Eq. (3.24) in Eq. (3.23)

$$\begin{aligned} \mathscr{L}(\boldsymbol{y}, \boldsymbol{p}) &= \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right]^{y_i} \left[1 - \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right]^{1-y_i} \\ &= \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta} y_i)}{[1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})]^{y_i}} \frac{1}{[1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})]^{1-y_i}} \right\} \\ &= \frac{\exp(\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\beta} y_i)}{\prod_{i=1}^{n} [1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})]} \end{aligned} \tag{3.25}$$

and, defining $t_j = \sum_{i=1}^{n} \boldsymbol{x}_i^T y_i$ the likelihood function becomes

$$\begin{aligned} \mathscr{L}(\boldsymbol{y}, \boldsymbol{p}) &= \frac{\exp(\sum_{j=0}^{p} \beta_j t_j)}{\prod_{i=1}^{n} [1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})]} \\ &= \frac{\exp(\boldsymbol{t}^T \boldsymbol{\beta})}{\prod_{i=1}^{n} [1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})]}. \end{aligned} \tag{3.26}$$

Applying the criterion of Neyman factorization it can be seen that the components of the vector $\boldsymbol{t}$ form a set of jointly sufficient statistics for the vector $\boldsymbol{\beta}$.
We will now introduce a general method for simultaneous inferences. In this case it is an asymptotic method, whose sampling properties are valid for samples of sufficiently large size. This method is based on the natural logarithm of likelihood function Eq. (3.26)

$$l(\boldsymbol{\beta}|\boldsymbol{t}) = \sum_{j=0}^{p} t_j \beta_j - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})] \tag{3.27}$$

This equation depends only on the sufficient statistics.
In inferential procedures based on the maximum likelihood it is necessary to determine the first and second derivatives of $L(\boldsymbol{\beta}|\boldsymbol{t})$, in our case we have:

$$\frac{\partial l(\boldsymbol{\beta}|\boldsymbol{t})}{\partial \beta_j} = t_j - \sum_{i=1}^{n} \frac{x_{ij}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}, \quad j=0,\ldots,p \tag{3.28}$$

$$\frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{t})}{\partial \beta_u \partial \beta_\nu} = -\sum_{i=1}^{n} \frac{x_{iu}x_{i\nu}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{[1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}, \quad u,\nu=0,\ldots,p. \tag{3.29}$$

Note that the second derivatives are no longer dependent variables on the t_j. Then we calculate the information matrix of the parameters $\boldsymbol{\beta}$, whose generic term is defined $F_{u,\nu}(\boldsymbol{\beta}) = \mathrm{E}\left\{-\frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{t})}{\partial\beta_u\partial\beta_\nu}|\boldsymbol{\beta}\right\}, \quad u,\nu=0,\ldots,p.$ In our case the information matrix will be defined by the following expression

$$\begin{aligned}\mathbf{F}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{[1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2}\boldsymbol{x}_i\boldsymbol{x}_i^T \\ &= \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X},\end{aligned} \tag{3.30}$$

where $\boldsymbol{W} = \mathrm{diag}\left\{\frac{\partial h(\eta_i)}{\partial \eta_i}\right\}$.

The maximum likelihood estimators for the Logistic Model

In the following discussion we will determine the estimators for the logistic model. Under some regularity conditions, the maximum point of this function is the solution of the system of nonlinear equations.

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{t})}{\partial \beta_j} = 0 \quad j=0,\ldots,p. \tag{3.31}$$

A well-known theorem on the uniqueness of the maximum likelihood estimators in the presence of sufficient statistics, states that the solution to Eq. (3.31), if it exists, is unique. The system Eq. (3.31) can be re-written:

$$t_j - \sum_{i=1}^{n} \frac{x_{is}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} = 0, \quad s=0,\ldots,p \tag{3.32}$$

remembering that $p_i = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}$ we get

$$\begin{aligned} t_j - \sum_i x_{ij} p_i &= 0 \\ \sum_{i=1}^{n} x_{ij} y_i &= \sum_{i=1}^{n} x_{ij} p_i, \quad j = 0, \ldots, p. \end{aligned} \tag{3.33}$$

Defining with $\hat{\boldsymbol{\beta}}(\mathbf{t})$ the system solution Eq. (3.32) in terms of $\boldsymbol{\beta}$, and with $\hat{\boldsymbol{\beta}}(\mathbf{T})$ the some solution seen as random variable in function of $\mathbf{t}$. In the theory of maximum likelihood estimation is demonstrated that

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} MVN_{p+1}(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta})), \tag{3.34}$$

For sufficiently numerous samples we can approximate the sampling distribution of $\hat{\boldsymbol{\beta}}(\mathbf{T})$ thought a multinormal distribution $MVN_{p+1}(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$ whose mean vector is provided by the maximum likelihood estimate of $\boldsymbol{\beta}$ and whose dispersion matrix is given by the inverse of the information matrix Eq. (3.30) of the maximum likelihood estimate calculated in the defined point. On the basis of literature, it can be stated that the maximum likelihood estimators are consistent and asymptotically efficient. It is possible to construct confidence regions, asymptotically optimal, for the parameters $\hat{\boldsymbol{\beta}}$ using the asymptotic multinormality estimators in question.

3.2.2 Descriptive measures of fit

In ordinary linear regression, the primary measure of the fit quality is the value of R^2, which is an indicator of the percentage of variance related to the dependent variable explained by the model. The R^2 index is only appropriate for linear regression, with continuous dependent variables. It would be useful to have a similar index for logistic regression. To get around this problem, a number of so-called '*Pseudo-*R^2' indicators have been developed by taking different conceptual approaches regarding the meaning to be given to the pseudo-R^2 for logistic regression models.
The most common pseudo-R^2s are:

- Cox and Snell's Pseudo-R^2: it is based on the evaluation of the proportion of

unexplained variance that is reduced by adding k variables to the*null* or *empty* model i.e. with constant term alone, where l_{null} is the log-likelihood for the empty model, and l_k is the log-likelihood for the model with k explanatory variables and n is number of cases:

$$R^2 = 1 - \left[\frac{l_{null}}{l_k}\right]^{\frac{2}{n}}. \tag{3.35}$$

The major problem with this measure is that

$$\text{Max } R^2 = 1 - [l_{null}]^{\frac{2}{n}} < 1$$

making it difficult to be interpreted.

- Nagelkerkes $Pseudo - R^2$ (called also adjusted $Pseudo - R^2$): it is a modified version of Cox and Snell's pseudo-R^2 that varies from 0 to 1. Nagelkerke's pseudo-R^2 divides Cox and Snell's Pseudo-R^2 by its maximum value:

$$R^2 = \frac{1 - \left[\frac{l_{null}}{l_k}\right]^{\frac{2}{n}}}{1 - [l_{null}]^{\frac{2}{n}}}. \tag{3.36}$$

Therefore Nagelkerke's pseudo-R^2 value will normally be higher than the Cox and Snell's one.

- Likelihood ratio tests: is defined as

$$D = -2 \cdot \ln\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right] \tag{3.37}$$

where D is deviance of the model. This quantity plays the same role that the residual sum of squares in linear regression. Large values suggest that the independent variables are helpful in predicting the response. In Eq. (3.37) the numerator is the value of the likelihood when the parameters are set equal to maximum likelihood estimates. It summarize the extent to which the sample data are fitted by this model (generally called *fitted or current model*). The denominator is the likelihood of the reference model that is the *saturated or full*

model i.e. the one for which the fitted values coincide with the all observations. It has a number of parameters equal to the number of observations. In the case of binary regression the denominator is equal to one.

3.2.3 Model validation: the ROC curve

Receiver operating characteristics (ROC) graphs are useful for organizing classifiers and visualizing their performance. It is a technique for visualizing, organizing and selecting classifiers based on their performance. A very common technique for the validation of a logistic regression model is to use a classification table obtained by crossing the classification of the response variable with a dichotomous variable whose values are derived from the probabilities estimated by the model. The dichotomous variables are obtained by defining a threshold and comparing the estimated probabilities with this threshold which is often 0.5. If the probabilities exceed this threshold the derived variable is classified 1, otherwise 0.
In this approach, the estimated probabilities are used to classify the units inside a group. It is implicitly assumed that if the model accurately predicts the group membership according to some criterion, then it is possible to say that the model estimates the data well. Unfortunately, it is very easy to construct theoretical situations in which the logistic regression model is the correct model, but the classification is bad, Hosmer and Lemeshow (2000) [36]. The goodness of classification is to be considered a complementary criterion with respect to the general goodness of the estimate; the aim is to verify whether the distances between the observed and estimated values are systematic or are in the range of *natural* variation of the model. The ROC graph is a two dimension graph that allows one to view and select the quality of classification models. This graph has long been used in the frame of signal detection theory during the Second World War, and today this technique is commonly adopted in the medical field, in the theory of decisions, in machine learning, data mining, etc.. Fawcett (2006) [26]. Formally, we consider that each unit corresponds to an observed value of the response variable in the observed set $\{\boldsymbol{p}, \boldsymbol{n}\}$ (class {positive, negative} or {1,0}). A classification model (or classifier) is a mapping from units to predicted classes $\{\hat{\boldsymbol{p}}, \hat{\boldsymbol{n}}\}$. Some classification models produce a continuous output (e.g., an estimate of an unit's class membership probability) to

which different threshold values may be applied to predict class membership.
Given an unit and a classifier, there are four possibilities:

1. the unit is positive and is classified as positive, then it is considered a *true positive, TP*;

2. the unit is positive and is classified as negative, then it is considered a *false negative, FP*;

3. the unit is negative and is classified as negative, then it is considered a *true negative, FN*;

4. the unit is negative and is classified as positive, then it is considered a *false positive, TN*.

It is possible to construct a Cross Table (or Confusion Matrix) that has for elements TP, FP,FN and TN, see Fig.3.2.
According to this matrix, is it possible to estimate some important index like the true positive rate (TP-rate, TPr) called also either *recall* or *sensitivity*, the false positivity rate (FP-rate, FPr) and the *specificity*:

$$\text{Sensitivity} = \text{TPr} \approx \frac{\text{positive case accurately classified}}{\text{total positive case}} = \frac{TP}{P} = \text{recall}$$

$$\text{FPr} \approx \frac{\text{case negative inaccurately classified}}{\text{total negative case}} = \frac{FP}{N}$$

$$\text{Specificity} = 1 - FPr = \frac{\text{true negative}}{\text{false positive} + \text{true positive}} = \frac{TN}{FP + TN}$$

and also:

$$\text{Precision} = \frac{TP}{TP + FP} = \text{positive predicted value}$$

$$\text{Accurancy} = \frac{TP + TN}{P + N}$$

$$\text{Measure F} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{sensibility}}}$$

		Observed Class	
		p	n
Estimated Class	p̂	True positive (TP)	False positive (FP)
	n̂	False negative (FN)	True negative (TN)
Total		P	N

Table 3.2: Confusion Matrix where $P = TP + FN$ and $N = FP + TN$.

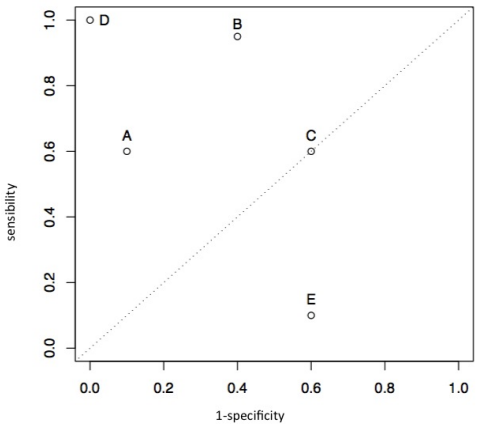


Figure 3.3: A simple ROC graph showing five discrete classifiers.

A ROC curve is a two-dimensional graph where the x-axis shows the false positive rates and the y-axis the true positive rates. A ROC graph depicts the relative trade-offs between benefits (true positives) and costs (false positives). Fig. 3.3 shows a ROC graph with five classifiers labeled A through E. The point (0,0) is the strategy of never producing a positive classification, therefore there is neither an error of false positive nor a gain of true positivity. The opposite strategy to produce more positive classifications corresponds to the point (1,1). The point D = (0,1) corresponds to a perfect classification. A point in a ROC graph is better than another if it is in the north-east (TPr is higher, FPr is lower) of the first. The classifiers that are positioned on the left side of the graph near the x-axis can be considered "conservative" because they do positive classifications only if there is large evidence; thereby committing a few false positive errors at the price of a low rate of true positivity. Classifiers on the top right of the graph can be considered "liberal" because they produce positive classifications with weak evidence, then they classify nearly all positives correctly, but often they have a high rate of false-positive, see Fig. 3.3. The point A is more conservative than B. The diagonal line $y = x$ is the strategy of random selection of the class. For example, if we have a random classifier that estimates the class positive in 70 % of the cases it is expected to get 70 % of the positive correct but also the false-positive rate would be 70 %, producing the point $C = (0.7, 0.7)$, see Fig. 3.3. So a random classifier will produce a point that moves back and forth on the diagonal, depending on the frequency with which it estimates the positive class.

To move from the diagonal, the classifier must exploit some information contained in the data. Note that any classifier which corresponds to points in the triangle below the diagonal develops a classifier quality worse than random selection, so this triangular area is usually empty in ROC graphs. In this case, we can consider the negated of a classifier, i.e. a classifier in which the classification decisions are swapped, so classifications of true positive and false negative errors become vice versa. Each classifier that produces a point in the triangle at the bottom may be negated to produce a point above the diagonal. In Fig. 3.3 point E is a classifier worse than random which corresponds to the negation of A.

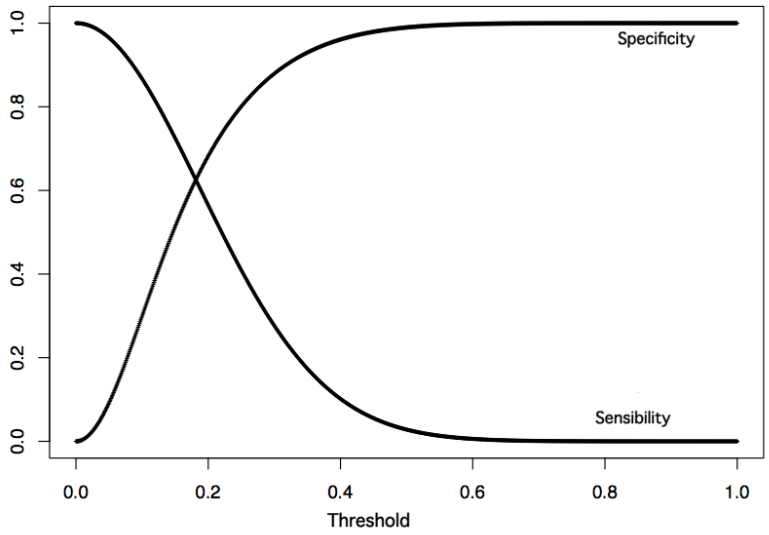


Figure 3.4: Illustrative graph of sensitivity and specificity vs. the threshold.

In order to use the classification tables for judging the goodness of a prediction model it is necessary to set a threshold: for each threshold we have a different classification table. If we plot in a graph the values of sensitivity and specificity as a function of the threshold value (the threshold values often are called *cutpoints*) we obtain, in general, curves such as those shown in Fig. 3.4. Since the goal is to choose an optimum threshold for classification purposes, we might choose the threshold that maximizes both sensitivity and specificity, and therefore, in the case shown in the figure, the optimal threshold would be that corresponding to the point of intersection between the two curves (about 0.2).

The ROC Curve is a different graphic representation obtained plotting the sensitivity, namely the rate of positive events correctly classified (TPr), as a function of the complement to 1 of the specificity that corresponds to the rate of false positives (FPr) (i. e. FPr = 1-specificity) for each threshold value. For example, the data previously shown in Fig. 3.4 produce the ROC curve in Fig. 3.5 where it is also shown the point corresponding to the optimal threshold.

A ROC curve is a two-dimensional description of the quality of discrimination of a classifier. In order to compare different classifiers it is necessary to synthesize

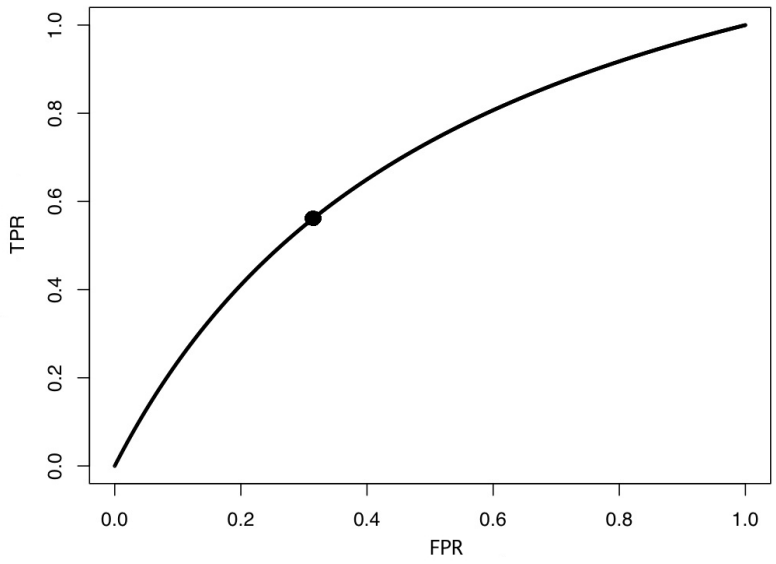


Figure 3.5: ROC curve.

the information contained in the ROC curve using a single value that represents quantitatively the expected *performance* of the classifier. A very common method is to calculate the Area Under the ROC curve, quantity commonly called *AUC, area under curve*, which is a measurement of the quality of discrimination of the estimation model. As AUC is a portion of an area of unit value, its value is between $[0, 1]$, however, because a random choice lies on the diagonal line, no realistic classifier should have a value of less than 0.5 AUC because if the value of AUC is equal to 0.5, the estimation model could not discriminate and it would correspond to a model of random choice. Increasing the AUC value also increases the discrimination ability of the model, and a value between $[0.7, 0.8]$ is considered acceptable, while a value between $[0.8, 0.9]$ is considered excellent.

