**“ALEXANDRU IOAN CUZA” UNIVERSITY OF IAŞI**

**FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION**

**DOCTORAL SCHOOL OF ECONOMICS AND BUSINESS ADMINISTRATION**

# Estimation procedures for multilevel models

## SUMMARY OF THE THESIS

Scientific coordinator

Prof. Univ. Dr. **Anca Laura Asandului**

PhD

**Hrițcu** Roxana-Otilia-Sonia

IAȘI

2015

The reality can be described and analyzed by using various models and types of data. Studying the social phenomena implies the administration and analysis of complex data sets. These data are hierarchical: the individual represents the first level and the various groups of individuals being are the upper level of the data hierarchy.

The natural nesting of invividuals in a society, the group and environment membership, generates not only a mutual connection between the individuals of the same group but also a reciprocal connection between the individuals and the group or between the individuals and society. The social context influences individuals opinions, actions and behaviour leading to a connection between the characteristics of the individual and the characteristics of the society or between the characteristics of the individual and his group. The individuals are interferring with their social groups and are influenced by them. Also, the groups are in turn influenced by the individuals who make up that group (Hox, 2010: 1).

One of the biggest challenges of the statistical analysis was the integration in a single model of analysis the available data at the individual level and at the group level.(de Leeuw J., Meijer E., 2008: 3). The single level data analysis cannot take into account and analyze simultaneously variables from different levels. Once the researchers have discovered the utility of multilevel models for grouped data, the multilevel analysis domain has developped more and more. The usage of multilevel models has been rapidly gowing and these models have been used in many areas and domains of research. The multilevel analysis is frequently used in social sciences, education, health, demography psychology, epidemology, biology, environment studies and many other domains that work with groupped data.

**The purpose, goals and assumptions of the thesis**

The aim of this research was to study the estimation procedures for multilevel models, from both theoretical and practical perspectives. Through a multilevel analysis we aimed to emphasize the characteristics, pros and cons for some estimation procedures for generalized multilevel models. For the data analysis we used three statistical analysis programs: SPSS, a program commonly used in research, the statistical programming language R and MLwiN , a program dedicated to multilevel analysis. The three programs use different methods of estimation of generalized multilevel models: Laplace approximation, adaptive quadrature Gauss-Hermite and restricted quasi-likelihood are implemented in R , restricted quasi-likelihood is implemented in SPSS, marginal and restricted quasi-likelihood are implemented in MLwiN. We also aimed at comparing the estimation results for a multilevel model through different procedures, via the 3 programs.

Complementary to the above presented scope, this research had certain established objectives that sustained and realized its scope, on one hand and functioned as validation keys during the reserach, on the other.

Therefore, a first objective of this research views the literature review regarding specific concepts and notions for the multilevel domain, tackling the researched theme, namely the multilevel models estimation. This objective is pursued along the first theoretical part of the thesis that includes the first five chapters and implies the synthesis and structuring of the literature information on multilevel models and their estimation. The importance of this objective for the scope of the research is the need to better know the notions and concepts related to multilevel models and their estimation. The theoretical framework is essential to be established in a doctotal thesis and assures a basis for the future research.

Our interest for generalized multilevel models leads to a second objective that relies in the implementation and analysis of a generalized multilevel model for a data set with two levels and a binary response variable. The data application presented in chapter 6 is a multilevel analysis on the opinion regarding the importance of work in various European Union countries dependind on the work independence level, marital status, level of education and gender.

The third objective aimes at comparing the results given by the traditional one level analysis approach with the multilevel modelling results. This objective is met with the development of the analysis for the opinion on work importance, an analysis implemented through multiple logistical regression and multilevel modelling. This objective is important for the research since this comparison points out the advantages and limits of each analysis procedure that is used foe our data analysis.

The forth objective that we consider is the comparison between the estimations through the marginal quasi-likelihood procedure (MQL) and those through the restricted quasi-likelihood procedure (PQL) for a two level model with a binary response variable. This objective is met in chapter 6 with the analysis of the multilevel model over the opinion regarding the importance of work. The importance of this objective represents the possibility to contribute to the quasi-likelihood procedures' evaluation for two level models and binary response variable.

Different statistical analysis programs work with different methods of estimation and therefore may lead to different estimates. Therefore, a fifth objective concerns comparing the estimation results for a generalized multilevel model using the statistical analysis software R, SPSS and MLwiN. This objective is met in the data analysis for the importance of work. The estimation is performed by different procedures of estimation implemented in each program: approximation Laplace, Gauss- Hermite adaptive quadrature and restricted quasi-likelihood in R, restricted quasi-likelihood in SPSS, marginal and restricted quasi-likelihood in MLwiN.

