UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Social Sciences

Methods for Analysing Complex Panel Data Using Multilevel Models with an Application to the Brazilian Labour Force Survey

by

Alinne de Carvalho Veiga

Thesis for the degree of Doctor of Philosophy

June 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES
SCHOOL OF SOCIAL SCIENCES

Doctor of Philosophy

METHODS FOR ANALYSING COMPLEX PANEL DATA USING MULTILEVEL MODELS WITH AN APPLICATION TO THE BRAZILIAN LABOUR FORCE SURVEY

by Alinne de Carvalho Veiga

Data sets commonly used in the social sciences are often obtained by sample surveys with complex designs. These designs usually incorporate a multistage selection from a population with a natural hierarchical structure. In addition, these surveys can also be carried out in a repeated manner including a rotating panel design, which is a source of planned non-response. Unplanned non-response is also present in panel data in the form of panel attrition and intermittent non-response.

Methods are developed to handle this type of data complexity. These methods follow the Multilevel Model framework which is reviewed. Longitudinal growth curve models accounting for the complex data hierarchy are fitted. Recognizing the need to account for the complex correlation structure present in the data, multivariate multilevel models are then adopted. Alternative modified correlation structures accounting for the rotating sample design are proposed. Multivariate multilevel models are fitted utilizing these alternative correlation structures. In addition, models estimated using robust methods are compared with those estimated using standard methods.

A method for calculating a set of longitudinal sample weights that accounts for attrition is proposed. Models utilising the conditional sample weights and longitudinal weights are fitted using the Probability-weighted Iterative Generalized Least Squares (PWIGLS) estimation method. Furthermore, an extension to PWIGLS for multivariate multilevel models is developed. Models fitted through different estimation methods are compared and the best approaches are suggested.

Data from the Brazilian labour force survey, *Pesquisa Mensal de Emprego* (PME) are used. The PME has a complex sampling design that includes a multistage selection of the sample units and a rotating panel design characterised as 4-8-4. The methods developed are used to investigate the labour income dynamics of employed heads of households in the PME survey.

Contents

Appendices

List of Figures

List of Tables

Declaration of Authorship

I, Alinne de Carvalho Veiga, declare that the thesis entitled "Methods for Analysing Complex Panel Data Using Multilevel Models with an Application to the Brazilian Labour Force Survey" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: ....

Date:...

Acknowledgements

This doctoral project was supported by the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), an agency of the Brazilian Ministry of Education, under the grant number BEX 2392-04-6.

I am deeply thankful to my supervisor Professor Peter W F Smith and deeply indebted to my advisor Dr. James J Brown. Their encouragement, guidance and constant support were indispensable throughout the development of my research. I am grateful to the staff of the Division of Social Statistics and the S3RI for providing a superb and friendly academic environment. Special thanks are due to the Postgraduate Research Office staff, to Dr. Sandy Mackinnon and to David Baker from iSolutions for all their support.

Many researchers have assisted in this work, and my special acknowledgements go to Dr. Denise Silva, Dr. Marcel Vieira and Dr. Solange Correa for their encouragement and advice. I want to thank Professor Danny Pfefferman and Professor John Micklewright for their valuable advice and also my research colleagues Tom King, Erofili Grapsa, Corrado Giulietti and Guy Abel. I am also grateful to my previous co-workers Dimitri Szerman, Carlos Henrique Corseuil, Adriana Fontes, Danielle Carusi and Frederico Finan; and to Cimar Azeredo Pereira, Giuseppe Antonacci and Mauricio Lila for providing assistance with the PME data. Thanks is also due to Ana Simoes.

The completion of a PhD abroad is a challenging and sometimes lonely process which has only been made more enjoyable *with a little help from my friends*: Maíra, Sol, Denise, Katinha, Luana, Bia, Flavio, Wagner and Felipe (my dear Brazilians); Vasthi, Eri and Renato (friends of all times); Marco, Carlo, Salomé, Lynda, Lorenzo, Aisha, Bernie, Hannah, Gi and Ana. Thanks for all the good times, you made part of an important period of my life and will be always with me where I go. My sincere gratitude goes to you all.

Last but not least, I am most grateful to my family, who have supported me and followed attentively my journey. My sister Paula who showed to be a real good friend at all time. The ocean that separated us during these years made our friendship grow stronger, thank you. Acknowledgement is due to my auntie Beatriz for all her support and companionship to my mother in some difficult times throughout the time I have been away. I deeply thank my mother, Olga, for all her patience, support, care and unconditional love and also thank my father, Pedro Paulo, for all his inspiration. Mum and Dad, you are my rock. Thanks is also due to my lovely little niece, Raiane Maria, who has kept me entertained, even if from a distance, and filled my heart with joy. This thesis is dedicated to my dearly missed grandmother Rita.

In memory of Rita Rodrigues de Carvalho (October 1911 to January 2009).

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Panel Data

Data sets commonly used in the social sciences are often obtained by sample surveys with complex designs. These designs usually incorporate a multistage sample selection from a population with a natural hierarchical structure. Furthermore, when these designs include different waves of data collection, i.e. when the same sample is followed over time, a complex longitudinal data set, or panel data set, is generated.

Panel data is also called longitudinal data in the social sciences or repeated measurements data in the medical sciences. According to Firebaugh (1997), panel and longitudinal surveys follow households, families or individuals over time while repeated measurements surveys follow birth cohorts over time. In economics panel surveys may sometimes refer to repeated cross-sectional surveys that follow cohorts or groups of individuals over time. These data are pooled or matched to form a panel but only allow the investigation of aggregate change. This is not the type of panel data this thesis deals with.