In practice it is extremely unusual to obtain AUC values greater than 0.9 (for example, in Fig. 3.5 we have AUC= 0.699). An important property is that the value AUC of a classifier is equivalent to the probability that the classifier assigns a higher rank to a unit randomly chosen positive rather than a negative unity chosen at random, that is equivalent to the *Wilcoxon test of ranks*, Piccolo (2002)[55], it is also connected to the *Gini coefficient*, Breiman et al. (1984) [10].

3.2.4 PLS-PM with binary endogenous latent variables

When we have non-metric data and the structure of the path diagram, depicting our model, it is composed of endogenous latent variables represented by a single manifest variable of binary type, it is thinkable to modify the inner model structure of the PLS methods by adopting the logistic regression to estimate the path coefficient. From the standard statistical point of view, the logistic regression models are adopted when it is necessary to drop the general assumptions of the classical regression model concerning the homoscedasticity request and both the normal distribution of the errors and of the response variables. When the endogenous variable is binary the logistic regression is usually adopted because:

1. the logistic transformation allows us to extend the unitary interval $[0, 1]$ to real axis $[-\infty, +\infty]$ in such a way to represent a linear combination of the explicative variables;

2. there is a probability interpretation associated with a Bernoulli distribution;

3. the result can be discussed by odds ratio;

4. it is easy to implement.

The advantages that we have using these models is that the logit allows us to obtain a probabilistic interpretation of the binary response and an analytical expression of the odds ratio, i. e. the ratio between the probability of the two realizations of the Bernoulli random variable:

$$\frac{p_i}{1 - p_i} = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{3.38}$$

where p_i is the probability that an event occurs. This ratio, for given $\boldsymbol{x}_i$, defines a sort of prediction in favor of a given choice. For example, if $p_i = 0.25$, the prediction is $\frac{p_i}{1-p_i} = 0.33$, that is 3 to 1 in favor of the event $y_i = 0$.
In Sect. 2.2.1 we discussed the structure of the inner model for the PLS methods i. e. the relationship among the LVs that are modeled through a multiple regression model, as shown in Eq. (2.1).
The method of computation of the path coefficients depends on the type of available

endogenous LVs.

In this thesis, we assume that the binary endogenous latent variables belong to the exponential family and then we have implemented the possibility of modeling them through logistic regression models replacing the linear multiple regressions Eq. (2.1) with:

$$\text{logit}(\boldsymbol{\xi}_j = 1|\boldsymbol{\xi}_{m\to j}) = \log\left[\frac{\Pr(\boldsymbol{\xi}_j = 1|\boldsymbol{\xi}_{m\to j})}{\Pr(\boldsymbol{\xi}_j = 0|\boldsymbol{\xi}_{m\to j})}\right] = \eta_j \tag{3.39}$$

where the linear predictor

$$\eta_j = \beta_{j0} + \sum_{m=1}^{M_j} \beta_{jm}\boldsymbol{\xi}_{m\to j} \tag{3.40}$$

and β_{jm} are the path coefficients expressing the impact on $\boldsymbol{\xi}_j$ of the predictor $\boldsymbol{\xi}_{m\to j}$, $(m = 1, \ldots, M)$.

For example, for the model described in Fig. 3.6 where the manifest variable x_{13} is binary, the inner model can be written as:

$$\xi_2 = \beta_0 + \beta_3\xi_1 + \nu \tag{3.41}$$

$$\text{logit}[\Pr(\xi_3 = 1|\xi_1, \xi_2)] = \beta_0 + \beta_1\xi_1 + \beta_2\xi_2. \tag{3.42}$$

It is possible to read this path coefficient in terms of odds ratio.

$$\text{logit}(\Pr(\xi_3 = 1|\xi_1, \xi_1)) = \log\ [\text{Odds}] = \log\left[\frac{\Pr(\xi_3 = 1|\xi_1, \xi_2)}{\Pr(\xi_3 = 0|\xi_1, \xi_2)}\right]$$

In this case, the validation of the logistic regression model can be carried out using the ROC curve graphical analysis and the AUC indicator, as described in Sect. 3.2.3.

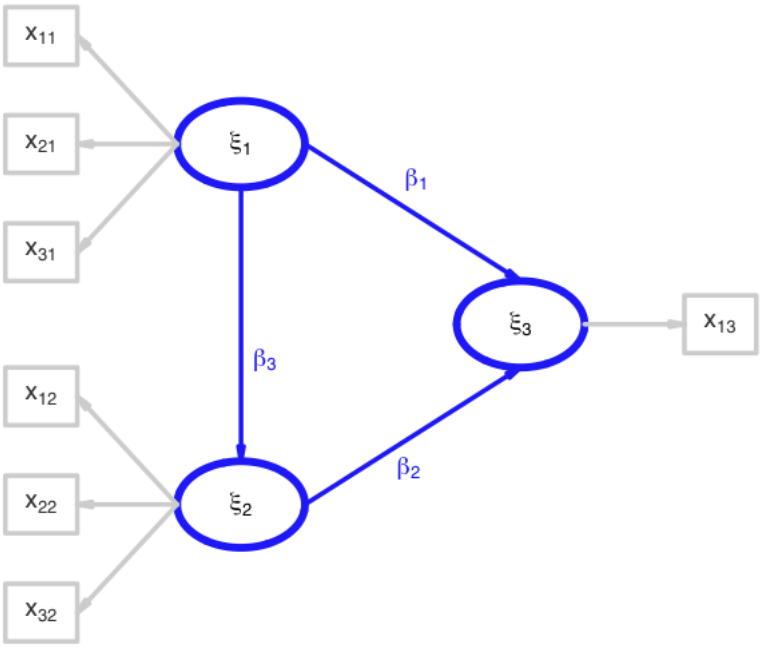


Figure 3.6: Path Diagram where the LV ξ_3 is a binary endogenous latent variable.

Chapter 4

Modeling the UNI.CO dataset

In this Chapter we will discuss how to model the UNI.CO dataset, which contains nominal and ordinal variables, by the Non-Metric PLSPM method in order to extract information about the integration into the labour market of Sapienza alumni. The general framework is the study of the subordinate and para-subordinate employment offered to the Sapienza University of Rome alumni by the Italian labour market based on the UNI.CO dataset. In the following we propose three models which have different purposes. The first one concerns the quantitative study of the job success of Sapienza University alumni in terms of quality of work. The second and the third models have been built in order to study the job career in terms of the contract type and quality evolution in the three years after graduation.
The class of the PLS methods (Esposito Vinzi et al. 2010 [22], Russolillo 2012 [56]) is the most suitable methodology to tackle this kind of problems because of their capability to:

- quantify the latent variables (LVs) representing unobservable constructs;
- provide an estimate of the LVs for each observation;
- work without distributional hypotheses.

The last point is important because in the social sciences it is often the case that the distributions of the variables are asymmetric and very far from the Gaussian distribution.

These analyses were performed using a specialized R package now available called *plspm* improved to handle non-metric data following the NM-PLS procedure based on the optimal scaling method. Details about this package are given in Sanchez, Trinchera and Russolillo (2013) [59] and Sanchez (2013) [58].

4.1 Modeling the job quality

In this Section, we propose a model in order to perform a preliminary quantitative study of the job success of Sapienza University of Rome alumni in terms of quality of work.[1]
The main purposes we addressed in the following are:

1. to study indicators of job success and to estimate their relationship with educational and job curricula.

2. to model job success as a latent variable in PLS-PM framework;

3. to assess the effectiveness of Non-Metric approach, Russolillo (2012) [56], to Partial Least Square Path Modeling in the analysis of variables observed on different measurement scales.

The main aim is to extract from the information contained in the database, indicators that have an impact on the job position after graduation.
The goal is to define and measure the possibility of getting a good job position, i.e., satisfactory, well paid, stable over time, with the possibility of improvements in career, consistent with university curriculum.
Measuring job qualifications is not an easy task, both in absolute or in relative terms, Fabbris (2012) [25]. In literature many indicators have been studied (e. g. index of job desirability Jencks et al. (1988) [39], job quality index Leschke (2008) [46]).
In this Section we are interested in studying the two new composite indicators Alleva

[1] A few preliminary results of this model have been presented at the Statistical Conference "Advances in Latent Variables, Methods, Models and Applications", Petrarca (2013) [53]. The extended version is in phase of publication, Petrarca (2014) [54]. The logistic regression extension is completely new.

and Petrarca (2013) [3] that are related to the possibility of the success in terms of Sapienza alumni best employment status. These indicators quantify the concept of job success using the definition of *optimal* and *quasi-optimal contract* based on the ISCO classification of job quality and on a minimum continuative duration of the job discussed in Sec. 1. In the International Standard Classification of Occupations (ISCO) a highly qualified position is identified with ISCO1 (managers) and ISCO2 (intellectual and scientific professions). The concept of a good job is rather theoretical and it needs a quantification in order to be inferred from data. Our suggestion is to measure the concept of a "good job" defining it as a latent variable in PLS-PM framework.

We take into account a sub-set of the UNI.CO archive: we consider only the alumni with a master's degree of Sapienza University who belong to the engineering disciplinary sector. Moreover we consider only alumni who enrolled in more than one contract during the three years after graduation (458 statistical units). In this way we have chosen to reduce the large dataset with a vast variability of the data to a specific disciplinary sector with a higher degree of homogeneity.

In this study we propose a model in which the Job Success impacts on the Educational and Job curricula. The set of manifest variables for each of the three latent variables representing Job Success, Educational Curriculum (Edu. Curr.) and Job Curriculum (Job Curr.) are described in Tab. 4.1. In the Job Success block only the two composite indicators: Optimal and Quasi-Optimal are included as manifest variables. In our model all the manifest variables are treated as reflective i. e. the LVs are to be considered as the cause of the MVs belonging to its own block. We performed a Non-Metric PLSPM analysis on the model by using the option centroid for the inner weight estimation (this choice only considers the sign of the correlations between a LV and its adjacent LVs). As shown in Fig. 4.1 our model relates Job Success with Educational Curriculum and Job Curriculum and also it analyses the relationship between Educational and Job curricula.

4.1.1 Discussion of the results

We performed a PLS-PM analysis on the model described previously. In our case the convergence of the algorithm has been achieved after only 9 cycles.

Table 4.1: Set of manifest variables for each latent variable. The number of levels for each ordinal variable are reported in brackets.

LVs	**MVs**	**Description**	**Scale**
Edu. Curr.	Age	Age at university graduation	Numerical
	Final grade	Final university grade	Numerical
	Average grade	Average graduation grade	Numerical
	Isee	Indicator of economic equivalent situation: it measures the economic status of the families	Ordinal (5)
Job Curr.	N_cn	Number of job relationships	Numerical
	gg_work	Number of worked days	Numerical
	gg_isco12	Number of worked days with high professional position	Numerical
	gg_al243	Number of worked days with an actual duration of the contract of at least 8 months	Numerical
	gg_CTI	Number of worked days with a permanent contract	Numerical
Job Success	Optimal	The graduate has got optimal contract	Nominal
	Quasi-optimal	The graduate has got a quasi-optimal contract	Nominal

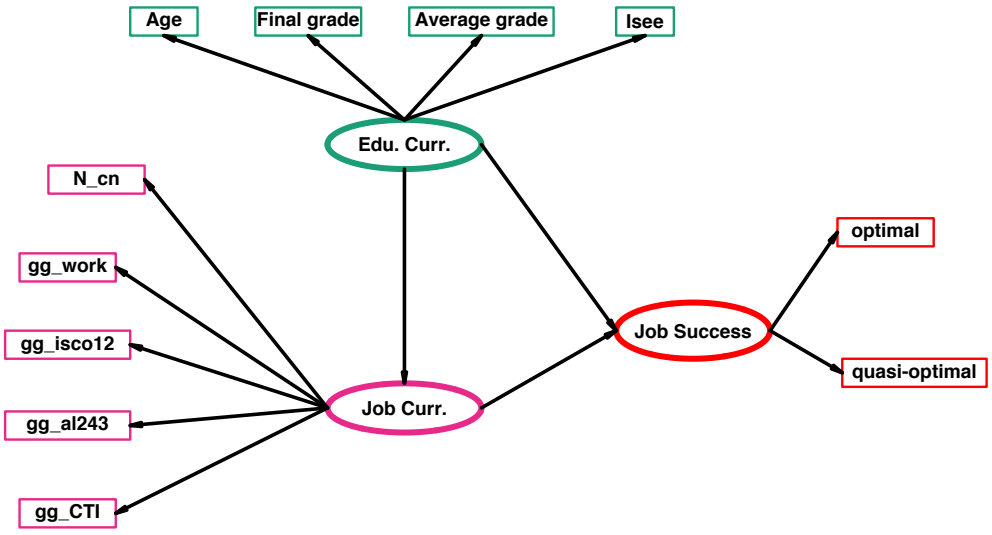


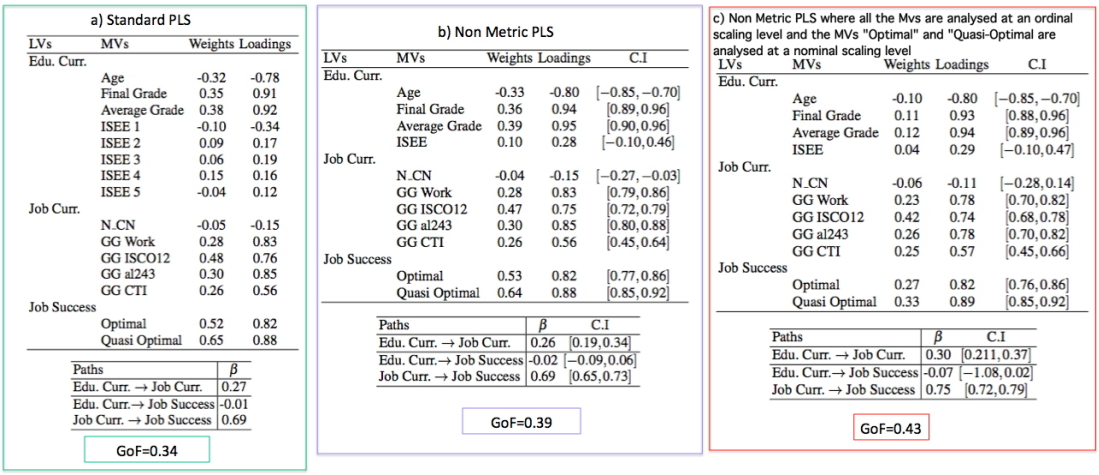
Figure 4.1: Path Diagram depicting our model.

In Fig. 4.2 we report the comparison of the results of the model shown in the path diagram Fig. 4.1 when three different scaling levels are used:

- case a) we used the standard PLS-PM model where the categorical variable *isee* was replaced by the five dummy variables *isee1*, *isee2*, *isee3*, *isee4* and *isee5* corresponding to its categories.

- case b) we performed a Non-Metric PLSPM analysis where the non-metric variable *isee* is properly quantified by the optimal scaling method, in order to overcame the binary coding drawbacks. In this case the variable *isee* is analysed at ordinal scaling level and the variables *optimal* and *quasi-optimal* at the nominal scaling level. For all the other variables we respected the natural scaling level and we analysed them at linear scaling level (i.e. we simply standardized them).