One last objective we consider is the use of bootstrap sampling to correct the bias of the quasi-likelihood estimators and to improve the accuracy of conclusions (Rasbash and others, 2014: 259). This objective is also reached by the multilevel analysis on the opinion regarding the importance of work.

## The structure of the thesis

The thesis is structured in such a way that the reader, regardless of his economic and econometrics knowledge, can follow the research process. The reader is, thereby, step by step initiated to the used notions and concepts and is accustomed to the applied procedures and their rationale.

As a result, after an introduction to the multilevel research purpose, the first chapter is a synthesis based on the literature review of the basic concepts used in multilevel research: multilevel data, multilevel model and multilevel analysis. The synthesis of theoretical notions is completeed by the multilevel data definition, multilevel data and multilevel models classification. After clarifying these basic concepts, we present examples from various areas of research that indicate the obvious utility of these models. Applying the multilevel models in reasearch areas is a vast subject that can be continuously extended due to the increased interest for this area of analysis and due to the rapid development of software applications and technologies in statistics analysis.

In the second chapter we present the necessity to apply the multilevel models and, consequently, the necessity for multilevel models estimation. We review essential information in the literature regarding the estimation of multilevel models, information which is presented in the third chapter for continuos response variable. The third chapter introduces procedures and algorithms proposed for obtaining the maximum likelihood estimators for liniar multilevel models; we sum up the debate with a comparision between these procedures by pointing out their advantages and limits. The fourth chapter presents the estimation of multilevel models for discrete answer variables; for generalized multilevel models we discuss about quasi-likelihood, numeric approximation and sampling methods used in multilevel statistics programs. Finally, we propose a comparison between these procedures, by taking into account their application distinctivness, and thus, pointing out the opportunity of their usage in generalized multilevel models estimation.

Considering that multilevel data analysis has been developping more and more recently due to software instruments for statisics analysis, we consider of interest to present some of these software programs. Thus, in the fifth chapter, out of the statistics software that offer multilevel analysis, we introduce R, a free programming language, SPSS a software used in the academic environment, and MLwiN, a multilevel analysis dedicated software. Consequently, we describe each software by taking into account its characteristics regarding the interface, working tools, and limits of multilevel analysis, like the use of a certain type of analysis, model or method of estimation. We round up the presentation with a synthesis of available estimation procedures for generalized multilevel models in each software.

The sixth chapter introduces the analysis of a hierarchically structured data set, with two levels, regarding the opinion on the importance of work in nine European Union's countries. The data analysis is performed using 3 statistical analysis programs R, SPSS and MlwiN, by using the various methods available in each program: Laplace approximation, Gauss-Hermite adaptive quadrature, restricted quasi-likelihood in R, restricted quasi-likelihood in SPSS, marginal and restricted quasi-likelihood in MLwiN. We also use bootstrapping sampling and we compare the results of each estimation procedure.

We finalyze the thesis by presenting the conclusions for the entire research, theoretical and practical - a multilevel analysis for the opinion regarding the importance of work. The conclusions of the research are rounded up by our recommendations based on the results for the opinion on the importance of work, by the limitations of study and future research proposals.

**Scientific contributions of the doctoral thesis**

In our opinion, the following contributins are added to the economic domain.

The first chapter presents essential theoretical elements used in multilevel research: multilevel data, multilevel model and multilevel analysis. The synthesis of theoretical concepts leads to the definition of multilevel data, clasification of multilevel data and multilevel models. As the multilevel area was recently approached by researchers and we find no clear synthesis and patterning of these basic theoretical concepts in the Romanian literature, the contribution of this research is important. By claryfing and structuring the multilevel domain concepts we offer a first image of this research area. Moreover, by introducing examples from various areas of research we indicate the obvious utility of multilevel models in all research domains.

In the second chapter, the contribution of this research is emphasized by the synthesis and structuring of national and international literature regarding the theoretical aspects of multilevel models' estimation. Based on books and articles and previous research in the multilevel area, we identify and systematize the terms, definitions and the particularities of multilevel model estimation, these being without doubt an added value to this research.