Heckman and Singer (1985) defined typical panel data as those that contain a large number of units observed over a short period of time. The different times the units are observed characterizes the waves of the data or the measurement occasions. Plewis (1985) observed that in the social sciences panel surveys are usually taken on fixed measurement occasions with regular spaces between each of the measurements. However, panel data can also result from a rotating panel scheme. These schemes involve the stratification of the selected sample into panels that are rotated in and out of the survey during a specific period of time. This results in a planned unbalanced panel data set, as not all the units are observed at every time point. In addition, panel data can be prospective, collecting current data in repeated interviews, or retrospective, collecting data from past events (Menard, 2002).

There are a number of advantages of panel data when compared to cross-sectional data sets. As already mentioned, in a longitudinal survey, the units of analysis have their information collected for two or more periods of time. Usually the same individuals or cohorts are followed over time, as opposed to a cross-sectional survey which collects information for a single period of time. Therefore, the analysis of panel data allows for the direct study of change (Plewis, 1985). It allows for the investigation of the process of change that the individuals and groups go through. Change can be due to growth or development reasons, historical reasons or due to group membership. These different processes relate to three linearly dependent effects: age-effect, period-effect and the cohort-effect (Menard, 2002). Longitudinal surveys also allow for the study of the stability over time. Stability, as defined in Plewis (1985), is the positive correlation between the repeated outcome measurements within the same individual. A higher positive correlation would indicate a high level of individual stability.

Panel data can also be used in the investigation of cross-sectional effects. For example it can be used for understanding the gross flow between each time period, as well as the investigation of net flows over the time of the survey (Firebaugh, 1997). In this sense, a panel data analysis permits the investigation of the within and between individuals variability, which is not possible in a cross-sectional analysis. It can also make possible the identification of directions and magnitude of causal or temporal relationships. In the social sciences panel data have great applicability in the development and assessment of public policies (Burkhauser and Smeeding, 2000). Furthermore, panel data analysis allows a better assessment of important dynamic economic processes such as the study of spells of unemployment, job turnover and income mobility (Baltagi, 2005).

There are also disadvantages in the collection of panel data. These data sets require special tools of analysis, which for example deal with the dependence over time between the responses within the same unit of analysis. Furthermore, the complexity of the panel survey design might certainly be seen as an disadvantage. Due to such complex designs, the methods of analysis typically used for cross-sectional data are not applicable for longitudinal data sets. In addition, if the survey design also involves the selection of units through unequal probabilities of selection, clustering and stratification, methods applied in the analysis need to account for these characteristics.

The analysis of panel survey data can also be affected by recall (Bailar, 1989) and measurement errors (Duncan, 2000). The former type of error relates to how much the respondents can recall past events and declare them correctly. Duncan (2000) stated that measurement errors were more serious in longitudinal studies than in cross-sectional studies. Furthermore, the length of time that the units participate in the survey may also be a source of bias which might reflect behaviour change. Self-selection bias is also a potential source of error in panel surveys (Solon, 1989) which may result in units ceasing their participation on the survey. Non-response or panel attrition is a main concern. Non-response can be determined by the design, or unplanned. Individuals can be omitted from the survey in a specific period, or wave, or can prematurely drop out from the survey characterizing attrition in a panel data set. The different types of non-response cause the reduction of the sample size also generating an unbalanced data set. This also requires special methods of analysis. Some panel surveys provide ways to compensate for panel non-response, for example by applying ad hoc imputation techniques or calculating sets of longitudinal sampling weights that account for the different patterns of non-response (Lepkowski, 1989).

1.1.2 Panel Data Analysis

There are different approaches available in the statistical and econometrics literature for the analysis of panel data. In the statistical literature, the different methods are concerned with the potential imbalance of the data caused by planned or unplanned non-response, as well as with the potential correlation between the repeated individual responses. One approach that handles incomplete longitudinal data is the marginal model (Diggle et al., 2002). This is based upon the generalized estimation equation (GEE) methodology that was presented by Liang and Zeger (1986). The GEE is an alternative to the maximum likelihood estimation which takes into account the dependence within individuals by incorporating a working covariance matrix into the estimation process. This method provides asymptotically normal and consistent coefficient estimates. Furthermore, this method produces standard errors that are robust against misspecification of the working covariance matrix. The focus of the marginal models is on the estimation of the mean structure of the response. The dependence among responses within individuals is treated as a nuisance. Therefore, the coefficients estimated by using models of this type have a marginal or population-average interpretation. Marginal models do not require assumptions about the distribution of the residuals. However, the mean structure of the model must be properly specified.

Marginal models have a parallel in the analysis of panel data in the econometrics literature, in which the methods are more concerned with the problem of omitted individual variables and their relationship with the observed variables rather than with the correlation between the repeated responses. Under the econometric framework fixed effects models (Hsiao, 2003) have a similar formulation to marginal models. Fixed effects models assume that the unobserved individual effects are fixed over time. These models are often estimated following the dummy-variable approach (Wooldridge, 2002), which accounts for as many dummies as the number of individuals in the sample minus one ($n-1$), capturing the time invariant effect. No assumptions are made on the distribution of these unobserved effects. Therefore, like the marginal models, inference can only be drawn conditionally on the effects that are in the sample (Hsiao, 2003). It is also worth mentioning that in a fixed effects model, only time varying covariates can be considered and, unlike the marginal models, the covariance structure is not considered in the estimation process. Hence, the repeated measures are considered conditionally independent and identically distributed, given the fixed effects. Fixed effects models are often estimated via ordinary least squares (Baltagi, 2005).

The random effects models under the econometrics framework have a similar formulation to the fixed effects model. However, in the random effects model, the unobserved individual effects are assumed to be uncorrelated with the observed variables. In this model, the unobserved individual effects are thought to be random draws from a population of individual effects. Therefore, inference about the population that generated the effects is permitted. For longitudinal data with a large number of time points, or occasions, random effects and fixed effects models should generate the same results (Hsiao, 2003). When the number of occasions is small the choice between either type of models should be based on the assumptions about the unobserved effects, as ignoring the potential endogeneity leads to biased estimators. However, due to their underlying modelling assumptions, the random effects model seems a more natural choice for panel survey data than the fixed effects models (Skinner, 2003).