- case c) we performed a Non-Metric PLSPM analysis with new transformations for all the original variables. In this case all MVs are analysed at ordinal scaling level, and the MVs *optimal* and *quasi-optimal* at nominal scaling level.



a) Standard PLS

LVs	MVs	Weights	Loadings
Edu. Curr.			
	Age	-0.32	-0.78
	Final Grade	0.35	0.91
	Average Grade	0.38	0.92
	ISEE 1	-0.10	-0.34
	ISEE 2	0.09	0.17
	ISEE 3	0.06	0.19
	ISEE 4	0.15	0.16
	ISEE 5	-0.04	0.12
Job Curr.			
	N_CN	-0.05	-0.15
	GG Work	0.28	0.83
	GG ISCO12	0.48	0.76
	GG al243	0.30	0.85
	GG CTI	0.26	0.56
Job Success			
	Optimal	0.52	0.82
	Quasi Optimal	0.65	0.88

Paths	β
Edu. Curr. → Job Curr.	0.27
Edu. Curr.→ Job Success	-0.01
Job Curr. → Job Success	0.69

GoF=0.34

b) Non Metric PLS

LVs	MVs	Weights	Loadings	C.I
Edu. Curr.				
	Age	-0.33	-0.80	$[-0.85, -0.70]$
	Final Grade	0.36	0.94	$[0.89, 0.96]$
	Average Grade	0.39	0.95	$[0.90, 0.96]$
	ISEE	0.10	0.28	$[-0.10, 0.46]$
Job Curr.				
	N_CN	-0.04	-0.15	$[-0.27, -0.03]$
	GG Work	0.28	0.83	$[0.79, 0.86]$
	GG ISCO12	0.47	0.75	$[0.72, 0.79]$
	GG al243	0.30	0.85	$[0.80, 0.88]$
	GG CTI	0.26	0.56	$[0.45, 0.64]$
Job Success				
	Optimal	0.53	0.82	$[0.77, 0.86]$
	Quasi Optimal	0.64	0.88	$[0.85, 0.92]$

Paths	β	C.I
Edu. Curr. → Job Curr.	0.26	$[0.19, 0.34]$
Edu. Curr.→ Job Success	-0.02	$[-0.09, 0.06]$
Job Curr. → Job Success	0.69	$[0.65, 0.73]$

GoF=0.39

c) Non Metric PLS where all the Mvs are analysed at an ordinal scaling level and the MVs "Optimal" and "Quasi-Optimal are analysed at a nominal scaling level

LVs	MVs	Weights	Loadings	C.I
Edu. Curr.				
	Age	-0.10	-0.80	$[-0.85, -0.70]$
	Final Grade	0.11	0.93	$[0.88, 0.96]$
	Average Grade	0.12	0.94	$[0.89, 0.96]$
	ISEE	0.04	0.29	$[-0.10, 0.47]$
Job Curr.				
	N_CN	-0.06	-0.11	$[-0.28, 0.14]$
	GG Work	0.23	0.78	$[0.70, 0.82]$
	GG ISCO12	0.42	0.74	$[0.68, 0.78]$
	GG al243	0.26	0.78	$[0.70, 0.82]$
	GG CTI	0.25	0.57	$[0.45, 0.66]$
Job Success				
	Optimal	0.27	0.82	$[0.76, 0.86]$
	Quasi Optimal	0.33	0.89	$[0.85, 0.92]$

Paths	β	C.I
Edu. Curr. → Job Curr.	0.30	$[0.211, 0.37]$
Edu. Curr.→ Job Success	-0.07	$[-1.08, 0.02]$
Job Curr. → Job Success	0.75	$[0.72, 0.79]$

GoF=0.43

Figure 4.2: Comparison among the analyses performed adopting different scaling levels.

In case a) the weights of all the binary variables representing the *isee* variable are small because there is a strong relation between the categorical variables and the Educational Curriculum. The non-metric analysis makes it clear that MV *isee* is not important in the construction of the Educational Curriculum.
The interpretation of the inner relations does not change substantially among the results of these three analysis: the impact on Job Success of Job Curriculum ($\hat{\beta} = 0.75$) is positive and higher than the one of Educational Curriculum ($\hat{\beta} = -0.07$); but the global model fit (GoF) improves from 0.34 to 0.43 from the standard PLS to the Non-Metric PLS analyses.

In what follows we discuss the results coming from the Non-Metric PLS analysis (case c). The values of the validation of the outer model with corresponding 95% confidence intervals built by means of 1000 bootstrap samples are reported in Tab.4.2. In the block of Educational Curriculum we have for all the MVs high loadings with the exception of the manifest variable *isee* (0.29). Thus we could consider removing this variable from the model. Moreover the empirical validation of the model shows that this value is not significant. The block of Educational Curriculum is positively affected by all the its MVs with the exception of Age that is negatively correlated. This is a trivial fact because the older the alumni the less is the study

Table 4.2: Main results of the measurement (outer) model: the weights and loadings (λ) are shown. For the loadings the corresponding 95% confidence intervals built by means of 1000 bootstrap samples are reported.

LVs	MVs	Weights	λ	Std.Error	perc.025	perc.975
Edu. Curr.						
	Age	-0.10	-0.80	0.12	-0.85	-0.70
	Final Grade	0.11	0.93	0.13	0.88	0.96
	Average Grade	0.12	0.94	0.13	0.89	0.96
	ISEE	0.04	0.29	0.16	-0.10	0.47
Job Curr.						
	N_CN	-0.06	-0.11	0.12	-0.28	0.14
	GG Work	0.23	0.78	0.03	0.70	0.82
	GG ISCO12	0.42	0.74	0.03	0.68	0.78
	GG al243	0.26	0.78	0.03	0.70	0.82
	GG_CTI	0.25	0.57	0.05	0.45	0.66
Job Success						
	Optimal	0.27	0.82	0.03	0.76	0.86
	Quasi Optimal	0.33	0.89	0.02	0.85	0.92

success.

In the block of Job Curriculum we find a similar situation to the one found in the previous block. Only for *n_cn* we have a very small values of the loadings (0.11). Also in this case the bootstrap procedure indicates a non significant value. We have checked that removing *isee* and *n_cn* from the model, the GoF increases from 0.42 to 0.46. All the MVs in this block are positively correlated with the own LV with the exception of *n_cn* that is negative. In the block of Job Success we have high loadings for all the MVs.

The *optimal* and *quasi optimal* indicators are discriminant to the construction of the Job Success block, see Tab. 4.2. In fact, the weights of the MVs quantified at nominal scaling, which reflect the variability of the corresponding LV explained by the categories of the MVs, have high values particularly in the case of the Quasi-Optimal indicator. Results of the structural model with corresponding 95% confidence intervals built by means of 1000 bootstrap samples are reported in Tab. 4.3. The path coefficient from Educational Curriculum to Job Curriculum is moderately small (0.30) indicating a feeble influence of educational curriculum on job experiences. In

Table 4.3: Results of the structural model with corresponding 95% confidence intervals built by means of 1000 bootstrap samples. The R^2 and the path coefficients (β) are shown.

Paths	R^2	β	Std.Error	perc.025	perc.975
Edu. Curr. $\rightarrow$ Job Curr.	0.09	0.30	0.04	0.21	0.37
Edu. Curr.$\rightarrow$ Job Success	0.54	-0.07	0.03	-1.08	0.02
Job Curr. $\rightarrow$ Job Success		0.75	0.02	0.72	0.79

the case of the regression of Job Success in respect of Educational Curriculum and Job Curriculum we see that, while the Job Curriculum influences the Job Success very much (0.75), the Educational Curriculum has a small coupling with Job Success and a negative sign (-0.07). The bootstrap intervals for the path coefficient of Educational Curriculum to Job Success contain the value zero, so this coefficient is not a significant 5% confidence level, see Tab. 4.3. It is also interesting to note that the indirect effect of Educational Curriculum to Job Success i. e. the path: Educational Curriculum $-$Job Curriculum$-$Job Success, gives a positive contribution of 0.17 which is not negligible.

The results of this regression suggest to analyse a simpler inner model in which Educational Curriculum is linked with Job Curriculum and Job Curriculum with Job Success.

In Tab. 4.3 we also reported the R^2 values of the endogenous latent variables for each regression in the structural model. We have $R^2 = 0.09$ for the regression where the endogenous variable is Job Curriculum and a higher value $R^2 = 0.54$ in the case of the endogenous variable Job Success. The value 0.09 for the first R^2 is rather low but it is confirmed by the bootstrap procedure as well as the corresponding path coefficient. In order to evaluate these values, it should be taken into account that high values of R^2 are not expected because our endogenous manifest variables (Optimal and Quasi-Optimal) in the block of Job Success are binary and they are analysed together with nominal, ordinal and numerical variables. The values of the main goodness indices obtained from our model are reported in Tab. 4.4. The average redundancy for Job Success indicates that Edu. and Job Curricula predict 40% of the variability Job Success indicators whereas the average redundancy for Job Curriculum indicates that Edu. Curriculum predicts lower value of 3% of the vari-

Table 4.4: Results of the main indices for the evaluation of the model. Average Communalities (Av.C), Average Redundancy (Av.R), AVE and GoF are shown.

LVs	Type	Av.C	Av.R	AVE
Edu. Curr.	Exogenous	0.62		0.62
Job Curr.	Endogenous	0.42	0.03	0.42
Job Success	Endogenous	0.73	0.40	0.73
GoF	0.42			

ability of Job Curriculum. The AVE index shows good values for all our constructs except for Job Curriculum. Finally, we obtained that the whole prediction power of the model is GoF=0.43.

In Fig. 4.3 we report, for all the variables, the plots of the raw values versus the scaling values obtained at the end of the convergence of the iterative procedure. These plots show that all the non-metric manifest variables are properly quantified using monotone transformations of the quantitative MVs.

Following the indications coming from the previous discussion, we analysed the simpler inner model in which Educational Curriculum is linked with Job Curriculum and Job Curriculum with Job Success. This path follows the natural temporal sequence from Educational Curriculum to Job Curriculum and then to Job Success of a standard student. We analysed this new model where the concept of Job Success is represented only from the Quasi-Optimal indicator and in which we have not considered the manifest variables *isee* and *n_cn*. In this case we have relaxed the requirement by the Job Success by considering only the more flexible Quasi-Optimal indicator. The results of this model are substantially unchanged, in fact, all the manifest variables and all the paths between the blocks are confirmed by the bootstrap procedure and the GoF index increases to 0.48. Fig. 4.4 reports the main results of this model.

In the case of a block q with only one MV and this is binary, the weight and the loading are 1 and the LV ξ_q has also a binary structure. We have studied this last model also modifying the inner model structure by adopting the logistic regression to estimate the path coefficient between Job. Curr. and Job Success. In Sect. 2.2.1 we discussed the structure of the inner model. In the PLS-PM and in the Non-Metric PLS, the relationship among the LVs are modeled through a multiple

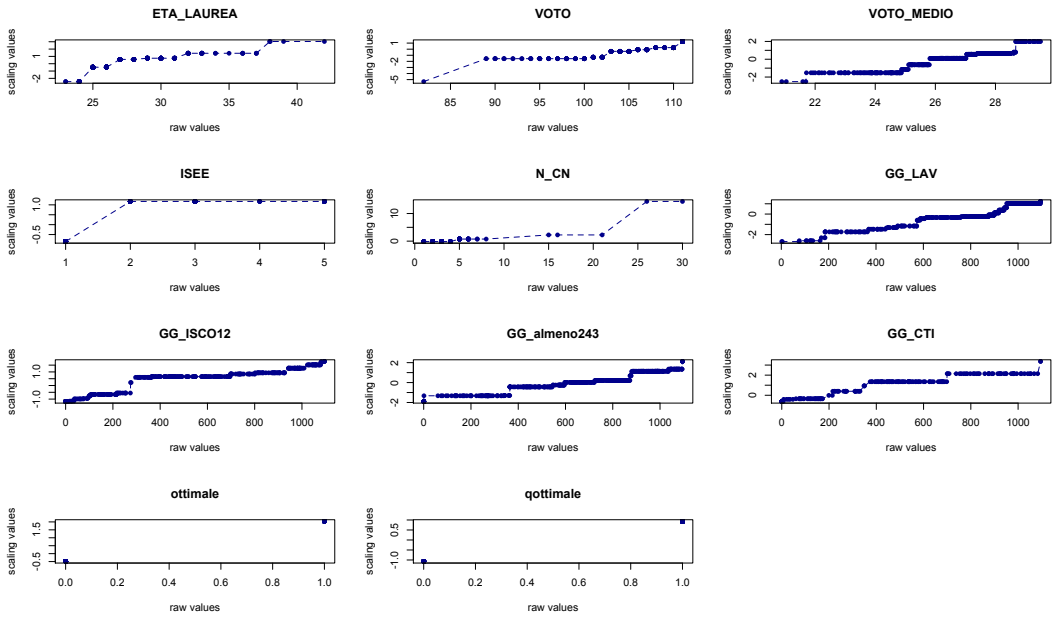


Figure 4.3: Values of the original variables plotted versus the corresponding optimal scaling values.

regression model, as shown in Eq. 2.1. The method of calculation of the path coefficients impacts on the type of available endogenous LVs. In this thesis we have implemented the possibility of modeling binary endogenous latent variables through logistic regression model:

$$\text{logit}(\boldsymbol{\xi}_j = 1|\boldsymbol{\xi}_{m\to j}) = \log\left[\frac{\Pr(\boldsymbol{\xi}_j = 1|\boldsymbol{\xi}_{m\to j})}{\Pr(\boldsymbol{\xi}_j = 0|\boldsymbol{\xi}_{m\to j})}\right] = \beta_{j0} + \sum_{m=1}^{M_j} \beta_{jm}\boldsymbol{\xi}_{m\to j} \tag{4.1}$$

So in this case the inner model can be written as:

$$\text{Job Curr.} = \beta_0 + \beta_1 \text{ Edu. Curr.} + \nu \tag{4.2}$$

$$\text{logit}[\Pr(\text{Job Success} = 1|\text{ Job Curr.})] = \beta_0 + \beta_1 \text{ Job Curr.} \tag{4.3}$$

where

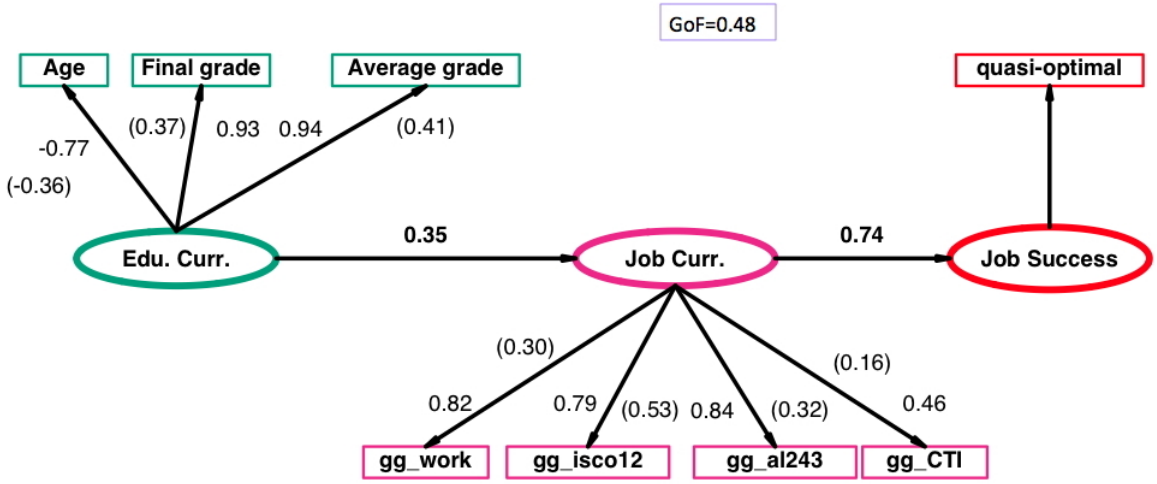


Figure 4.4: Path Diagram depicting our model. In this graph, for the outer model the weights in brackets and the standard loadings are reported. For the inner model the path coefficients and GoF are shown.

Figure 4.5: Path Diagram depicting our model. For the inner model the values of path coefficients obtained from the model is shown. In this case the path coefficient between Job Curr. and Job Success is calculated using a logistic regression.

$$\begin{aligned} \text{logit}(\Pr(\text{Job Success} = 1 |\ \text{Job Curr.})) &= \log\ [\text{Odds}] = \\ &= \log \left[\frac{\Pr(\text{Job Success}{=}1 | \text{Job Curr.})}{\Pr(\text{Job Success}{=}0 | \text{Job Curr.})} \right] \end{aligned} \tag{4.4}$$

It is possible to read this path coefficient in terms of odds ratio. Increasing of one-unit the Job Curr. score, the Job Success odds increases about 30 time, see Fig. 4.5.

We use the ROC curve to validate the model with the logistic regression, as described in Sect. 3.2.3. Fig. 4.6 shows the ROC curve for our model. The value of the Area Under the Curve is high and equal to 0.94 indicating a good classification of the data.

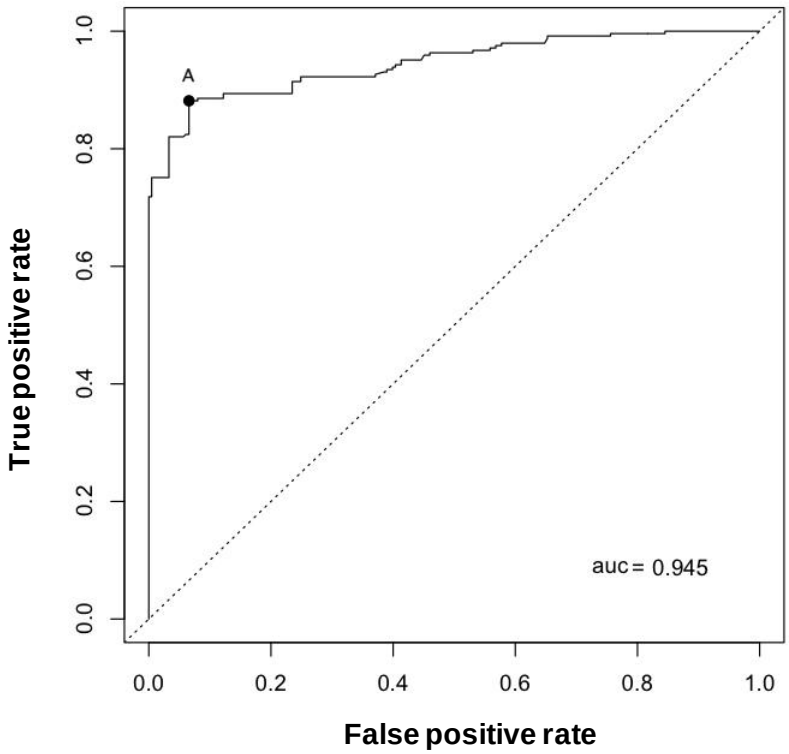


Figure 4.6: Validation model: the ROC curve. The value of the Area Under the Curve (AUC) is reported.