The contribution of the third and fourth chapters signifies the presentation in Romanian of the procedures for the estimation of liniar multilevel models and of generalized multilevel models. A unitary presentation of these procedures described in the international literature means a clear contribution to the Romanian literature. We also propose, within the third chapter, a comparison between the estimation procedures for liniar multilevel models. This comparison is a result of our research and has a role in establishing and understanding the theoretical notions. Last but not least, in the end of the fourth chapter we propose a comparison of the generalized multilevel models procedures. We consider that the structuring of the estimation procedures' characteristics toghether with the emphasis of pro and cons for using these procedures could offer a real support in choosing the estimation procedure based on research purpose.

The contribution added to multilevel models area is also emphasize by the presentation of statistics analysis softwares introduced in the fifth chapter. We considering our choice of programs to be appropriate: the R software, a free programming language for statistics analysis, SPSS frequently used in the academic environment, and MLwiN a dedicated multilevel analysis software. The presentation of the three softwares with different features means catching the different interface particularities, analysis instruments or limitations of multilevel analysis. Also, the summary of the research results regarding the estimation procedures in all three programmes for the generalized multilevel models, is an useful instrument to choose the needed software and the estimation procedure when the research implies multilevel data. Based on this synthesis, the researcher can decide to compare the research results with different procedures of the same software or to use the same estimation procedure in different softwares.

The sixth chapter is clearly a major contribution that this thesis brings to the economic area. We analyse the multilevel data regarding the opinion on the importance of work. The multilevel model had two levels of analysis, the respondents and the country of residence, and the response variable is binary (work can be or cannot be important in each individual’s life). We introduce the multilevel model to determine the impact of several variables on the opinion regarding the importance of work and the country's influence on this opinion. The multilevel data on the importance of work is modelled by logistic regression and the multilevel model, for bothe the null model and the complete model, using 3 statistical analysis programs: R , SPSS and MLwiN .

By applying the estimation procedures of the three statistical software programs, to logistic regression and multilevel model, we can compare: the estimation results for the single level model and for two level model, the estimation results of different estimation procedures, as well as the estimation results of the same estimation procedures implemented different programs. The bootstrap sampling for the complete multilevel model is performed with parameters. The simulation is implemented the MQL1 and PQL2 estimation by using IGLS and afterwards RIGLS, by varying the number of replicas in the sampling from 100 to 500 and to 1000 and the number of samplings from 5 to 8.

**Research limits and future directions of study**

The modelling of multilevel data is a complex process which appeals for a good understanding of groupped data notions and of the implications assumed by being part of a group. The single level analysis model does not take into account the data correlation between multilevel structures and the results are biased estimators, big standard errors, and consequently, incorrect tests and conclusions. The multilevel analysis is an approach that takes into consideration both the individuals and the groups to which they belong.

With these information in mind, our thesis aimed at studying the estimation procedures for the multilevel models, from both a theoretical and a practical perspective. By data analysis we intended to emphasize the characteristics of some estimation procedures for generalized multilevel models, namely the quasi-likelihood estimation procedures. The statistical analysis was performed using the R programming language, a free software, SPSS, commonly used in academia and MLwiN, a program dedicated to multilevel analysis.

From a theoretical perspective, a limit of the literature in the multilevel domain, is a framework characterized by the lack of definitions and uniform methodologies at the international level. This irregularity is a result of the recent development of the multilevel domain and of the increased new software technologies.

One of the limits of this research would be the relatively small number of analyzed countries, so a small number of analyzed groups due to reduced possibility of EU countries choice in the World Values Survey database. Concurrently, the WVS questionnaires on different periods have diverse type of questions and countries and cannot be correlated for a multi-annual analysis. Last, but not least, the lack of data regarding the importance of work makes impossible the comparison of our results with others from the multilevel domain. A methodological limit regards the bootstrapping sampling application: the accuracy of conclusions and the correction of bias depend on the number of re-samplings and the result would depend hence on the decision of the analyst to stop or to increase the number of re-samplings.

The future research directions are, as the author sees them, multiple and can be aimed at the extension and completion of the present research or at the comparison of this research to others in the area. A first future direction of research would be the application of the same estimation models and procedures on a two years database. Thus, we would have a multilevel model and we could test the results obtained on a yearly basis through a two level multilevel model, and also the results of a three level model.

Alternatively, the analysis of a longitudinal model or a multivariate multilevel model for data regrading the importance of work may be a development direction for complementing this research study.

As long as the reseacher knows about the existence of multilevel data, he can recognize them everywhere (Kreft, Jan De Leeuw, 1998: 1). So, we can allege, without a doubt, that applying the multilevel models is helpful and recommendable to be used in all areas of research that work with groupped data.