Methods for the analysis of longitudinal data as discussed so far usually consider only a time and an individual level effect. However, panel survey data originated from a multistage sampling design, which involves the clustering of basic units in higher level units, present a hierarchical structure. Each level of the data structure might be of research interest but also represents a potential source of random data variability. Therefore, the data hierarchy needs to be considered in the analysis. The multilevel modelling framework (Goldstein, 2003) is the standard choice for clustered data of this type (Steele, 2008). Multilevel models are random effects models that account for the clustering originated from multistage sampling design, also allowing for the investigation of the effects of the data structure on the outcome variable (Plewis and Fielding, 2003). These models rely on the same strong assumptions about the conditional distribution of the random effects and the observed variables as the models in the econometrics literature. Despite this, multilevel models are perhaps the most appropriate for multistage survey data.

The multilevel modelling framework encompasses the random intercept model, the random slope model, or growth curve model, and the multivariate multilevel model. It handles unbalanced and unequally spaced panel data, provided that the non-response mechanism is at random (Little and Rubin, 2002). In addition, this framework allows for the inclusion of variables at the different levels of the data hierarchy. Therefore, it allows for the inclusion of time varying and time constant covariates, as well as covariates for the higher level units. Under this framework the repeated measurements are the level one units nested within individuals, the level two units, which are nested within higher level units. The inclusion of variables with coefficients varying randomly across the units relaxes the underlying assumption that the repeated measures are exchangeable within individuals (Goldstein, 2003). This characterizes the growth curve model, which models the panel data as dependent of some measurement of time (Steele, 2008). Furthermore, time, which is usually taken as a continuous variable in these types of models, can be considered as having a linear or polynomial effect on the outcome. A further extension of the random effects model under the multilevel model framework is to consider the repeated outcomes as a multivariate outcome. Multivariate multilevel models allow the modelling of the correlation structure between the individual responses usually treating time as a discrete variable. Models under this framework are usually estimated via maximum likelihood or iterative generalized least squares.

Multilevel models, as described above, do not incorporate the sampling design in contrast to those described in Pfeffermann et al. (1998). The method presented by Pfeffermann et al. (1998) accounts for the unequal selection probabilities in a two-level random coefficients model. Their proposed method modifies the iterative generalized least squares estimation procedure to account for the sampling weights, whence called probability-weighted iterative generalized least squares. Other authors proposed similar methods for cross-sectional multilevel models. However, having the multivariate model as basis, Folsom (1989) and Skinner and Holmes (2003) proposed methods to account for the sampling weights in a longitudinal data analysis. These methods are not yet vastly explored in the research literature and are still under debate.

1.1.3 Labour Force Surveys

The majority of the Labour Force Surveys in the world include some kind of rotating sample design (McLaren and Steel, 2000). These rotating designs can have a consecutive pattern of the form *in-for-d* or a non-consecutive pattern of the form *a-b-a(d)* as defined in McLaren and Steel (2000). When a non-consecutive pattern is adopted, the selected sample units are in the survey for a consecutive occasions. There will be a gap of b between the middle occasions for a total span of $d = a + b + a$ and $T = a + a$ time points. These non-consecutive rotating designs are usually symmetric (Mehran, 2007). However, this does not need to be the case, and instead these designs could follow an pattern such as *a-b-c(d)*. Examples of often used rotation patterns are the 2-2-2 (Italy and Israel for example), *in-for-6* (Canada, Spain and Portugal), *in-for-5* (the UK), *in-for-8* (Australia) and 4-8-4 (U.S. and Brazil) (Mehran, 2007).

The Brazilian labour force survey whose official[1] name translates to Monthly Employment Survey is a probabilistic household sample survey conducted by the Brazilian Institute for Geography and Statistics[2]. It is conducted every month with the main objective of investigating the characteristics of the Brazilian labour force. This survey has a complex multistage sample scheme characterised as a stratified two-stage cluster design with approximately equal probabilities of selection at the household level within each of the six metropolitan areas covered by the survey. In addition it has a non-consecutive but symmetric sample rotating panel design characterized as 4-8-4. This means that the selected sample units stay in the sample for four consecutive months, are out for eight months and return to the sample for other four consecutive months. Therefore, due to this rotation pattern there is a gap of eight months between the fourth and fifth interview for every unit in this survey.

[1] The official name of the survey is in Portuguese *Pesquisa Mensal de Emprego* usually abbreviated as PME. A list of abbreviations and notation is provided in the glossary at the end of this thesis.

[2] Instituto Brasileiro de Geografia e Estatítica - IBGE.

1.2 Aims and Outline of the Thesis

Motivated by the different data complexities, this thesis aims to further develop methods for analysing complex panel data. These methods follow the multilevel modelling framework. Recognizing that the models under this framework are based upon strong modelling assumptions, this thesis aims to demonstrate how some of these assumptions can be relaxed to better incorporate the data complexities in a single modelling exercise. The main data complexities considered are: *(i)* the hierarchical structure of the data set, *(ii)* the complex correlation structure between the repeated outcomes, *(iii)* the imbalance of the panel data due to panel non-response and *(iv)* the incorporation of the rotating design and the sampling weights into the analysis.

Following this introduction, this thesis contains seven more chapters. Chapter 2 presents a methodological review of the analysis of longitudinal data under the multilevel modelling framework. This chapter starts by presenting a general multilevel model for cross-sectional data which is extended to accommodate longitudinal data and further extended to the multivariate multilevel model. This chapter also presents: a review of the main covariance structures that can be modelled under the multivariate multilevel model formulation; a review of the robust methods for the estimation of the standard errors of the estimated multilevel model coefficients and a review of the different types of panel non-response mechanisms. The chapter finishes with a review of the methods to account for panel non-responses with special attention given to methods which calculate longitudinal sampling weights.