4.1.2 Concluding remarks

We have presented one of the first statistical analyses based on the data of the UNI.CO archive performed by Non-Metric PLSPM methods which have demonstrated a great adaptability to handle a large dataset with numerical, nominal and ordinal variables.

The high values of the measurement model have confirmed that the Optimal and Quasi Optimal indicators are discriminant to the construction of the Job Success block. We have seen that two variables (*isee* and *c_cn*) can be removed without reducing the capacity of the model to explain the variance and also the structural inner model can be reduced to a model with a simpler structure where the path

among the LVs becomes Educational Curriculum, Job Curriculum and Job Success. In the reduced model where the Job Success LV is represented only from *quasi-optimal* indicator, the logistic regression performed on the Job Success LV yielded a ROC curve with high AUC value indicating the good classification power of the *quasi-optimal* indicator.

4.2 Modeling the job career

In this Section, we propose two models in order to study the evolution of the job contract type and the job professional qualification in the three years after the master's degree.
We consider the comparison between the features of the following pairs of contracts:

- the first and the last contract (I and Z)
- the two most important contracts in terms of maximum actual duration (K1 and K2),

as discussed in Sect. 1.2.2. These comparisons allows us to monitor the evolution of the contract type and of the job professional qualification in order to classify Sapienza alumni in terms of improvement (or not) of their job status. With this in mind, in the case of the evolution of the job contract type we have built a binary variable (*contract_evol.*) that is equal to one if the alumni improve their job with the same contract type or an improved one according to the rules in Fig. 1.3 and zero otherwise. According to the discussion of Sect. 1.2.2, we consider as an evolution of the job contract type the following six cases:

1. the graduate changes from a non standard to a standard contract, i. e. from an atypical or a mixed cause or internship job contract to a standard job contract (case A1);
2. the graduate changes from a partially standard to a standard contract (case A2);

3. the graduate changes from a non standard to a partially standard contract, i. e. from an atypical or a mixed cause or internship contract to a partially standard job contract (case A3);

4. the graduate changes from a standard contract to a new standard contract (case B1);

5. the graduate changes from a partially standard contract to a new partially standard contract (case B2);

6. the graduate changes from a non standard contract to a new non standard contract (case B3).

Similarly, according to Fig. 1.5 in Sect. 1.2.2, the evolution of the job professional qualification is defined as a binary variable (*isco_evol.*) that is equal to one if the alumni improve their job with the same professional qualification or an improved one and zero otherwise.
Assuming the ISCO classification as discussed in Sect. 1.2.2, we consider as an evolution of the job professional qualification the following six cases:

1. the graduate's profession changes from a medium skilled to a highly qualified profession;

2. the graduate's profession changes from a low skilled to a highly qualified profession;

3. the graduate's profession changes from a low skilled to a medium skilled profession;

4. the graduate's profession changes from a highly qualified profession to a new one with the same qualification;

5. the graduate's profession changes from a medium skilled profession to a new one with the same qualification;

6. the graduate's profession changes from a low skilled profession to a new one with the same qualification.

The main purpose we addressed is to model in the PLS-PM framework the contract type and job professional quality evolution as latent variables - in the following called Contractual Evolution and Professional Evolution respectively - to study these variables in relation to educational curriculum, age, initial experience and tendency to evolve. For this study we selected from the UNI.CO dataset, independently from the disciplinary group, two sets of data:

- alumni who at the time of the master's degree had obtained an active contract (997 units), called WORKERS;
- alumni who at the time of the master's degree had not an yet obtained an active contract (4605 units), called NON-WORKERS.

This distinction allows us to design two different models which follow the time evolution scheme of the two alumni's types. Moreover this decision stems from the high degree of heterogeneity revealed by our preliminary analysis on the complete dataset. This great heterogeneity probably reflects the loss of clear social and cultural reference points that is typical of the times of crisis.
These two datasets contain variables which are non metric. Below is reported the list of the variables, presented in blocks of indicators with their corresponding description:

Age LV associated with only one MV that represents the alumni's age at the master's degree.

* **Age:** Age at the master's degree. This variable is ordinal with the following modalities:
  - $1 \rightarrow (20 - 21]$ years;
  - $2 \rightarrow (22 - 23]$ years;
  - $3 \rightarrow (24 - 25]$ years;
  - $4 \rightarrow (26 - 27]$ years;
  - $5 \rightarrow (28 - 29]$ years;
  - $6 \rightarrow (30 - 35]$ years;
  - $4 \rightarrow [36+)$ years.

Educational Curriculum LV associated with the characteristics of the university educational career. The MVs related with this LV are:

* **Final Grade:** Final university grade. This variable is ordinal with the following modalities:
  - $1 \rightarrow (66 - 99]$ grade;
  - $2 \rightarrow (100 - 104]$ grade;
  - $3 \rightarrow (105 - 109)$ grade;
  - $4 \rightarrow (110{-}111)$ grade (where the grade 111 indicates 110 cum laude).
* **Average Grade:** Average graduation grade. This variable is ordinal with the following modalities:
  - $1 \rightarrow [18 - 20]$ grade;
  - $2 \rightarrow (20 - 22]$ grade;
  - $3 \rightarrow (22 - 24]$ grade;
  - $4 \rightarrow (24 - 26]$ grade;
  - $5 \rightarrow (26 - 28]$ grade;
  - $6 \rightarrow (28 - 29]$ grade;
  - $7 \rightarrow (29 - 30]$ grade.
* **Fcourse:** Binary variable that represents whether a student is on time (or not). This is nominal variable with the following modalities:
  - $0 \rightarrow$ The student is on time;
  - $1 \rightarrow$ Otherwise.

Initial Experience LV associated with the characteristics of the initial job experience in terms of contract type and professional qualification. The MVs related with this LV are:

* **isco_first:** The professional qualification of the first contract. This variable is ordinal with the following modalities:
  - $9 \rightarrow$ ISCO 9: Unqualified professions;
  - $8 \rightarrow$ ISCO 8: Workers at industries machineries and assembly lines;

- 7 → ISCO 7: Craftsmen and skilled workers;
- 6 → ISCO 6: Staff specialized in agriculture, forestry and fishing;
- 5 → ISCO 5: Commercial activities;
- 4 → ISCO 4: Office worker;
- 3 → ISCO 3 : Technical and intermediate occupations;
- 2 → ISCO 2: Intellectual and scientific professions;
- 1 → ISCO 1: Managers.

* **Contract_first:** The contract type of the first signed contract . This variable is ordinal with the following modalities:
  - 1 → Professional experience;
  - 2 → Mixed cause contracts (e.g. contracts that provide a training as a component of Apprenticeship, insert, CFL in public administration);
  - 3 → Atypical (e.g. fixed-term employment, employment with project contract) ;
  - 4 → Partially Standard (part-time permanent work);
  - 5 → Standard (full time permanent work).

Tendency to Evolve LV associated with the capability showed by a subject to improve the job career changing works during the three years after the master's degree. The MVs related with this LV are:

* **isco_best:** Best professional qualifications among the four contracts considered. This variable is ordinal with the following modalities:
  - 9 → ISCO 9: Unqualified professions;
  - 8 → ISCO 8: Workers at industries machineries and assembly lines;
  - 7 → ISCO 7: Craftsmen and skilled workers;
  - 6 → ISCO 6: Staff specialized in agriculture, forestry and fishing;
  - 5 → ISCO 5: Commercial activities;
  - 4 → ISCO 4: Office worker;
  - 3 → ISCO 3 : Technical and intermediate occupations;

  - 2 → ISCO 2: Intellectual and scientific professions;
  - 1 → ISCO 1: Managers.

* **Contract_best:** Best job contract type among the four contracts considered. This variable is ordinal with the following modalities:
  - 1 → Professional experience;
  - 2 → Mixed cause contracts (e.g. contracts that provide a training as a component of Apprenticeship, insert, CFL in public administration);
  - 3 → Atypical (e.g. fixed-term employment, employment with project contract) ;
  - 4 → Partially Standard (part-time permanent work);
  - 5 → Standard (full time permanent work).

Contractual Evolution LV associated with only one MV that represents whether the alumni improve their job with the same contract type or with an improved one.

* **contract_evol :** Dummy variable that represents the presence of a contract type evolution. This is a nominal variable with the following modalities:
  - 1 → if the graduate improve their job with the same contract type or an improved one;
  - 0 → otherwise.

Professional Evolution LV associated with only one MV that represents whether the alumni improve their job with the same professional qualification or with an improved one.

* **isco_evol:** Dummy variable that represents the presence of a professional qualification contract evolution. This is a nominal variable with the following modalities:
  - 1 → if the alumni improve their job with the same professional qualification or an improved one;
  - 0 → otherwise.

4.2.1 Alumni without an active contract at the master's degree

In the following discussion we analyse the dataset of the alumni who at the time of the master's degree did not have an active contract (4605 units). First of all, we assume a number of a priori hypotheses that represent the expected relationships according to the existing literature and to the temporal sequence of the events.

H1 : Educational Curriculum is associated with Age;

H2 : Educational Curriculum is positively associated with Initial Experience;

H3 : Educational Curriculum is positively associated with Tendency to Evolve;

H4 : Educational Curriculum is positively associated with Contractual Evolution;

H5 : Educational Curriculum is positively associated with Professional Evolution;

H6 : Age is associated with the Initial Experience;

H7 : Age is associated with Tendency to Evolve;

H8 : Age is associated with Contractual Evolution;

H9 : Age is associated with Professional Evolution;

H10 : Initial Experience is positively associated with Tendency to Evolve;

H11 : Initial Experience is positively associated with Professional Evolution;

H12 : Initial Experience is associated with Contractual Evolution;

H13 : Tendency to Evolve is positively associated with Contractual Evolution;

H14 : Tendency to Evolve is positively associated with Professional Evolution.

In this study a Non Metric Partial Least Squares Path modeling using the *plspm* R-package was used to assess the measurement model and the structural model. The validation of model paths, loadings and weights is carried out by a bootstrap procedure by $n = 1000$ re-samples. The path diagram representing the validated

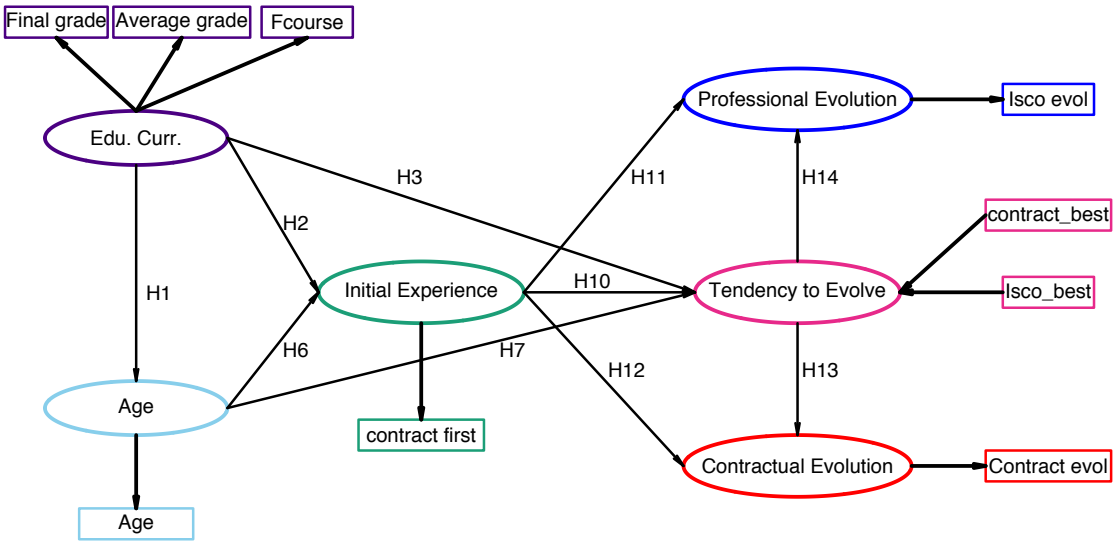


Figure 4.7: Path Diagram depicting our model for NON-WORKERS.

model is depicted in Fig. 4.7. We performed a Non-Metric PLSPM analysis on the model described previously by using the option centroid for the inner weight estimation (this choice only considers the sign of the correlations between a LV and its adjacent LVs). To update the external weights we used for Educational Curriculum the Mode A option because it is assumed to be a common factor that reflects itself in the MVs; Mode B is used for Tendency to Evolve because we considered that each manifest variable of MVs accounts for different dimensions of the underlying concept. As shown in Fig. 4.7 our model relates:

- Contractual Evolution with Initial Experience and with Tendency to Evolve;
- Professional Evolution with Initial Experience and Tendency to Evolve;
- Tendency to Evolve with Educational Curriculum, Initial Experience and Age;
- Initial Experience with Educational Curriculum and Age;
- Age with Educational Curriculum.

Note that the weight and the loading of *isco_first* are not significant and therefore this MV has been removed from the validated model. Hence, Initial Experience

block is composed only by the *contract_first* MV.

4.2.2 Discussion of the results

In Fig. 4.8 the optimal transformations of ordinal and nominal MVs are reported. These plots show that all the variables are properly quantified by monotone transformations of the quantitative MVs. The more the dashed lines deviate from linearity, the more the raw variables deviate from the assumption of linearity. So, for example, the transformation of *isco_best* significantly improves the ability of the model to identify relationships with other variables, while the raw variable *final_grade* is already nicely linear. Note that the plot of the quantification for the MV *Fcourse* shows that the sign of this variable has been changed, in other words we are considering the variable *in course*.

The results of the outer model with corresponding 95% confidence intervals built

Table 4.5: Main results of the measurement model for NON-WORKERS with corresponding 95% confidence intervals built by means of 1000 bootstrap samples. The weights and the loadings are shown for the blocks with more than one MV. The others: Age, Initial Experience, Contractual Evolution and Professional Evolution have therefore weights and loadings equal to one.

LVs	**MVs**	**Weigths**	**Std.Error**	**C.I.**
Edu. Curr.				
	Fcourse	0.44	0.02	[0.41, 0.47]
	final_grade	0.40	0.01	[0.38, 0.41]
	average_grade	0.37	0.01	[0.36, 0.39]
Tendency to Evolve				
	best_contract	0.97	0.00	[0.97, 0.98]
	isco_best	0.18	0.01	[0.15, 0.21]
LVs	**MVs**	**Loadings**	**Std.Error**	**C.I.**
Edu. Curr.				
	Fcourse	0.72	0.01	[0.70, 0.75]
	final_grade	0.89	0.01	[0.87,0.90]
	average_grade	0.88	0.01	[0.86, 0.89]
Tendency to Evolve				
	best_contract	0.98	0.00	[0.98, 0.99]
	isco_best	0.23	0.02	[0.20, 0.27]

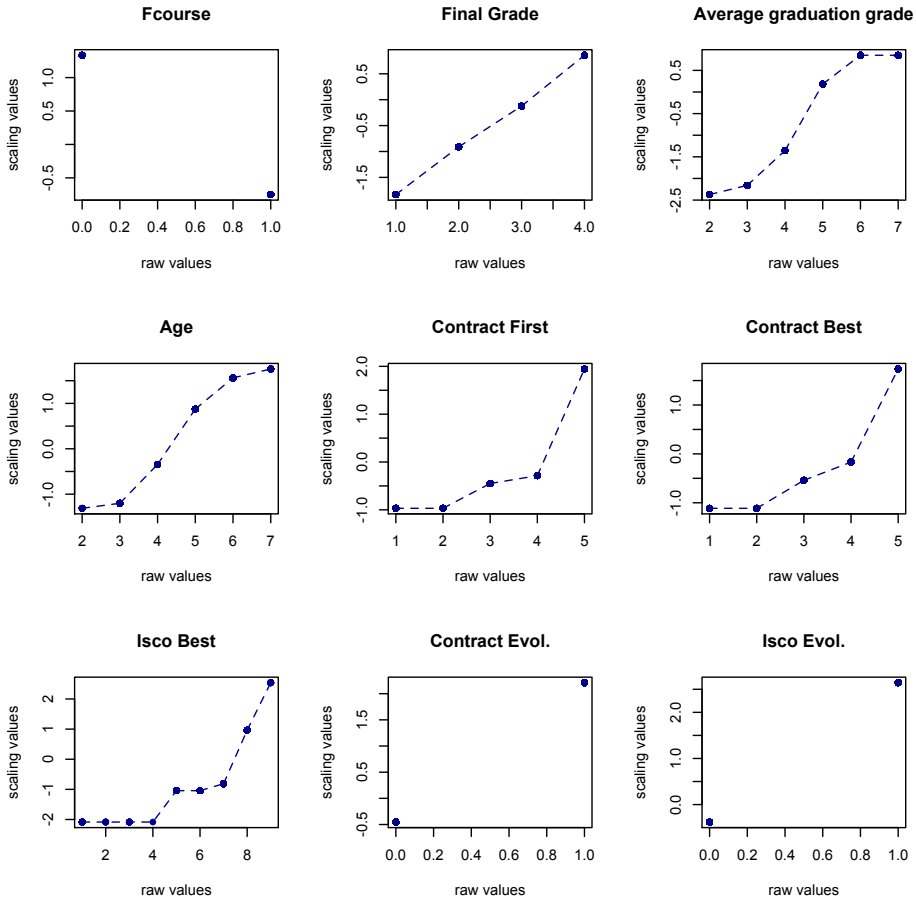


Figure 4.8: Values of the original variables plotted versus the corresponding optimal scaling values for NON-WORKERS model.

by means of 1000 bootstrap samples are reported in Tab. (4.5)[2]. The block Educational Curriculum, where Mode A has been adopted, is positively affected by its MVs which have all high loadings. Hence an increasing of the values of the *Fcourse*, *final_grade* and *average_grade* MVs, contributes positively to the Educational Curriculum.

In the block Tendency to Evolve, where Mode B has been adopted, *contract_best*

[2] In the case of the Age, Initial Experience, Contractual Evolution and Professional Evolution blocks which are represented by only one manifest variable, the weights and the loadings are equal to 1 and they are not reported in this table.

and *isco_best* have high and mildly high value of the weights respectively. This is reasonable because nowadays it is easier and more desirable to change work in order to improve the contract type rather than getting an advancement of the job professional qualification with a short-term contract.
The structural model can be written as:

$$\begin{aligned} E[\text{Age}|X] &= \beta_0 + \beta_1 \text{Edu. Curr.} \\ E[\text{Initial Exper.}|\boldsymbol{X}] &= \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Edu. Curr} \\ E[\text{Tend. to Evolve}|\boldsymbol{X}] &= \beta_0 + \beta_1 \text{Edu. Curr.} + \beta_2 \text{Age} + \beta_3 \text{Initial Exper.} \end{aligned}$$

$$E[(\text{Contrac. Evol.} = 1|\boldsymbol{X})] = \frac{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\}}{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\} + 1}$$

$$E[(\text{Prof. Evol.} = 1|\boldsymbol{X})] = \frac{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\}}{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\} + 1}$$

Results of the structural model using the logistic regression together with the corresponding 95% confidence intervals are reported in Tab. 4.6 [3]. We have analysed the inner model structure using the enhanced NM-PLSPM procedure where we have modified the structural model by adopting the logistic regression for the latent binary variables Contractual Evolution and Professional Evolution. In what follows we will discuss the picture of the relationships among the five LVs as it appears from the value of the path coefficients.
Age has a negative impact on Educational Curriculum ($\hat{\beta} = -0.53$). The higher the Educational Curriculum success, smaller is the age at the master's degree, as it could be expected.
Initial Experience impacts on Educational Curriculum ($\hat{\beta} = 0.07$) and, with negative sign, on Age ($\hat{\beta} = -0.15$). We found a similar situation for the relationships between Tendency to Evolve and Educational Curriculum ($\hat{\beta} = 0.02$), and Tendency to Evolve and Age ($\hat{\beta} = -0.05$). It is reasonable that the age at the master's degree mildly influences Initial Experience and also Tendency to Evolve. The negative sign

[3] In this table, we reported the 95% confidence intervals obtained from the regressions. We have checked the agreement with the 95% confidence intervals built by means of 1000 bootstrap re-samples.

Table 4.6: Results of the structural model for NON-WORKERS with corresponding 95% confidence intervals. The R^2 and the path coefficients ($\hat{\beta}$) are shown. The pseudo Nagelkerke R^2 is indicated with $*$ in the R^2 column, and O.R. indicates the odds ratio values.

Paths	$\boldsymbol{R^2}$	$\hat{\boldsymbol{\beta}}$	**Std.Err.**	**C.I.**	**O.R.**
Edu. Curr. →Age	0.28	-0.53	0.01	[-0.55, -0.51]	-
Edu. Curr. →Initial Exper.	0.03	0.07	0.02	[0.03, 0.10]	-
Age → Initial Exper.		-0.15	0.02	[-0.18, -0.12]	-
Edu. Curr. →Tend. to Evolve	0.78	0.02	0.01	[0.01, 0.04]	-
Age → Tend. to Evolve		-0.05	0.01	[-0.06, -0.03]	-
Initial Exper. → Tend. to Evolve		0.87	0.01	[0.85, 0.88]	-
Tend. to Evolve → Contrac. Evol.	0.77*	2.79	0.18	[2.47,3.17]	16.28
Initial Exper. → Contrac. Evol.		0.55	0.07	[0.42, 0.69]	1.73
Tend. to Evolve → Profess. Evol.	0.09*	1.01	0.07	[0.87, 1.15]	2.74
Initial Exper. → Profess. Evol.		-0.58	0.07	[-0.72, -0.44]	0.56

is also comprehensible because it is expected that increasing Age reduces the quality of Initial Experience and Tendency to Evolve. The path coefficients from Educational Curriculum to Initial Experience and to Tendency to Evolve are very small, ($\hat{\beta} = 0.07$) and ($\hat{\beta} = 0.02$) respectively, indicating a feeble influence of educational curriculum on Initial Experience and on Tendency to Evolve. Moreover, it is worth noting that Tendency to Evolve largely impacts on Initial Experience ($\hat{\beta} = 0.87$) (rather than on Age ($\hat{\beta} = -0.05$) and on Educational Curriculum ($\hat{\beta} = 0.02$)). Therefore alumni with a good initial job have the propensity to consider the first job as the starting point of a career for which the mobility is an important feature. The two latent predictors Initial Experience and Tendency to Evolve positively impact on the response of Contractual Evolution, $\hat{\beta} = 0.55$ and $\hat{\beta} = 2.79$ respectively. Holding Tendency to Evolve at a fixed value, for every one-unit increase in Initial Experience score, the odds of observing an evolution of the type of contract changes by a factor 1.73. On the other hand, holding Initial Experience at a fixed value, for every one-unit increase in Tendency to Evolve score, the odds of observing an evolution of the type of contract changes by a factor 16.28. Therefore alumni with a starting job of good quality and an own tendency to evolve have good chances to improve their contract status.

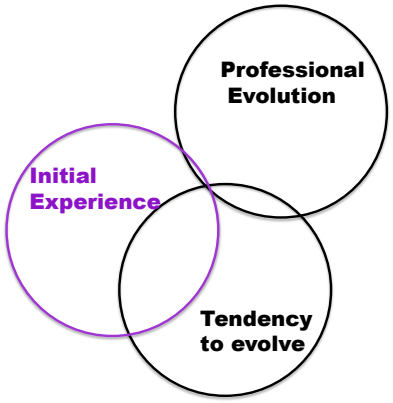


Figure 4.9: Diagram describing schematically the role of Initial Experience like suppressor variable in the Professional Evolution regression.

The two latent predictors Initial Experience ($\hat{\beta} = -0.58$) and Tendency to Evolve ($\hat{\beta} = 1.01$) impact in opposite sense on the response of Professional Evolution. It is rather clear that Professional Evolution should be positively influenced by Tendency to Evolve due to individual initiative to get a better job status. Taking into account the Professional Evolution block, holding Tendency to Evolve at a fixed value, for every one-unit increase in Initial Experience score, the odds of the evolution of the professional qualification of signed contracts for Sapienza alumni changes by a factor 0.56 showing that a good Initial Experience can inhibit the tendency to compete for a job of better quality. On the other hand, holding Initial Experience at a fixed value, for every one-unit increase in Tendency to Evolve, the odds of the possibility of an evolution of the professional qualification of signed contracts for Sapienza alumni changes about 2.74 times. This result highlights how crucial the capability of alumni to change their job in order to improve their career is.

Checking the coherence between the path coefficients and the corresponding correlation signs, we found that the correlation between Initial Experience and Professional Evolution is low ($r = 0.09$) while its path coefficient is negative and moderately high ($\hat{\beta} = -0.58$). Moreover the correlation between Initial Experience and Tendency to Evolve is very high ($r = 0.88$). In the linear multiple regression theory, this situation is described as a *negative classical suppression* between the exogenous variables, see Cohen and Cohen (1975) [18], Krus and Wilkinson (1986) [44] and Baron and

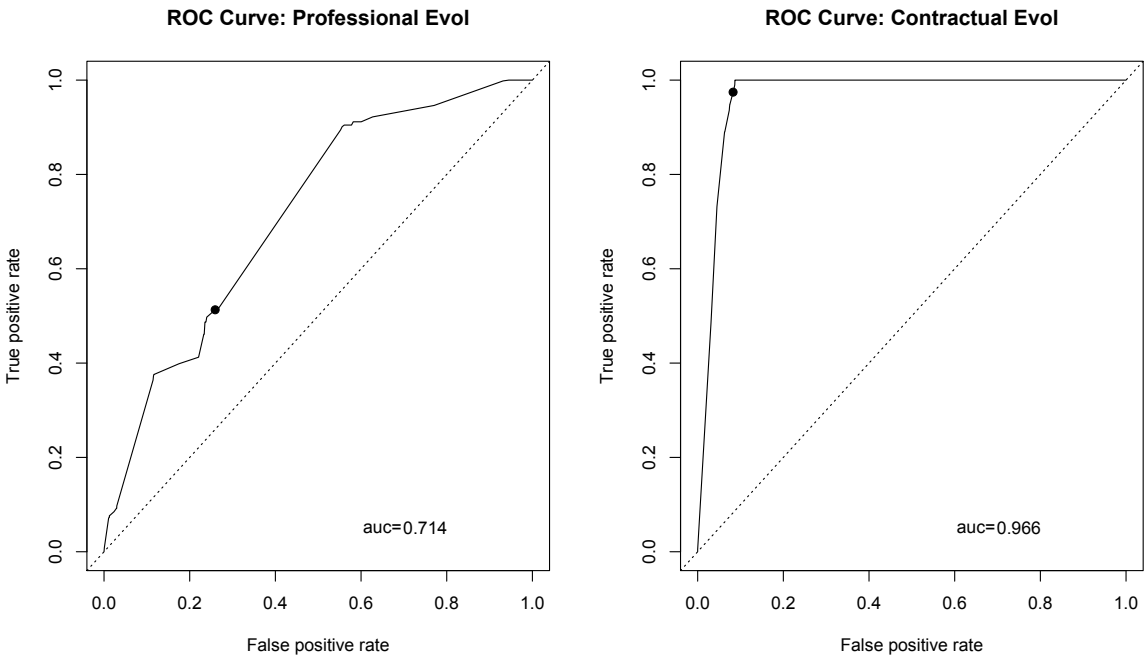


Figure 4.10: Validation model for NON-WORKERS: the ROC curve. The value of the area under the curve (AUC) is reported.

Kenny (1986) [6]. Here, Initial Experience increases the quality of the regression even though it has a small correlation with Professional Evolution. A sketch of the situation is depicted in Fig. 4.9 where the overlap in variance of the correlated variables is shown. The area enclosed by a circle represents the variance of its variable, that is the same for the three variables because they are standardized. The area of overlap among the circles represents the degree of linear relationship. What happens is that Initial Experience "suppresses" some of what would otherwise be variance in Tendency to Evolve. The general idea is that there is some kind of noise in Tendency to Evolve that is not particularly correlated with Professional Evolution, but it is strongly correlated with Initial Experience. By including Initial Experience in the regression, this noise is suppressed leaving Tendency to Evolve as an improved predictor of Professional Evolution and raising slightly the R^2 value of the regression. In other words, the capability to change job conditionally to the information of the initial job experience better predicts the professional evolution.

We use the ROC curve to validate the model with the logistic regression: Fig. 4.10 shows the ROC curves for our models. The values of the Area Under the Curve are high and equal to 0.71 for the logistic model where the endogenous LV is Professional Evolution and equal to 0.97 for Contractual Evolution.

4.2.3 Alumni with an active contract at the master's degree

In the following we analyse the dataset, that is complementary to the previous one, of alumni who at the time of the master's degree had an active contract (997 units). First of all, we assume a number of a priori hypotheses that represent the expected relationships according to the existing literature and to the temporal sequence of the events.

H1 : Initial Experience is positively associated with Educational Curriculum;

H2 : Initial Experience is associated with Age;

H3 : Initial Experience is positively associated with Tendency to Evolve;

H4 : Initial Experience is associated with Contractual Evolution;

H5 : Initial Experience is associated with Professional Evolution;

H6 : Educational Curriculum is associated with Age;

H7 : Educational Curriculum is positively associated with Tendency to Evolve;

H8 : Educational Curriculum is positively associated with Contractual Evolution;

H9 : Educational Curriculum is positively associated with Professional Evolution;

H10 : Tendency to Evolve is positively associated with Contractual Evolution;

H11 : Tendency to Evolve is positively associated with Professional Evolution.

In this study a Non Metric Partial Least Squares Path modeling using the *plspm* R-package was used to assess the measurement model and the structural model. The validation of model paths, loadings and weights is carried out by a bootstrap

procedure by $n = 1000$ re-samples. The path diagram representing the validated model is depicted in Fig. 4.11. We performed a Non-Metric PLSPM analysis on this model by using the option centroid for the inner weight estimation (this choice only considers the sign of the correlations between a LV and its adjacent LVs). To update the external weights we used for Educational Curriculum the Mode A option; Mode B is used for Tendency to Evolve.
For this model the temporal structure of the inner model has been designed to be suitable to the fact that this sample of data is related to alumni with professional experience obtained before getting the master's degree and already active at the time of the master's degree. Hence, Educational Curriculum is seen as a consequence of Initial Experience (contrary to the temporal sequence of the NON-WORKERS model). The second part of the model (the right part on the path diagram in Fig. 4.11) i.e. the scheme of relationship among Contractual Evolution, Tendency to Evolve and Professional Evolution is the same as before. As we will see, the final results of this model are very similar to the NON-WORKERS results. Anyway in the following, for sake of completeness, we repeat the complete scheme of analysis adopted for NON-WORKERS.
As shown in Fig. 4.11 our model relates:

- Contractual Evolution with Initial Experience and with Tendency to Evolve;
- Professional Evolution with Initial Experience and Tendency to Evolve;
- Tendency to Evolve with Educational Curriculum and Initial Experience;
- Educational Curriculum with Initial Experience.

Note that, also in this model, the weight and the loading of *isco_first* are not significant and therefore this MV has been removed from the validated model. Hence, Initial Experience block is composed only by the *contract_first* MV. The only difference from the model for NON-WORKERS is that for WORKERS the Age block has been removed because all the links starting from the Age block have been found to be not significant.

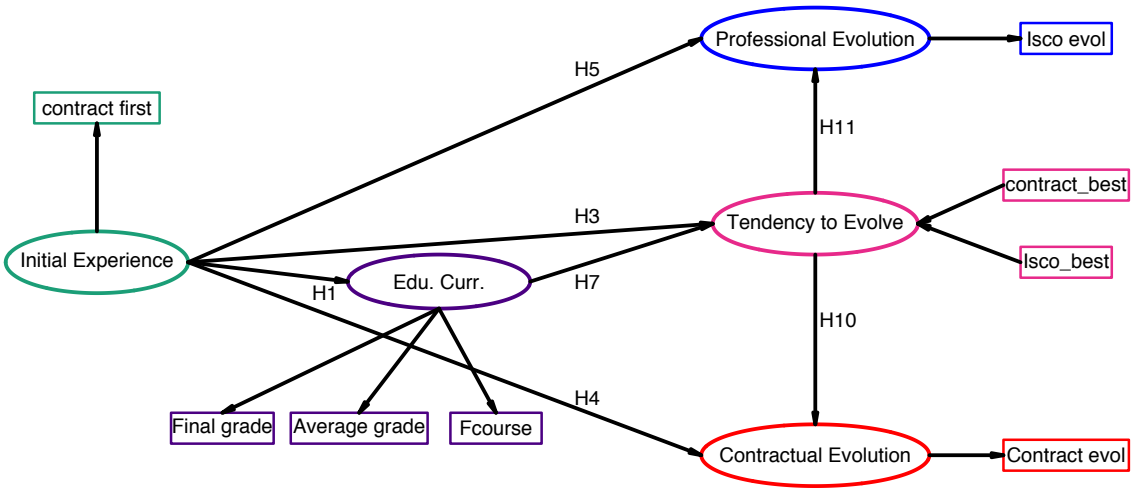


Figure 4.11: Path Diagram depicting our model for WORKERS.

4.2.4 Discussion of the results

In Fig. 4.12 the optimal transformations of ordinal and nominal MVs are reported. These plots show that the variables are properly quantified by monotone transformations of the quantitative MVs. The more the dashed lines deviate from linearity, the more the raw variables deviate from the assumption of linearity. So, for example, the transformation of *contract_first* significantly improves the ability of the model to identify relationships with other variables. Also in this case, the quantification for the MV *Fcourse* has reversed the sign of this variable that therefore takes the meaning of being *in course*.

The results of the outer model for the WORKERS with corresponding 95% confidence intervals built by means of 1000 bootstrap samples are reported in Tab. 4.7[4]. The block Educational Curriculum, where Mode A has been adopted, is positively affected by its MVs which have all high loadings. Hence an increasing of the values of the *Fcourse*, *final_grade* and *average_grade* MVs, contributes positively to the Educational Curriculum. In the block Tendency to Evolve, where Mode B has been

[4] In the case of the Age, Initial Experience, Contractual Evolution and Professional Evolution blocks which are represented by only one manifest variable, the weights and the loadings are equal to 1 and they are not reported in this table.

Table 4.7: Main results of the measurement model WORKERS with corresponding 95% confidence intervals built by means of 1000 bootstrap samples. The weights and the loadings are shown for the blocks with more than one MV. The others: Initial Experience, Contractual Evolution and Professional Evolution have therefore weights and loadings equal to one.