Data from the Brazilian labour force survey are used in this thesis to demonstrate the different applications of the methods developed. This survey and its main design features are described in Chapter 3. The design of the Brazilian labour force survey is quite complex and encompasses the complexities aimed to be accounted for in this thesis. One difficulty found relates to the linkage of the data across time for this survey and this is also discussed in Chapter 3.

The methods developed in this thesis are used to investigate the labour income dynamics of employed heads of household in the Brazilian labour force survey. Chapter 4, therefore, presents a brief review of the Brazilian economy and labour market. Special attention is given to the review of models for the determinants of the labour income. Based on this review, Chapter 4 also presents an initial cross-sectional model for labour income determinants. This model serves as the basis for the other applications in this thesis.

Longitudinal growth curve models accounting for the complex data hierarchy are fitted in Chapter 5 assuming that the process of change varies between heads of households. Therefore, the assumption of exchangeability between residuals within the same individual is relaxed. These models are further extended to multivariate multilevel models which account for the complex correlation structure of the data. Therefore, the assumption of uncorrelated residuals within individuals is also relaxed. Alternative modified correlation structures that account for the rotating panel design are presented. Multivariate multilevel models are then fitted utilizing the alternative structures, which are compared and the best formulation is identified.

Chapter 6 presents a detailed description of the method proposed in Pfeffermann et al. (1998), the probability-weighted iterative generalized least squares estimation (PWIGLS) method. This chapter also presents an extension of the PWIGLS method for the fit of multivariate multilevel models. This developed framework allows the fitting of multivariate models imposing the alternative correlation structures presented in Chapter 5.

Chapter 7 presents a method for calculating a set of multilevel longitudinal weights that accounts for the panel attrition patterns of the Brazilian labour force survey. Models utilising the developed set of multilevel longitudinal sampling weights are fitted through PWIGLS. Furthermore, Chapter 7 also presents a comparison between the models fitted through PWIGLS and an equivalent model fitted using IGLS to a longitudinal data set.

Chapter 8 presents the key conclusions of this thesis. This chapter also presents a summary of the research contributions and identified pieces of further research.

Chapter 2

Multilevel Models for the Analysis of Longitudinal Data

2.1 The Multilevel Modelling Framework

Longitudinal data in the social sciences often originate from surveys with a complex multistage sampling scheme. The different stages of the sampling scheme can represent the different levels of the hierarchical structure of the data which usually represents the natural grouping in the population. Most of the time, the selection of the units of analysis in these schemes involve some kind of cluster sampling techniques. The selected clusters are thought to be a random sample of the population of clusters. However, clustered data are not expected to be independent (Kish and Frankel, 1974). In other words, units selected using clustered multistage designs are not independent as the selection of the secondary sampling units is conditioned on the selection of the primary sampling units. Furthermore, individuals within the same cluster may share similar characteristics and behaviours and some of these characteristics might not even be observed by the researcher. Such data complexity invalidates the use of standard estimation methods for regression models once the observations are no longer assumed to be conditionally independent and identically distributed (IID). These methods are based on the ordinary least squares (OLS) estimation and their use for the analysis of multilevel data still generate unbiased estimates. However, these estimates would be inefficient and with standard errors biased downwards (Maas and Hox, 2004). In addition, the levels of a hierarchical data set may be of research interest themselves. These groupings might contain potential random influences on the units of analysis representing different sources of data variability (Snijders and Bosker, 1999; Goldstein, 2003). For such data, Goldstein et al. (1994) stated that the use of methods within the multilevel model framework is the appropriate approach to follow. To account[1] for the clusters and the hierarchical structure of data of this type yields statistically efficient estimates of the regression coefficients with correct estimates of the standard errors also producing more "conservative" confidence intervals and test statistics (Goldstein, 2003). The multilevel approach also allows for the investigation of particular clusters of interest, comparison between clusters and inference on the level of variability between them.

Multilevel models were developed from the 1980s with the work of Aitkin et al. (1981), as mentioned in Goldstein (2003) and Longford (1993). They are an extension of the classical multiple regression model still keeping the linearity and normality assumptions but relaxing the independence (due to clusters) and the constant variance (due to random effects of covariates) assumptions. Before the development of multilevel models, data with some hierarchical structure were analysed either using aggregating or disaggregating techniques. Both techniques ignore the multilevel structure. The former used individual data aggregated to cluster averages and proportions. This is appropriate to make inferences at the cluster level but not at the individual level avoiding the ecological fallacy, that is when cluster level effects are generalized to the individuals (Luke, 2004). The latter used cluster data disaggregated to the individual level. This inflates the sample size and inference is made at the individual level. Consequently, this leads to the incorrect calculation of standard errors and incorrect conclusions. Another usual method to analyse clustered data was to fit separate regression models, one for each cluster. Goldstein (2003) stated that for data where the total number of clusters is small with a relatively large number of units within cluster, this procedure may produce efficient results. However, this procedure cannot be used if the variation between clusters needs to be investigated. Moreover, it could not be applied in longitudinal analysis as longitudinal data are typically formed by a large number of individuals with relatively few measurements.

Longitudinal data can be analysed in a multilevel modelling framework. For example, individuals on a longitudinal data set represent the level two units and their repeated measurements the level one units. In a more complex context, where individuals are selected through a more complex sampling design, the highest level of the analysis represent the clusters within which the individuals are nested, such as households, the middle level the individuals and the bottom level their repeated

[1] Accounting for clusters here does not mean that the sampling design is being accounted for as recommended in Pfeffermann et al. (1998).

measurements. In this sense, longitudinal multilevel models are an extension of cross-sectional multilevel models, which is the starting point of the next subsection. It is worth mentioning that in the statistical literature multilevel models are also called: Hierarchical linear models (Bryk and Raudenbush, 1992); Random coefficient models (Longford, 1993), Mixed effects models (as mentioned in Snijders and Bosker (1999)) among others.

2.1.1 Cross-sectional Multilevel Modelling

Consider the following model representation:

$$y_{ij} = \boldsymbol{x}^T_{(1)ij}\boldsymbol{\beta}_{(1)} + \boldsymbol{x}^T_{(2)j}\boldsymbol{\beta}_{(2)} + u_j + e_{ij} \; . \tag{2.1}$$

This is a two-level random intercept model. In a cross-sectional data set, the level two units are the n clusters, assumed to be independent and represented in the model by the subscript j ($j = 1, 2, \ldots, n$). The level one units are the n_j individuals, represented in the model by the subscript i ($i = 1, 2, \ldots, n_j$), that are nested within cluster j. The total number of individuals in this cross-sectional data is $m = \sum_j n_j$.

The outcome variable, y_{ij} in model 2.1, is a continuous variable at the individual level, hence, the pair of subscripts ij. The idea of the multilevel approach is to model the effects of individual and cluster characteristics on the outcome variable in a single model representation, as in equation 2.1. The individual characteristics are represented by the vector of explanatory variables at the individual level $\boldsymbol{x}_{(1)ij}$. For a single cluster j the matrix of $p_{(1)}$ explanatory variables is represented as:

$$X_{(1)j} = \begin{pmatrix} 1 & x_{1j2} & \cdots & x_{1jp_{(1)}} \\ 1 & x_{2j2} & \cdots & x_{2jp_{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_j j2} & \cdots & x_{n_j jp_{(1)}} \end{pmatrix} ,$$

where $\boldsymbol{x}_{(1)ij1}$ is a vector of ones for the intercept. Associated to this matrix is the vector of fixed regression coefficients at the individual level

$$\boldsymbol{\beta}^T_{(1)} = \left(\beta_{(1)1},\ \beta_{(1)2},\ \ldots,\ \beta_{(1)p_{(1)}}\right) .$$

Model 2.1 also includes the vector of fixed regression coefficients at the cluster level

$$\boldsymbol{\beta}_{(2)}^T = \left(\beta_{(2)1},\ \beta_{(2)2},\ \dots\ , \beta_{(2)p_{(2)}}\right),$$

associated with the vector of explanatory variables $\boldsymbol{x}_{(2)j}$. The explanatory variables at the cluster level are also called contextual variables. For a single cluster j the vector of $p_{(2)}$ contextual variables is represented as

$$\boldsymbol{x}_{(2)j}^T = \left(x_{(2)j1},\ x_{(2)j2},\ \dots\ , x_{(2)jp_{(2)}}\right).$$

Additionally, consider $\boldsymbol{x}_{ij}$ as being the vector of all $p = p_{(1)} + p_{(2)}$ explanatory variables for both of the levels and $\boldsymbol{\beta}$ the vector of all p fixed regression coefficients. This is the fixed part of the model 2.1.

The model 2.1 contains two error terms: u_j, at the cluster level, and e_{ij}, at the individual level. These form the random part of the model. The presence of more than one error term is what makes the multilevel model different from the classical regression model. Multilevel models have at least one error term for each level of analysis being considered (Snijders and Bosker, 1999). The u_j in model 2.1 are also called cluster specific effects or cluster residuals and here it is assumed that

$$u_j{\sim}N(0, \sigma_u^2). \tag{2.2}$$

Under the formulation of model 2.1 and conditioning on the set of explanatory variables $\boldsymbol{x}_{ij}$, the random intercepts u_j are assumed to be uncorrelated with the covariates. The u_j are also assumed to be independent from the individual level residuals e_{ij}. These are equivalent to the error terms of the classical linear regression, also called the raw residuals. Here it is assumed that conditioned on the set of explanatory variables $\boldsymbol{x}_{ij}$, the e_{ij} are mutually independent and that

$$e_{ij}{\sim}N(0, \sigma_e^2). \tag{2.3}$$

The conditional distribution assumed implies that e_{ij} and the total set of all explanatory variables $\boldsymbol{x}_{ij}$ are uncorrelated.

The mutual independence assumption of the e_{ij} implies the absence of correlation between level one residuals of two individuals within the same cluster. Therefore, given the random intercepts u_j and the covariates $\boldsymbol{x}_{ij}$, these two individuals are assumed to be independent (Skrondal and Rabe-Hesketh, 2004). The implicit assumptions can be summarized as

$$Cov\left(u_j, e_{ij}\right) = 0,$$

$$Cov\left(e_{ij}, e_{i'j}\right) = 0, \qquad \forall i \neq i^{'}$$

$$Cov\left(y_{ij}, y_{i'j} | u_j, \boldsymbol{x}_{ij}, \boldsymbol{x}_{i'j}\right) = 0, \qquad \forall i \neq i^{'}.$$

The variance components in 2.2 and 2.3, σ_u^2 and σ_e^2 respectively, represent the between and the within cluster variabilities that are simultaneously modelled in 2.1. The total variability of the observations under the model can be calculated from the variance of the composite residuals

$$r_{ij} = u_j + e_{ij}.$$

Under the already stated assumptions, it is easy to see that the variance of the composite residuals r_{ij} for the random intercept model in 2.1 is:

$$Var(r_{ij}) = Var(u_j + e_{ij}) = Var(u_j) + Var(e_{ij}) = \sigma_u^2 + \sigma_e^2.$$