LVs	**MVs**	**Weights**	**Std.Error**	**C.I.**
Edu. Curr.				
	Fcourse	0.40	0.04	[0.31, 0.48]
	final grade	0.40	0.02	[0.37, 0.45]
	average grade	0.40	0.02	[0.36, 0.46]
Tendency to Evolve				
	best_contract	0.91	0.01	[0.89, 0.94]
	isco_best	0.29	0.03	[0.23, 0.34]
LVs	**MVs**	**Loadings**	**Std.Error**	**C.I.**
Edu. Curr.				
	Fcourse	0.69	0.04	[0.62, 0.76]
	final Grade	0.90	0.02	[0.86, 0.92]
	average grade	0.89	0.02	[0.84, 0.92]
Tendency to Evolve				
	best_contract	0.96	0.01	[0.94, 0.97]
	isco_best	0.43	0.04	[0.37, 0.49]

adopted, *contract_best* and *isco_best* have high and mildly high value of the weights respectively. These results are almost the same to those obtained for these blocks in the NON-WORKERS model. The only remarkable difference is that the weight of *isco_best* MV for WORKERS has a higher value (0.29) than NON-WORKERS, likely due to a longer working career of these alumni.

We have analysed the inner model structure using the enhanced NM-PLSPM procedure where we have modified the structural model by adopting the logistic regression for the latent binary variables Contractual Evolution and Professional Evolution.

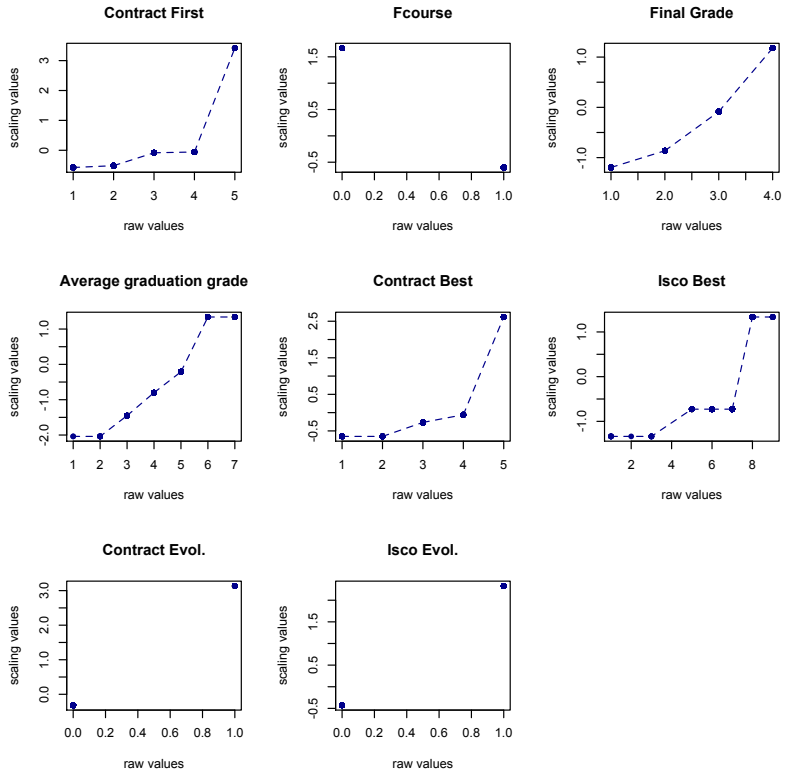


Figure 4.12: Values of the original variables plotted versus the corresponding optimal scaling values for WORKERS model.

The structural model can be written as:

$$
\begin{aligned}
E[\text{Edu. Curr.}|X] &= \beta_0 + \beta_1 \text{Initial Exper.} \\
E[\text{Tend. to Evolve}|\boldsymbol{X}] &= \beta_0 + \beta_1 \text{Initial Exper.} + \beta_2 \text{Edu. Curr.}
\end{aligned}
$$

$$
E[(\text{Contrac. Evol.} = 1|\boldsymbol{X})] = \frac{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\}}{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\}+1}
$$

$$
E[(\text{Prof. Evol.} = 1|\boldsymbol{X})] = \frac{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\}}{\exp\{\beta_0 + \beta_1 \text{ Initial Exper.} + \beta_2 \text{ Tend. to Evolve}\}+1}
$$

In what follows we will discuss the picture of the relationships among the four LVs

Table 4.8: Results of the structural model for NON-WORKERS with corresponding 95% confidence intervals. The R^2 and the path coefficients ($\hat{\beta}$) are shown. The pseudo Nagelkerke R^2 is indicated with $*$ in the R^2 column, and O.R. indicates the odds ratio values.

Paths	$\boldsymbol{R^2}$	$\hat{\boldsymbol{\beta}}$	**Std.Err**	**C.I.**	**O.R.**
Initial Exper. → Edu. Curr.	0.06	0.25	0.03	[0.19, 0.32]	-
Edu. Curr. → Tend. to Evolve	0.55	0.09	0.02	[0.05, 0.14]	-
Initial Exper. → Tend. to Evolve		0.72	0.03	[0.66, 0.76]	-
Initial Exper. → Contrac. Evol.	0.84*	0.61	0.13	[0.36, 0.88]	2.32
Tend. to Evolve → Contrac. Evol.		2.53	0.36	[1.94, 3.41]	12.55
Initial Exper. → Profess. Evol.	0.11*	-0.49	0.11	[-0.71, -0.28]	0.61
Tend. to Evolve → Profess. Evol.		0.85	0.11	[0.64, 1.064]	2.33

as it appears from the value of the path coefficients, see Tab. 4.8 [5].
Tendency to Evolve largely impacts on Initial Experience ($\hat{\beta} = 0.72$) rather than on Educational Curriculum ($\hat{\beta} = 0.09$). Also for WORKERS the influence of Educational Curriculum on Tendency to Evolve is a feeble influence, whereas alumni with a good initial job obtained before the end of the university studies have a natural propensity toward the mobility. The strength of Initial Experience over Tendency to Evolve is significantly lower than in the case of NON-WORKERS ($\hat{\beta} = 0.87$ against $\hat{\beta} = 0.72$) indicating a reduced motivation of these alumni to take the risk of changing.
The two latent predictors Initial Experience and Tendency to Evolve positively impact on the response of Contractual Evolution, $\hat{\beta} = 0.61$ and $\hat{\beta} = 2.53$ respectively. Holding Tendency to Evolve at a fixed value, for every one-unit increase in Initial Experience score, the odds of observing an evolution of the type of contract changes by a factor 2.32. On the other hand, holding Initial Experience at a fixed value, for every one-unit increase in Tendency to Evolve score, the odds of observing an evolution of the type of contract changes by a factor 12.55. Therefore, like NON-WORKERS, alumni with an active contract of good quality and an own tendency to evolve have good chances to improve their contract status.
The two latent predictors Initial Experience ($\hat{\beta} = -0.49$) and Tendency to Evolve

[5] In this table, we reported the 95% confidence intervals obtained from the regressions. We have checked the agreement with the 95% confidence intervals built by means of 1000 bootstrap re-samples.

($\hat{\beta} = 0.85$) impact in opposite sense on the response of Professional Evolution. It is rather clear that Tendency to Evolve influences positively the Professional Evolution because the individual initiative to get a better job status. The negative sign of the path coefficient of Initial Experience to Professional Evolution is reasonable because, having in mind that the initial job contract was already active at the moment of the master's degree, these alumni tend to keep the current contractual position even though in this way it is more difficult to improve their professional qualification. Probably they believe that it is better to try to improve their professional qualification without changing the contract type.
Taking into account the Professional Evolution block, holding Tendency to Evolve at a fixed value, for every one-unit increase in Initial Experience score, the odds of the evolution of the professional qualification of signed contracts for Sapienza alumni changes by a factor 0.61 On the other hand, holding Initial Experience fixed, for every one-unit increase in Initial Experience score, the odds of observing an evolution of the Professional Evolution changes by a factor 2.33.
Checking the coherence between the path coefficients and the corresponding correlation signs, we found that the correlation between Initial Experience and Professional Evolution is very low ($r = 0.04$) while its path coefficient is negative and moderately high ($\hat{\beta} = -0.49$). Moreover the correlation between Initial Experience and Tendency to Evolve is high ($r = 0.74$). Also in this case, we have a *negative classical suppression*, as shown in Fig. 4.9.
We use the ROC curve to validate the model with the logistic regression, Fig. 4.13 shows the Roc curves for our models. The values of the Area Under the Curve are high and equal to 0.81 for the logistic model where the endogenous LV is Professional Evolution and equal to 0.99 for Contractual Evolution.

4.2.5 Concluding remarks

We have shown that the great content of information included in the UNI.CO archive can be exploited to follow the career development of Sapienza alumni when adopting suitable models based on NM-PLSPM. The two models that have been proposed in order to study the job career of Sapienza alumni following the three years after

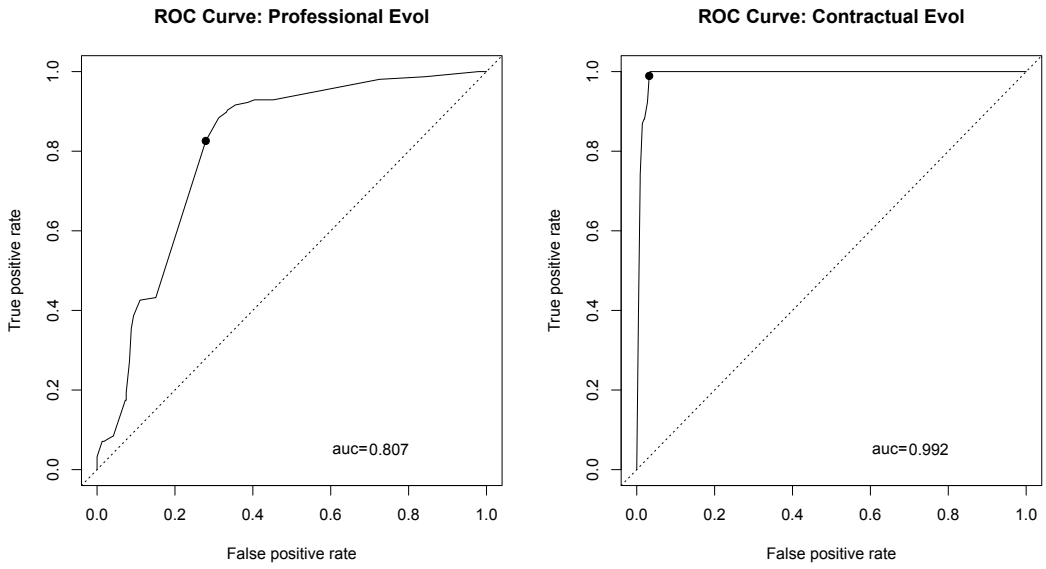


Figure 4.13: Validation model for WORKERS: The ROC curves. The value under the curve (AUC) is reported.

the master's degree, have showed the capability of NM-PLSPM to analyse a large amount of data. These models are essentially composed by four concepts: the university evaluation, the initial job experience, tendency to evolve and the career evolution. The career evolution is monitored by two constructs, one for the job professional evolution and the other for the job contractual evolution. We have found that in both the models there is a strong influence of the initial job experience and of tendency to evolve over the job career, while the university evaluation at the level of the master's degree is less important. The Age has a marginal effect in the model for NON-WORKERS while it disappears for WORKERS. The weight of *isco_best* MV for WORKERS has a higher value than NON-WORKERS, likely due to a longer working career of these alumni.

The improved version of the *plspm* R-package used to analyse a large dataset, as the UNI.CO archive, has shown a rapid convergence of the algorithms both in the case of Mode A and Mode B. The computational time necessary for the validation through the bootstrap procedure with 1000 re-samples was not heavy.

Conclusions

In this thesis the theme of the integration of Sapienza University of Rome alumni into the employee and para-subordinate labour market is presented analysing the data of the UNI.CO archive. This archive makes available administrative data which by their nature are objective and devoid of any emotional influence as it happens in archives based on sample surveys which are usually biased. Moreover, other sample surveys supply information about the status of alumni at some fixed point in time, while the data of the UNI.CO archive allow us to also study the evolution of alumni throughout the observation period and therefore to analyse the changes in job qualifications.

The study of the integration into the labour market of Sapienza alumni has been analysed by adopting the Non-Metric PLSPM method which allow us to handle a large dataset with numerical, nominal and ordinal variables. This technique recently proposed by Russolillo (2012) [56] and now available in the *plspm* R-package, is characterized by the possibility to analyse at the same time variables observed in different measurement scales, to investigate the non linearity and to work without distributional hypotheses. The quantifications of the original data obtained by Non-Metric PLSPM are properly carried out using monotone transformations showing the remarkable fact that it is possible to discard the hard assumption of linearity in favor of the milder assumption of monotonicity.

In this thesis, we have modified the inner structure of the Non-Metric PLSPM to implement the possibility of modeling binary endogenous latent variables through a logistic regression and then we have introduced the ROC curve for the model validation. This implementation is applicable in the case of blocks with only one MV and this is binary, therefore LVs have a binary structure and the weight and the

loading are equal to one.

We have addressed the study of the integration into the labour market of Sapienza alumni by developing a few models.

The first one concerns the quantitative study of the job success of the master's degree alumni of the Sapienza University who belong to the engineering disciplinary sector in terms of quality of work. In particular, we studied indicators of job success to estimate their relationships with educational and job curriculum. These indicators have been constructed defining the concept of *optimal contract* based on a contract with the characteristics of permanent position, highly qualified professional position identified using the ISCO classification and the actual duration more than or equal to 8 months. The concept of *quasi-optimal contract* is similar to the optimal definition without the condition of permanent position. The high values of the measurement model have confirmed that the *optimal* and *quasi-optimal* indicators are discriminant to the construction of the Job Success block.

The second and the third models have been built in order to study the job career in terms of the contract type and quality evolution in the three years after graduation. For these studies we selected from the UNI.CO dataset, independently from the disciplinary group, two sets of data: alumni who at the time of the master's degree had not yet obtained an active contract (NON-WORKERS studied by the second model) and alumni who at the time of the master's degree had obtained an active contract (WORKERS studied by the third model). The main purpose we addressed is to model, in the PLS-PM framework, the contract type and job professional quality evolution as latent variables to be studied in relation to educational curriculum, age, initial experience and tendency to evolve. The two models adopted consist essentially of four concepts: the university evaluation, the initial job experience, tendency to evolve and the career evolution. The career evolution is monitored by two constructs, one for the job professional evolution and the other for the job contractual evolution. We have found that in both the models there is a strong influence of the initial job experience and of the tendency to evolve over the job career, while the university evaluation at the level of the master's degree is less important. The age at master's degree has a marginal effect in the model for NON-WORKERS while it disappears for WORKERS.

It is remarkable that the implementation of the logistic regression, in all the presented models, has shown a good classification measured in terms of high AUC values.

The overall frame that arises from these studies is that when there is a natural individual aptitude by the graduates to change job, this tendency plays a positive important role with the aim of obtaining a satisfactory job; on the other hand, the university path of Sapienza alumni does not seem to have a large influence on their job career.

It is clear that our analysis can be considered as a starting point for further studies aimed to investigate the relationships between the world of labour and that of the university education in order to improve the efforts that should be made to integrate these two worlds.

Appendix A

A.1 Exponential Family

Definition: *The distribution of a random variable Y belongs to the simple exponential family if its density function (discrete or continuous) with respect to a σ-finite measure takes the following form:*

$$f(Y;\theta,\phi) = \exp\left\{\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right\} \tag{A.1}$$

where $c(Y,\phi) \geq 0$ is a measurable function.
The parameter $\theta \in \Theta \subset \mathbb{R}$ is called *canonical parameter* of the family, while $\phi > 0$ is a constant parameter called *dispersion parameter* or *nuisance*. Given a fixed value to ϕ, we assume that the set Θ represents the space of canonical parameters, that is the set of θ satisfies the relation

$$0 < \int \exp\left\{\frac{y\theta}{a(\phi)} + c(y,\phi)\right\} dy < \infty,$$

where the $c(y,\phi)$ is a measurable function, independent from θ and for a fixed value of ϕ it is a finite quantity. Finally, $a(\phi)$, is called *scale function*, it is a positive and monotone function; and the usual assumption that can always be taken after reassignment is that $a(\phi) = \phi$.

The exponential family has the following important properties:

$$\mathrm{E}(Y) = \mu(\theta) = \frac{\partial b(\theta)}{\partial \theta} \in \mathcal{M} \subset \mathbb{R} \tag{A.2}$$

$$\mathrm{Var}(Y) = v(\theta) = a(\phi) b''(\theta) \geq 0. \tag{A.3}$$

In generalized linear models, the linearity of the relationship is maintained on the scale used for the transformation of the average value of the response variable.
In the case of $\boldsymbol{q}$-dimensional random variables $\boldsymbol{Y}$ the exponential family takes the following form.
The density function (discrete or continuous) with respect to a σ-finite measure can be written:

$$f(\boldsymbol{y}; \boldsymbol{\theta}, \phi) = \exp\left\{ \frac{y^T \boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(\boldsymbol{y}, \phi) \right\} \tag{A.4}$$

where the canonical parameter of the exponential family is the vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$. The set Θ represents the space of the canonical parameters and it satisfies the relation

$$0 < \int \exp\left\{ \frac{\boldsymbol{y}^T \boldsymbol{\theta}}{a(\phi)} + c(\boldsymbol{y}, \phi) \right\} d\boldsymbol{y} < \infty.$$

The exponential family has the following important properties:

$$\mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{Y}) = \mu(\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathcal{M} \subset \mathbb{R}^q \tag{A.5}$$

$$\mathrm{cov}_{\boldsymbol{\theta}}(\boldsymbol{Y}) = V(\boldsymbol{\theta}) = a(\phi) \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}, \tag{A.6}$$

where it is assumed that the matrix $V(\boldsymbol{\theta})$ is positive definite. Considering the inverse function $\boldsymbol{\theta}(\boldsymbol{\mu})$ it is possible to define the *variance* and *covariance* functions:

$$v(\boldsymbol{\mu}) = \frac{\partial^2 b(\boldsymbol{\theta}(\boldsymbol{\mu}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}, \qquad \mathrm{cov}(\boldsymbol{Y}) = a(\phi) v(\boldsymbol{\mu}), \tag{A.7}$$

that are functions of the mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y})$ and also of the dispersion parameter ϕ.