As already mentioned, individuals within the same cluster are not expected to have independent observations. In a random intercept model, this dependency is induced by the u_j as they are shared between individuals in the same cluster as explained in Skrondal and Rabe-Hesketh (2004, page 51). The level of dependence can be measured by the intra-cluster correlation given by the conditional correlation between units within the same cluster as:

$$Cor(y_{ij}, y_{i'j} | \boldsymbol{x}_{ij}, \boldsymbol{x}_{i'j}) = Cor(u_j + e_{ij}, u_j + e_{i'j}) \qquad \forall i \neq i^{'} \tag{2.4}$$

$$= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = \rho. \tag{2.5}$$

This intra-cluster correlation can also be interpreted as the amount of the total residual variability that is due to between clusters variability or as how similar the individuals within the same cluster are (Snijders and Bosker, 1999). Therefore it is a measure of the dependence that exists in the hierarchical data. Singer and Willett (2003) advised starting the model building process by fitting an empty random intercept model and calculating ρ. This gives an estimate of the average correlation between randomly selected pairs of residuals in the same cluster $(r_{ij}, r_{i'j})$. For values of ρ close to zero, some authors (Snijders and Bosker, 1999; Bryk and Raudenbush, 1992) suggested that the random intercept model could be simplified, leading to the classical regression model without the common random intercepts u_j. However, large values of ρ justify the use of the multilevel modelling framework and the model building process should then continue.

Now consider an extension of the random intercept model in equation 2.1 to the following model representation:

$$y_{ij} = \boldsymbol{x}_{(1)ij}^T \boldsymbol{\beta}_{(1)} + \boldsymbol{x}_{(2)j}^T \boldsymbol{\beta}_{(2)} + \boldsymbol{z}_{ij}^T \boldsymbol{u}_j + e_{ij} \ . \tag{2.6}$$

This is a two-level random coefficients model. The difference between the model in equation 2.1 and the model in equation 2.6 is that the latter includes the vector with a sub-set of explanatory variables $\boldsymbol{z}_{ij}$ for which the coefficients are considered as random at the cluster level. The vector $\boldsymbol{z}_{ij}$ is associated with the vector of random effects at the cluster level $\boldsymbol{u}_j$ that now includes the random intercepts and the random slopes. The random slopes are thought as interaction terms between the random intercepts and the explanatory variables whose effects are considered to vary among the clusters. The same assumptions adopted for the formulation of model 2.1 are also adopted for model 2.6. However, due to the presence of random slopes, the model in equation 2.6 now has more than two error terms. There is one error term at the individual level (e_{ij}), one error term for the intercept (u_{0j}) and additional error terms for the random slopes (u_{kj}). For simplicity of illustration, consider that the design matrix Z, for a cluster j, is of the form

$$Z_j = \begin{pmatrix} 1 & z_{1j} \\ 1 & z_{2j} \\ \vdots & \vdots \\ 1 & z_{n_j j} \end{pmatrix} . \tag{2.7}$$

Therefore, it contains the vector of ones for the random intercepts and one explanatory variable. In this case, the vector of cluster random effects is represented by

$$\boldsymbol{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} .$$

This vector is then assumed here to have a bivariate normal distribution with mean vector zero and covariance matrix Σ_u, as in

$$\boldsymbol{u}_j \overset{iid}{\sim} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} , \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right) .$$

The element σ_{u01}, in the covariance matrix Σ_u, represents the covariance between the random intercepts and the random slopes. It shows the relationship between u_{0j} and u_{1j} giving an idea of whether high intercepts are associated with higher slopes, for example (Rabe-Hesketh and Skrondal, 2005).

The model in equation 2.6 indicates that there is heterogeneity among clusters (Snijders and Bosker, 1999). It retains the assumption that individuals in the same cluster are correlated, but this correlation now depends on the vector $\boldsymbol{z}_{ij}$. The conditional variance of the observations within the same cluster will also depend on $\boldsymbol{z}_{ij}$ as can be seen below. The vector of composite residuals for the cluster j under the random coefficients model in equation 2.6 is given as:

$$\boldsymbol{r}_j = Z_j \boldsymbol{u}_j + \boldsymbol{e}_j,$$

where $\boldsymbol{e}_j$ is the vector of level one raw residuals in the cluster j.

The covariance matrix V of the total composite residuals $\boldsymbol{r}$ is a block-diagonal matrix. Each block in V represents one of the n clusters and is defined as

$$V_j = Z_j \Sigma_u Z_j^T + I_{n_j} \sigma_e^2 \ , \tag{2.8}$$

where I_{n_j} is a $n_j \times n_j$ identity matrix. The matrix V_j for a generic cluster j of size n_j can then be written as:

$$V_j = \begin{pmatrix} \sigma^2_{r_1 j} & & & \\ \sigma_{r_{21 j}} & \sigma^2_{r_2 j} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r_{n_j} 1 j} & \sigma_{r_{n_j} 2 j} & \dots & \sigma^2_{r_{n_j} j} \end{pmatrix}.$$

The random coefficient model 2.6 also assumes that the vector of composite residuals $\boldsymbol{r}$ is $N(0, V)$. For the case where the design matrix Z_j is as in equation 2.7 each component of the diagonal of V_j, the variances, can be written as

$$\sigma^2_{u0} + 2\sigma_{u01} z_{ij} + \sigma^2_{u1} z^2_{ij} + \sigma^2_e,$$

and the off-diagonal terms, the covariances, as

$$\sigma^2_{u0} + \sigma_{u01}(z_{ij} + z_{i'j}) + \sigma^2_{u1} z_{ij} z_{i'j} \qquad \forall i \neq i^{'}.$$

This shows that, under the random coefficients model, the assumption of homoscedasticity of the residuals is relaxed. The variance has a quadratic relationship with $\boldsymbol{z}_{ij}$. However, when the components of Σ_u are very small and close to zero the total variance is close to the constant variance component at the individual level, σ_e^2 (Singer and Willett, 2003). One other important point to observe is that under the model representation 2.6 the intra-cluster correlation coefficient can no longer be calculated using the formula in equation 2.5 as it also depends on $\boldsymbol{z}_{ij}$. However, ρ can be calculated for specific pairs of units as:

$$\rho_{ii'j} = \frac{\sigma_{r_{ii'j}}}{\sqrt{\sigma^2_{r_{ij}}\sigma^2_{r_{i'j}}}} \qquad \forall i \neq i'. \tag{2.9}$$

Because the parameters of the random part, or the variance components, of the model in equation 2.6 are often of great interest in a multilevel analysis, Goldstein (2003) called the model in 2.6 the variance components model. It is worth mentioning that this model, like the model in equation 2.1, also has a set of contextual variables. The inclusion of such variables is of interest because in multilevel data it is believed that the outcome can be influenced by the context as well (Luke, 2004). These variables are cluster level covariates and can express cluster level information, as well as cluster averages or proportions of level one variables. They do not need to be produced from the same data set that provides the level one variables. They can be provided from external sources. Snijders and Bosker (1999) advised for the inclusion of cluster variables when the random coefficients show dependency with the level one variables. In addition, the inclusion of contextual variables can explain some of the level two unexplained variability.

2.1.2 Estimation Methods

The use of standard methods such as the OLS to estimate the model in equation 2.6 is no longer appropriate due to the correlated cluster data and the presence of more than one residual term. Goldstein (2003) presented some of the alternative methods that are appropriate for the estimation of multilevel models. Other authors such as Longford (1993); Skrondal and Rabe-Hesketh (2004); Singer and Willett (2003) and Snijders and Bosker (1999) also presented some of the same methods and this sub-section aims to briefly describe them (for a full explanation refer to these authors).

Longford (1993) stated that the methods utilized in the estimation of multilevel models are based upon **maximum likelihood** (ML) estimation. Singer and Willett (2003) stated that the ML estimation is an appealing method because of its large sample properties. The ML estimates are asymptotically consistent, normal and efficient. However, these properties are unlikely to hold in relatively small samples. It is also worth mentioning that the ML estimation is based on the assumption that the residuals are normally distributed.

The ML estimators are those that maximize the sample likelihood, which is the joint probability function of observing the specific sample of the data. The sample likelihood contains all the unknown parameters, both of the fixed ($\boldsymbol{\beta}$) and random (variance components) parts for the model in equation 2.6. The mean is determined by the fixed part of the model and the variance is determined by the random part (Singer and Willett, 2003). The method uses the whole sample and in the case of multilevel data each cluster j contributes with n_j terms to the sample likelihood. However, ML estimation methods need numerical integration algorithms in order to find the values of the parameters which maximize the log-likelihood function. Some algorithms such as the EM, Newton-Raphson and Fisher scoring are used and are described in detail in Longford (1993) and Skrondal and Rabe-Hesketh (2004). These algorithms need starting values that are usually the OLS estimates. One drawback of the ML estimation is that it does not take into account the uncertainty of estimating the fixed effects when estimating the parameters of the random effects. This represents an important drawback of the ML method (Singer and Willett, 2003). In other words, the ML estimation does not take into account the loss of degrees of freedom due to the estimation of the fixed parameters. Hence, the estimates of the parameters in the random part of the model are biased downwards.

Restricted maximum likelihood (REML) estimation is an alternative to the ML for the estimation of the variance components. REML, which is based on the residual likelihood, estimates the variance components of a model such as the one in equation 2.6 accounting for the loss of degrees of freedom due to the estimation of the parameters in the fixed part of the model (Singer and Willett, 2003; Snijders and Bosker, 1999). However, REML is not a complete substitute for ML. It is more sensitive to outliers than ML (Skrondal and Rabe-Hesketh, 2004) and for unbalanced data, usually the case when dealing with longitudinal data, REML can also generate biased estimators. Snijders and Bosker (1999) added that for a large enough sample, with large number of clusters, not much difference will be expected between ML and REML estimators. It is also worth mentioning that when assessing the goodness-of-fit via likelihood ratio tests, when REML is the estimation method adopted, only the random part of the model can be tested. This test is described in the next sub-section.

Iterative generalized least squares (IGLS) (Goldstein, 1986) is another of the algorithms used in the estimation of multilevel models. IGLS is based upon the **generalized least squares** (GLS) estimation method that minimizes a weighted function of the residuals (Goldstein, 2003). GLS is a more flexible estimation method because it does not necessarily need the residuals to be normally distributed and it accommodates heteroscedastic or autocorrelated residuals (Skrondal and Rabe-Hesketh, 2004). However, if the assumption of the normal distribution of the residuals holds the IGLS algorithm yields ML estimates (Goldstein, 1986). Some of the authors already mentioned in this section such as Goldstein (2003), Longford (1993) and Skrondal and Rabe-Hesketh (2004) presented the IGLS method. It iteratively estimates the random effects and the fixed effects until their convergence. (See Goldstein (1986) for further details.) The IGLS method is described in detail in the sub-section 6.2.1 following the notation presented in Pfeffermann et al. (1998). It is worth mentioning that like ML estimation, IGLS produces biased estimates for the random part of the model especially in small samples (Goldstein, 2003). An alternative method which yields estimates equivalent to REML is **restricted iterative generalized least squares** (RIGLS). One more point to be considered is that, because IGLS and RIGLS are iterative algorithms, they may not converge easily for very small data sets or highly unbalanced data sets (Singer and Willett, 2003).

Goldstein (2003) also presented some information on Bayesian multilevel model estimation based on Markov Chain Monte Carlo (MCMC) methods such as the Gibbs Sampler and the Metropolis Hastings algorithm, but details of these methods are beyond the scope of this thesis.