A.2 Generalized Linear Models

The generalized linear models can represent a number of real life situations, they are essentially based on three assumptions.
Consider the vector of n observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ and a vector $(p+1)$-dimensional covariates (known) $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$; the structure of a GLM is completely specified by the following hypotheses, McCullagh and Nelder (1989) [51] and Nelder and Wedderburn (1972) [52]:

- the *random component*: $\boldsymbol{y}_i$ is the realization of a random variable $\boldsymbol{Y}_i$, $\quad i = 1, \ldots, n$ that has density function belonging to the exponential family. The $\boldsymbol{Y}_i$ are independent and identically distributed random variables. The mean and the covariance matrix of the response variable will be indicated respectively with $\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i = \mu(\boldsymbol{\theta}_i)$ and with $\mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{V}_i = V(\boldsymbol{\theta}_i)$.

- the *systematic component*: the $(p+1) < n$ quantitative explanatory variables $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ produce a theoretical value, called *linear predictor*, $\eta_i = \sum_{j=0}^{p} \beta_j x_{ij}$, which can be written in vector form $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ con $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^{(p+1)}$, where β_j are the parameters to be estimated from the data and $x_{i0} = 1$.
  A more general form for the linear predictor is $\boldsymbol{\eta}_i = \boldsymbol{Z}_i \boldsymbol{\beta}$, where $\boldsymbol{Z}_i = Z(\boldsymbol{x}_i)$ is the *design matrix* of size $q \times (p+1)$ and it is a function of the covariates.

- the *link function*: it is a function that connects the random and systematic component; if $\boldsymbol{\mu}_i$ is the mean value of the random variable $\boldsymbol{Y}_i$, then it is linked to the linear predictor $\boldsymbol{\eta}_i$ through the link function $g(\cdot)$:

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i \iff \boldsymbol{\mu}_i = h(\boldsymbol{\eta}_i)$$

  where the *response function* $h(\cdot) = g^{-1}(\cdot)$ and $g(\cdot)$ are monotone and differentiable functions.
  Given the link function $g : \mathcal{M} \to \mathbb{R}^q$ and the response function $h : \mathbb{R}^q \to \mathcal{M}$,

$$\boldsymbol{\mu}_i = h(\boldsymbol{\eta}_i),\ \boldsymbol{\eta}_i = h^{-1}(\boldsymbol{\mu}_i) = g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i \Rightarrow h^{-1} \equiv g. \tag{A.8}$$

The link function connects $\boldsymbol{\mu}_i$ with the linear component model and allows to extend the range of variation of the expected value of the distributions belonging to the exponential family, which may be limited to a portion of $\mathbb{R}^q$, to all $\mathbb{R}^q$, so that the parameter estimates should not be bound by the type of the response variable that we are using.
Each link function defines an equivalence class whose elements are identical if related by a linear transformation.
For each exponential family distribution, there is a particular link function identified by the equality $\boldsymbol{\theta} = \boldsymbol{\eta}$; that is $g(\boldsymbol{\mu}_i) = g(b'(\boldsymbol{\theta}_i)) = \boldsymbol{\theta}_i = \boldsymbol{\eta}_i$, which implies $b'(\cdot) = g^{-1}(\cdot)$, called *canonical link*. In Tab. A.1 a few canonical links are reported.

Table A.1: Common link functions and mean functions

Distribution	**Name**	**Link Function**	**Mean Function**
Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Exponential	Inverse	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Gamma			
Inverse Gaussian	Inverse squared	$X\beta = \mu^{-2}$	$\mu = (X\beta)^{-1/2}$
Poisson	Log	$X\beta = ln(\mu)$	$\mu = exp(X\beta)$
Binomial	Logit	$X\beta = ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{exp(X\beta)}{1+exp(X\beta)}$
Multinominal			

A.2.1 Parameters estimation

Parameters estimation of generalized linear models is achieved using the *Maximum Likelihood* method. This principle assumes that among the possible $\boldsymbol{\theta}$ values those corresponding to the maximum probability of generating the observed data are to be preferred, Piccolo (2002) [55].
Given a random sample $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^T$ and $\boldsymbol{Y} \sim f(\boldsymbol{y}; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$, the maximum likelihood estimation for $\boldsymbol{\theta}$ is the value $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{Y})$ which maximizes the likelihood function $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y})$.
For the GLM, the likelihood function associated with the exponential density for

q-dimensional random variables (A.4) takes the following form

$$\mathscr{L}(\boldsymbol{\theta}_i, \phi; \boldsymbol{y}_i) = \prod_i f(\boldsymbol{y}_i; \boldsymbol{\theta}_i, \phi) = \exp\left\{\sum_i \left[\frac{\boldsymbol{y}_i^T \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{a(\phi)} + c(\boldsymbol{y}_i; \phi)\right]\right\}. \quad \text{(A.9)}$$

Hence, the log-likelihood function is the sum of the individual likelihood $l_i(\cdot)$

$$l(\boldsymbol{\theta}_i, \phi, \boldsymbol{y}_i) = \sum_i l_i(\boldsymbol{\theta}_i, \phi, \boldsymbol{y}_i) = \sum_i \left[\frac{\boldsymbol{y}_i{}^T \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{a(\phi)} + c(\boldsymbol{y}_i; \phi)\right]. \quad \text{(A.10)}$$

The maximum likelihood estimation is obtained by the value $\boldsymbol{\theta} \in \Theta$ characterized by

$$\boldsymbol{V}'_n(\boldsymbol{\theta}) = \frac{\partial \log \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \frac{\partial\, l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = 0; \quad \boldsymbol{V}''_n(\boldsymbol{\theta}) = \frac{\partial^2 \log \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \cdot \partial \boldsymbol{\theta}^T} = \frac{\partial^2\, l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \cdot \partial \boldsymbol{\theta}^T} < 0,$$

where $\boldsymbol{V}''_n(\boldsymbol{\theta}) < 0$ indicates that the matrix is negative definite.
In general, to solve this problem, it is necessary to resort to numerical methods. The most common procedures are: the *Newton-Raphson* and the *Fisher scoring* methods. This last procedure is included in the class of the *iteratively weighted last squares* methods see A.2.2.
The expression (A.10) is a function of the parameters $\boldsymbol{\beta}$ through the relation $g(\boldsymbol{\mu}_i) = g\left(\frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right) = \boldsymbol{\eta}_i = \boldsymbol{Z}_i\boldsymbol{\beta}$ that corresponds to (A.5), while the function $c(\boldsymbol{y}_i; \phi)$ represents a constant term independent from $\boldsymbol{\beta}$.
Deriving the log-likelihood function with respect to $\boldsymbol{\beta}$ and using the chain rule we get:

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}}\right)^T \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial l_i}{\partial \boldsymbol{\theta}_i},$$

we obtained the *score function* $\boldsymbol{s}(\boldsymbol{\beta})$:

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{\beta}) &= \frac{\partial l(\cdot)}{\partial \boldsymbol{\beta}} = \sum_i \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}}\right)^T \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial l(\cdot)}{\partial \boldsymbol{\theta}_i} \\ &= \frac{1}{a(\phi)} \sum_i \boldsymbol{Z}_i^T \boldsymbol{\Delta}_i \left[\boldsymbol{y}_i - b'(\boldsymbol{\theta}_i)\right] \end{aligned} \quad \text{(A.11)}$$

where $\boldsymbol{\Delta}_i = \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\eta}_i}$.
An important property of the score function is that its expected value is equal to zero, $\mathrm{E}[\boldsymbol{s}(\boldsymbol{\beta})] = 0$.
Remembering that $b'(\boldsymbol{\theta}_i) = \boldsymbol{\mu}_i$, and that $b''(\boldsymbol{\theta}_i) = v(\boldsymbol{\mu}_i)$, it is possible to obtain $\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_i} = v(\boldsymbol{\mu}_i) = \frac{\boldsymbol{V}_i}{a(\phi)}$, where the parameter ϕ can be interpreted either as a scaling parameter of the likelihood function or as a dispersion parameter of the variance function; from the definition of the linear predictor $\boldsymbol{\eta}_i = \boldsymbol{Z}_i\boldsymbol{\beta}$ it is possible to obtain $\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} = \boldsymbol{Z}_i$. Then (A.11) can be written as

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{\beta}) &= \sum_i \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}}\right)^T \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i}\frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\mu}_i}\frac{\partial l(\cdot)}{\partial \boldsymbol{\theta}_i} \\ &= \sum_i \boldsymbol{Z}_i^T \boldsymbol{D}_i \, \boldsymbol{V}_i^{-1}[\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] \end{aligned} \tag{A.12}$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i}$ represents the *Jacobian* of the trasformation $\boldsymbol{\mu}_i = h(\boldsymbol{\eta}_i)$ calculated in $\boldsymbol{\eta}_i$.
For the canonical link, the relationship $\boldsymbol{\theta}_i = \boldsymbol{\eta}_i$ holds, therefore we have that $\boldsymbol{\Delta}_i = \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\eta}_i} = \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is the q-dimensional identity matrix, and $\boldsymbol{D}_i = [v(\boldsymbol{\mu}_i)]$, so using the expressions (A.11) and (A.12) the conditions of extremum can be written unless of a multiplicative constant not depending on $\boldsymbol{\beta}$, as:

$$\boldsymbol{s}(\boldsymbol{\beta}) \propto \sum_i \boldsymbol{Z}_i^T[\boldsymbol{y}_i - b'(\boldsymbol{\theta}_i)] = \boldsymbol{0}. \tag{A.13}$$

Another advantage of using the canonical link is that for them a sufficient statistic exists with the same size of the parameter vector, $\boldsymbol{\beta}$. This sufficient statistic is equal to $\sum_i \boldsymbol{Z}_i^T\boldsymbol{y}_i = \boldsymbol{Z}^T\boldsymbol{y}$, where the $(n \times q, p+1)$-dimensional matrix $\boldsymbol{Z}$ is given by $\boldsymbol{Z} = \{\boldsymbol{Z}_i\} \otimes \boldsymbol{1}_n = (\boldsymbol{Z}_i^T, \ldots, \boldsymbol{Z}_n^T)^T$ and $\boldsymbol{y}$ has dimensions $(n \times q, 1)$ and is given by $\boldsymbol{y} = \{\boldsymbol{y}_i\} \otimes \boldsymbol{1}_n$. As function of $\boldsymbol{\beta}$, the matrix of the negative second derivative of the log-likelihood function is called the *Observed Fisher Information Matrix*:

$$\boldsymbol{F}_{obs}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\cdot)}{\partial \boldsymbol{\beta} \cdot \partial \boldsymbol{\beta}^T}. \tag{A.14}$$

The *Expected Fisher Information Matrix* is defined as

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\beta}) &= \mathrm{E}(\boldsymbol{F}_{obs}(\boldsymbol{\beta})) \\ &= \sum_i \boldsymbol{Z}_i^T \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T \boldsymbol{Z}_i. \end{aligned} \tag{A.15}$$

In the case of the canonical links, the expressions (A.14) and (A.15) coincide and they can be written as

$$\mathrm{E}\left(-\frac{\partial l(\cdot)}{\partial \boldsymbol{\beta} \cdot \partial \boldsymbol{\beta}^T}\right) = \frac{1}{a(\phi)} \sum_i \boldsymbol{Z}_i^T b''(\boldsymbol{\theta}_i) \boldsymbol{Z}_i = \frac{1}{[a(\phi)]^2} \sum_i \boldsymbol{Z}_i^T \boldsymbol{V}_i \boldsymbol{Z}_i \tag{A.16}$$

It is important to note that ϕ is irrelevant for the estimation of the model parameters given that the function $a(\phi)$ assumes a role of a simple multiplicative factor independent of the parameters $\boldsymbol{\beta}$. When the model hypothesis $h(\boldsymbol{\mu}_i) = \boldsymbol{Z}_i\boldsymbol{\beta}$ is validated, the estimation $\hat{\boldsymbol{\beta}}$ is consistent even if the variance is not correctly specified. In general, the maximum likelihood estimations $\hat{\boldsymbol{\beta}}$ are not computed searching the maximum of the log-likelihood function $l(\cdot)$ but are computed solving the following homogeneous system

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = 0 \tag{A.17}$$

which correspond to points in the parameters space for which $\boldsymbol{F}_{obs}(\hat{\boldsymbol{\beta}})$ is positive definite. In many models, the log-likelihood function $l(\cdot)$ is strictly concave, e.g. the logistic model, with respect to $\boldsymbol{\beta}$ therefore the local and global maxima coincide.
Inferential methods for GLMs rely on asymptotic properties of ML estimators. Under 'regularity assumptions' [1] , the following properties hold:

1. *Asymptotic existence and uniqueness*: the probability that $\hat{\boldsymbol{\beta}}$ exists and is (locally) unique tends to 1 for $n \to \infty$;

2. *Consistency*: if $\boldsymbol{\beta}$ denotes the 'true' value, then for $n \to \infty$ we have $\hat{\boldsymbol{\beta}} \xrightarrow{a} \boldsymbol{\beta}$ in probability (weak consistency) or with probability 1(strong consistency);

[1] In the standard theory of asymptotic for ungrouped data, the 'regularity assumptions' require only that, for $n \to \infty$, $\boldsymbol{F}^{-1}(\boldsymbol{\beta}) \to 0$ in conjunction with continuos property of $\boldsymbol{F}(\boldsymbol{\beta})$.

3. *Asymptotic normality* : the distribution of the (normed) maximum likelihood estimator becomes normal for $n \to \infty$ i.e. for large n

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})) \tag{A.18}$$

$\hat{\boldsymbol{\beta}}$ is approximately normal with asymptotic covariance matrix

$$\text{cov}(\hat{\boldsymbol{\beta}}) \overset{a}{=} \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$ is the inverse of the Fisher matrix computed in $\hat{\boldsymbol{\beta}}$.

For an unknown scale parameter ϕ, all results remain valid if it is replaced by a consistent estimate $\hat{\phi}$.

A.2.2 Numerical methods

- The *Newton-Raphson* method: is a numerical algorithm with the aim to search the zeros of any function. Therefore it is employed to find the maximum of the log-likelihood function, i.e. the zeros of the first derivative. Let us assume that this function depends on the vector $\boldsymbol{\theta}$, $g(\boldsymbol{\theta}) = \boldsymbol{V}'_n(\boldsymbol{\theta})$ is the vector of the first derivative (the gradient) and with $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{V}_n{}''(\boldsymbol{\theta})$ the matrix of the second derivative, i.e. the Hessian Matrix. Using the first order Taylor approximation, it is possible to linearize $g(\boldsymbol{\theta})$ around a value $\boldsymbol{\theta}_0$,

$$g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_0) + \boldsymbol{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \dots.$$

Let us call $\boldsymbol{t}_n$, the estimation value $\boldsymbol{\theta}$: $g(\boldsymbol{t}_n) = 0$ then $\boldsymbol{\theta}$ con $\boldsymbol{t}_n$

$$0 = g(\boldsymbol{\theta}_0) + \boldsymbol{H}(\boldsymbol{\theta}_0)(\boldsymbol{t}_n - \boldsymbol{\theta}_0) + \ldots \Longrightarrow \boldsymbol{t}_n = \boldsymbol{\theta}_0 - [\boldsymbol{H}(\boldsymbol{\theta}_0)]^{-1} g(\boldsymbol{\theta}_0) + \dots.$$

Therefore, the approximation of the solution for the maximum likelihood estimation is

$$\boldsymbol{t}_n = \boldsymbol{\theta}_0 - [\boldsymbol{H}(\boldsymbol{\theta}_0)]^{-1} g(\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0 - [\boldsymbol{V}''_n(\boldsymbol{\theta}_0)]^{-1} \boldsymbol{V}'_n(\boldsymbol{\theta}_0).$$

Given an initial fixed value $\boldsymbol{t}^{(0)}$, the subsequent values are obtained by iterations

$$\boldsymbol{t}^{(i)} = \boldsymbol{t}^{(i-1)} - [\boldsymbol{V}_n{}''(\boldsymbol{t}^{(i-1)})]^{-1}\boldsymbol{V}_n{}'(\boldsymbol{t}^{(i-1)}) \quad i = 1, 2, \ldots \tag{A.19}$$

- The *Fisher Scoring* method exploits the proprieties of the maximum likelihood function replacing in Eq. (A.19) the Hessians $\boldsymbol{V}_n{}''(\boldsymbol{t}^{(0)}), \boldsymbol{V}_n{}''(\boldsymbol{t}^{(1)}), \boldsymbol{V}_n{}''(\boldsymbol{t}^{(2)}), \ldots$ with the respectively expected values $\mathrm{E}[\boldsymbol{V}_n{}''(\boldsymbol{T}^{(0)})], \mathrm{E}[\boldsymbol{V}_n{}''(\boldsymbol{T}^{(1)})], \mathrm{E}[\boldsymbol{V}_n{}''(\boldsymbol{T}^{(2)})], \ldots$. When the average values are a function of the parameters, the latter should be replaced by numeric values determined in each iteration. These quantities can be evaluated by using the first derivative:

$$\mathrm{E}\left(\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right) = -\mathrm{E}\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \log L(\boldsymbol{\theta})}{\partial\theta_j}\right).$$

It is possible to demonstrate that the Newton Raphson and the Fisher Scoring Method coincide when the random variables belong to the exponential family and the canonical link is adopted.
Therefore the Fisher scoring iterative algorithm in terms of $\boldsymbol{\beta}$, can be expressed in this way. It starts from an initial value $\hat{\boldsymbol{\beta}}^{(0)}$ and iterations are defined by the following expression:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{s}(\hat{\boldsymbol{\beta}}^{(k)}), \quad k = 0, 1, \ldots \tag{A.20}$$

The algorithm terminates when a stopping criterion is satisfied, e.g.

$$\frac{||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||}{||\hat{\boldsymbol{\beta}}^{(k)}||} \leq \epsilon$$

where $\epsilon > 0$ is a value chosen a priori generally small.
It can be shown that these iterative algorithms can be reformulated in such a way to be included in the class of the *Iteratively Weighted Least Squares, IWLS.*

Defining $\boldsymbol{W}_i$ the generalized weights, and $\boldsymbol{e}_i$ the generalized residuals:

$$\boldsymbol{W}_i = \frac{1}{a(\phi)}\boldsymbol{\Delta}_i b''(\boldsymbol{\theta}_i)\boldsymbol{\Delta}_i^T = \boldsymbol{D}_i \boldsymbol{V}_i^{-1}\boldsymbol{D}_i^T \tag{A.21}$$

$$\boldsymbol{e}_i = \frac{1}{a(\phi)}\boldsymbol{W}_i^{-1}\boldsymbol{\Delta}_i[\boldsymbol{y}_i - b'(\boldsymbol{\theta}_i)] = \boldsymbol{D}_i^{-1}[\boldsymbol{y}_i - b'(\boldsymbol{\theta}_i)]. \tag{A.22}$$

and denoting the *working or adjusted observations* as

$$\begin{aligned} \tilde{\boldsymbol{y}}(\boldsymbol{\beta}) &= (\tilde{y}_1(\boldsymbol{\beta}), \ldots, \tilde{y}_n(\boldsymbol{\beta}))^T \\ \tilde{\boldsymbol{y}}_i(\boldsymbol{\beta}) &= \boldsymbol{z}_i^T\boldsymbol{\beta} + \boldsymbol{D}_i^{-1}(\boldsymbol{\beta})[\boldsymbol{y}_i - \mu_i(\boldsymbol{\beta})] \end{aligned} \tag{A.23}$$

the expression that defines the iterations of the Fisher scoring (A.20) takes the form of an iterative weighted least squares algorithm:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(k+1)} &= \hat{\boldsymbol{\beta}}^{(k)} + (\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{e}^{(k)} \\ &= (\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\tilde{\boldsymbol{y}}^{(k)}, \end{aligned} \tag{A.24}$$

with $\boldsymbol{W} = \{\boldsymbol{W}_i\} \otimes \boldsymbol{I}_n$, $\boldsymbol{e} = \{\boldsymbol{e}_i\} \otimes \mathbf{1}_n$, $\tilde{\boldsymbol{y}} = \{\tilde{\boldsymbol{y}}_i\} \otimes \mathbf{1}_n$ and where $\boldsymbol{W}^{(k)}$ and $\tilde{\boldsymbol{y}}^{(k)}$ are the estimation of $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ when $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}^{(k)}$.

In this algorithm we have that the generalized weights, the generalized residuals and the transformed observations depend on the linear predictor, and then on the regression parameters implying that quantities (A.24) must be recalculated at each iteration.

In general, the convergence is rapid and the iteration stops near the maximum in a few steps (< 10). The default value for the maximum number of iteration is fixed to 25 in the *glm* R-function. It can happen that the procedure diverges i.e. successive differences $||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||$ increase. This can indicate a bad initial value $\hat{\boldsymbol{\beta}}^{(0)}$ or more often the nonexistences of a maximum of the likelihood function within the admissible parameters set. When this set coincides with R^p one of the $\boldsymbol{\beta}$ components goes to infinity.

In the case of the logistic regression the occasional problem of the failure of the likelihood maximization is generally associated with data patterns characterized

by *complete* or *quasi-complete separation*. A *complete separation* occurs when the outcome variable completely separates a combination of predictor variables. In other words, the outcome variable is predicted perfectly. Complete separation occurs for small samples while a quasi-separation is a more frequent phenomenon. For these patterns the maximum likelihood estimate does not exist.

Formally, in case of separation a sub-vector of the covariates exists by which all the observations can be classified as either zero or one. A geometrical interpretations, due to Agresti (2002) [1], is the following: *"a hyperplane can pass through the space of predictor values such that on one side of that hyperplane* $Y = 0$ *for all observations, while in the other side* $Y = 1$ *always"*.

In the case of *glm* R-function, the problem is pointed out by warning messages like: "*failed to converge*" and "*fitted probabilities numerically 0 or 1 occurred*". In these cases a few strategies to overcome the problem are discussed in literature, Albert and Anderson (1984) [2]. The more promising method seems to be the so called *penalized maximum likelihood method* by Firth (1993) [27] and its improvements but the issue is still under debate, Zorn (2005) [72].

Bibliography

[1] Agresti, A., Categorical Data Analysis, 2nd Ed. New York: Wiley (2002).

[2] Albert, A. and Anderson, J. A. (1984), On the Existence of Maximum Likelihood Estimates in Logistic Regresson Models, Biometrika **71**(1):1-10.

[3] Alleva, G. and Petrarca, F. (2013), New indicators for investigating the Integration of Sapienza graduates into the labour market, working papers n. **120**/2013 del Dipartimento Memotef, ISSN 2239-608X.

[4] Alleva, G., Petrarca, F., Renda, E., Lucisano, P. and Magni, C. (2012),*Potenzialità della matrice UNI.CO. per lo studio delle caratteristiche della domanda di lavoro dei laureati della Sapienza: primi risultati e possibili sviluppi*, Workshop on"Monitoring of the dynamics of the professional graduates". Italia Lavoro, 24th September 2012.

[5] AlmaLaurea (2013), Condizione occupazionale dei Laureati. XV Indagine 2012, www.almalaurea.it/universita/occupazione.

[6] Baron, R. M., and Kenny, D. A. (1986), *The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations*. Journal of Personality and Social Psychology, Vol. 51, No. 6, 1173-1182.

[7] Bollen, K. A.(1989), *Structural equations with latent variables*. Wiley, New York. MR0996025.

[8] Bollen, K. A. and Ting, K. F. (1993), *Confirmatory Tetrad Analysis*, Sociological Methodology **23**, 147-175.

[9] Breiman, L. and Friedman,J., *Estimating optimal transformations for multiple regression and correlation.* JASA **80**, 580597, (1985).

[10] Breiman, L., Friedman,J., Olshen, R. and Stone, C., *Classification and regression trees*, Wadsworth International Group, Belmont, CA, 1984.

[11] Capecchi, S., Iammario, M. and Piccolo, D. (2012), *Modelling Job Satisfaction in AlmaLaurea surveys.* AlmaLaurea working paper n.50, ISSN 2239-9453.

[12] Carpita, M., (2011) Laureati Stella: Rapporto statistico 2008-2010, ISBN 978-88-88971-25-4, CILEA .

[13] Carroll, JD., *Generalization of canonical analysis to three or more sets of variables*, Proceedings of the 76th Convention of the American Psychological Association, vol. 3, 1968; 227228.

[14] CENSIS (2012), 46esimo Rapporto sulla situazione sociale del Paese/2012.

[15] Chin, W. W. (1998), *The partial least squares approach for structural equation modeling* in G. A. Marcoulides (Ed.), Modern methods for business research (pp. 295-236). London: Lawrence Erlbaum Assiciates.

[16] Ciriaci, S. and Muscio, A.,(2011) *University choice, research quality and graduates' employability: Evidence from Italian national survey data.* AlmaLaurea working paper n.48, ISSN 2239-9453.

[17] Clark, A.(1998), *What Makes a Good job? Evidence From OECD Countries*, CSLS Conference on the State of Living Standards and the Quality of life in Canada, Château Laurier Hotel, Ottawa, Ontario.

[18] Cohen, J. and Cohen, P., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* New York: Wiley, (1975).

[19] D'Ambra, L. and Lauro, C. N. (1982), *Analisi in componenti principali in rapporto a un sottospazio di riferimento.* Rivista di Statistica Applicata **15**(1), 51-67.

[20] De Leeuw, J., Young, F.W., Takane Y.: *Additive structure in qualitative data: an alternating least squares method with optimal scaling features.* Psychometrika 41, 471503 (1976).

[21] Efron, B (1982),*The Jackknife, the bootstrap, and other resampling plans*, SIAM, Philadelphia. MR0659849

[22] Esposito Vinzi, V., Chin W.W., Henseler, J. and Wang, H. (Eds.) (2010).*Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications*, Springer Verlag, Berlin Heidelberg. MR2742562

[23] Esposito Vinzi, V. and Russolillo, G.(2013): Partial Least Squares algorithms and methods. WIREs Comput Stat 2013, 5:1-19. doi:10.1002/wics.1239.

[24] European Commission, *Measuring the quality of employment in the EU.* in Employment in Report, Chapter 4, Directorare-General for Employment, Social Affairs and Equal Opportunities, Brussels, 2008.

[25] Fabbris, L., *Indicators of higher education effectiveness*, McGraw-Hill Education, Milano, 2012.

[26] Fawcett, T. (2006), *An introduction to ROC analysis.* Pattern recognition letters, **27**, 861–874.

[27] Firth, D. (1993), *Bias reduction of maximum likelihood estimates.* Biometrika, **80**: 27-38.

[28] Fornell, C. & Larcker, D. F. (1981), *Evaluating structural equation models with unobservable variables and measurement error*, Journal of Marketing Research **18**, 39-50.

[29] Fornell, C. and Bookstein, F. L. (1982), *Two structural equation models: LISREL and PLS applied to consumer exit-voice theory.* Journal of Marketing Research, **19**, 440-452.

[30] Frey, B, S. and Stutzer, A. (2002), *What can economists learn from happiness research?*, Journal of Economic Literature, Vol. xl, pp. 402-435.

[31] Glang, M. *Maximierung der Summe erklärter Varianzen in linearrekursiven Strukturgleichungsmodellen mit multiple Indikatoren: Eine Alternative zum Schätzmodus B des Partial-Least-Squares-Verfahren*, PhD Thesis, Universität Hamburg, Hamburg, Germany, 1988.

[32] Green, F. (2006), *Demanding work: the paradox of job quality in the affluent economy*, Princeton University Press.

[33] Gudergan, S. P., Ringle, C. M., Wende, S. & Will, A. (2008), *Confirmatory tetrad analysis in PLS path modeling*, Journal of Business Research **61 (12)**, 1238-1249.

[34] Hanafi M. *PLS path modeling: computation of latent variables with the estimation mode B*. Comput. Stat 2007, 22:275292.

[35] Horst, P., *Relations among m sets of measures*. Psychometrika, 1961, 26:129 149.

[36] Hosmer,D. W. and Lemeshow, S., *Applied logistic regression*, 2nd edition, John Wiley and Sons, 2000.

[37] Hotelling, H. (1933). *Analysis of a complex of statistical variables into components*, Journal of Educational Psychology **24**.

[38] ISTAT (2012), Report: I Laureati e il Lavoro. Indagine 2011 sui laureati del 2007.

[39] Jencks, C., Perman, L. and Rainwater, L. (1988), *What is a good job? A new measure of labor-market success*, The American Journal of Sociology, 1988, **93(6)**: pp. 1322-1357.

[40] Jöreskog, K. (1970), A general method for analysis of covariance structure, *Biometrika* **57**, 239-251. MR0269024

[41] Kaplan, D.(2000), *Structural equation modeling: foundations and extensions*. Thousands Oaks, California: Sage.

[42] Kettenring, JR. *Canonical analysis of several sets of variables.* Biometrika 1971, 58:433451.

[43] Krämer, N., *Analysis of high-dimensional data with partial least squares and boosting*, PhD Thesis, Technische Universität Berlin, Berlin, Germany, 2007.

[44] Krus, D.J. and Wilkinson, S.M. (1986).*Demonstration of properties of a suppressor variable.* Behavior Research Methods, Instruments, and Computers, 18, 21-24.

[45] Kruskal, J. (1964), *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*, Psychometrika, **29** (1), 1-27.MR0169712.

[46] Leschke, J. and Watt, A., *Job Quality in Europe*, WP 2008-07, ETUI-REHS, Brussels, 2008.

[47] Leschke, J. and Watt, A. with Finn, M., *Putting a Number on Job Quality? Constructiong a European Job Quality Index*, WP 2008-03, ETUI-REHS, Brussels, 2008.

[48] Lohmöller, J. (1989),*Latent variable path modeling with partial least squares.* Physica-Verlag, Heildelberg. MR1014472.

[49] Lyttkens,. E., Areskoug, B., Wold, H., *The convergence of NIPALS estimation procedures for six path models with one or two latent variables*, Technical Report, University of Goö teborg, 1975.

[50] Mathes H. *Global optimisation criteria of the pls- algorithm in recursive path models with latent variables.* In: Haagen K, Bartholomew DJ, Deister M, eds. Statistical Modelling and Latent Variables. Amsterdam: Elsevier Science, 1993.

[51] McCullagh, P. and Nelder, J.. *Generalized Linear Models*, Second Edition, Boca Raton: Chapman and Hall/CRC.(1989). ISBN 0-412-31760-5.

[52] Nelder, J. and Wedderburn, R. (1972). *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General) (Blackwell Publishing) **135** (3): 370-384.

[53] Petrarca, F. (2013), *Non-Metric PLS Path Modeling: the job success of Sapienza graduates*, nel libro elettronico "Advances in Latent Variables" Eds. Brentari E., Carpita M.,Vita e Pensiero, Milan, Italy, ISBN 978 88 343 2556 8.

[54] Petrarca F. (2014), *Non-Metric PLS Path Modeling: integration into the labour market of Sapienza graduates*, has been accepted for publication in Studies in Theoretical and Applied Statistics, manuscript number, STAS-D-13-00033R1.

[55] Piccolo, D., *Statistica*, Il Mulino, 2002.

[56] Russolillo, G. (2012): *Non-Metric Partial Least Squares*. Electronic Journal of Statistics. Vol. 6, 1641-1669.

[57] Russolillo, G., (2013), *Beyond the measurement scale: the Non-Metric Partial Least Squares approach*, Oral presentation at the 6th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2013), 14-16 December 2013.

[58] Sanchez, G. (2013): PLS Path modeling with R.
`http://www.gastonsanchez.com` PLSPM Path Diagram

[59] Sanchez, G., Trinchera, L. and Russolillo, G., (2013): R package of PLSPM.
`http://cran.r-project.org/web/packages/plspm/index.html`

[60] Schmid, G. and Gazier, B. (2002), The dynamics of full employment. Social integration through transiotional labour markets, Cheltenham, Edward Elgar.

[61] Stevens, S. S. (1946), *On the Theory of scales of measurement*, Science **103 (2684)**, 677-680.

[62] Tenenhaus, M.(2009), *A criterion based PLS approach to structural equation modelling*, in Programme and Abstract of the 6th International Conference in Partial Least Squares Methods, p.3.

[63] Tenenhaus, M. & Hanafi, M., *A bridge between PLS path modeling and multi-block data analysis*. In: Esposito Vinzi V, Chin W, Henseler J, Wang H, eds. Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications. Heidelberg, Germany: Springer Verlag; 2010.

[64] Tenenhaus, A., Tenenhaus, M., *Regularized generalized canonical correlation analysis*. Psychometrika 2011, 76:257 284.

[65] Tenenhaus, M., Amato, S. and Esposito Vinzi, V. (2004),*A global goodness-of-fit index for PLS structural equation modelling*. Proceedings of the XLII SIS Scientific Meeting, Vol. Contributed Papers, CLUEP, Padova, pp. 739-742.

[66] Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., and Lauro, C. (2005), *PLS path modeling*. Computational Statistics and Data Analysis, 48, 159-205.

[67] Tilly, C., *The good, the bad, and the ugly: Good and Bad Jobs in the United States at the Millennium*, Russell Sage Foundation, New York, 1996.

[68] Tucker, L. R., (1958). *An Inter-Battery Method of Factor Analysis*, Psychometrika **23**, 111-136.MR0099737

[69] Van de Wollemberg, A. L., (1977). *Redundancy analysis. An alternative for canonical correlation analysis*, Psychometrika **42**, 207-219.

[70] Wold, H. (1982),*Soft modeling: the basic design and some extensions.*In K. G. Jöreskog, and H. Wold, (Eds.), *Systems under indirect observation*, Part II (pp. 1-54). Amsterdam: North-Holland.

[71] Young, F. W. (1981), *Quantitative analysis of qualitative data*. Psychometrika **44** (4), 357-388. MR0668307

[72] Zorn, C. (2005), *A Solution to Separation in Binary Response Models*, Political Analysis 13(2): f157-170.