2.1.3 Model Selection, Checking and Interpretation

This subsection presents some suggested steps to follow when selecting a multilevel model. This process involves the selection of significant fixed effects and significant random effects. There are no rules of thumb for the selection of a multilevel model, but some authors (Bryk and Raudenbush, 1992; Snijders and Bosker, 1999; Hox, 2000) provided some suggestions.

A good starting point is an empty random intercept model. This allows for the investigation of the amount of variability explained in each of the levels and the estimation of the intra-cluster correlation ρ (Snijders and Bosker, 1999). The next step would be to include the set of level one explanatory variables. These could include main effects and level one interaction terms, and when selecting the significant interaction terms the hierarchical principle should be employed, i.e.

the main effects of the interaction terms should also be retained in the model. In other words, the main effects of the significant interaction terms should also be kept in the model. After a level one model is selected two alternative steps could be followed. Either the inclusion of level two variables or the inclusion of random slopes could be addressed. Snijders and Bosker (1999) suggested that a good practice is to perform these steps separately and that the significant effects of each of the two steps could afterwards be tested together in the final model.

Snijders and Bosker (1999) recognized the difficulty in testing for significant random slopes. For this reason, it is advised that these should be tested only for covariates that show a strong fixed effect or for those that are substantively expected to vary between clusters. Random slopes should not be tested for level one interactions and they should be tested one at time. Care must be taken when testing for the inclusion of random slopes as their variances are likely to be close to zero. It should be kept in mind that the omission of an important random effect will impact on the hypothesis testing of the fixed part of the model and that, because the estimation methods involve numerical iteration, the inclusion of many random slopes may lead to convergence problems. If a random slope is found to be significant it means that there are still unexplained cluster differences. Contextual variables can then be included in the model in order to try to explain more of the unexplained cluster variability. In addition, the inclusion of level two variables is highly advisable when the random intercept is thought to be correlated with some of the covariates. It is also advisable to include in the model cross-level interactions between the variable with the random slope and the level two variables when the random slope seems to be correlated with a level two variable (Snijders and Bosker, 1999).

Model selection is a dynamic process that should involve both theoretical and empirical considerations. Overall, Snijders and Bosker (1999) advised refraining from including non-significant effects. It is worth reinforcing that different selection procedures can result in different selected models. As in the classical linear regression case the objective of the multilevel model selection is still to find the most parsimonious model that best represents the relationship between outcome variables and explanatory variables. Different tests of hypotheses can be used to assist in the model selection. These tests are usually applicable in the comparison of nested models from the same sample of data, and are listed below.

The **Wald test** is the general single parameter test that can be employed to test whether a fixed effect, say β_k, is significantly different from zero or not. It tests the hypotheses:

$$H_0 : \beta_k = 0$$

$$H_1 : \beta_k \neq 0 \ .$$

Under the null hypothesis, the Wald test statistic is:

$$T_{Wald}(\hat{\beta}_k) = \left(\frac{\hat{\beta}_k}{S.E.(\hat{\beta}_k)} \right)^2 ,$$

where $S.E.(\hat{\beta}_k)$ is the standard error of the estimate of the fixed effect $\hat{\beta}_k$ being tested. For large samples and under the null hypothesis, $T_{Wald}(\hat{\beta}_k) \sim \chi^2_1$ and the test procedure leads to the rejection of H_0 if $T_{Wald}(\hat{\beta}_k) > \chi^2_{1(1-\alpha)}$ for level significance α.

When multiple parameters need to be tested simultaneously, like in the case of a categorical covariate with several categories, the **multivariate Wald test** can be used. The hypotheses for the multivariate Wald test are:

$$H_0 : C\boldsymbol{\beta} = \mathbf{0}$$

$$H_1 : C\boldsymbol{\beta} \neq \mathbf{0} \ ,$$

where C is a matrix of linear combinations, or the contrast matrix. Each row of C is formed of sequences of $1's$ or $0's$, where 1 is relatively positioned to the parameters being tested from the vector of regression parameters $\boldsymbol{\beta}$. Rewriting $C\boldsymbol{\beta}$ as $\boldsymbol{\beta}_*$, representing a sub-vector of $\boldsymbol{\beta}$ the hypotheses for the multivariate Wald tests can now be written as:

$$H_0 : \boldsymbol{\beta}_* = \mathbf{0}$$

$$H_1 : \boldsymbol{\beta}_* \neq \mathbf{0} \ .$$

Under the null hypothesis, the multivariate Wald test statistic is

$$T_{Wald}(\hat{\boldsymbol{\beta}}_*) = \hat{\boldsymbol{\beta}}_*^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}_*} \hat{\boldsymbol{\beta}}_*,$$

where $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_*}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_*$, and DF is the number of rows of C, therefore the number of parameters being tested. The null hypothesis is rejected for large values of $T_{Wald}(\hat{\boldsymbol{\beta}}_*)$, i.e. when $T_{Wald}(\hat{\boldsymbol{\beta}}_*) > \chi^2_{DF(1-\alpha)}$.

It is worth mentioning that the multivariate Wald test is applicable to test fixed effects only. For testing multiple parameters including random effects an alternative is to use the **likelihood ratio test** (LRT). The LRT compares the log-likelihoods (l) of two nested models, a reduced model M_{red} and a model with the parameters being tested M_{full}. The hypotheses being tested for the LRT are:

$$H_0 : M_{\text{red}}$$

$$H_1 : M_{\text{full}} \ .$$

Under the null hypothesis, the test statistic of the LRT is

$$L^2 = -2 \times (l_{\text{red}} - l_{\text{full}}) \ , \tag{2.10}$$

where DF is determined by the difference between the number of parameters in the full and in the reduced model. The reduced model is rejected for large values of the likelihood-ratio test statistic L^2 (i.e. $L^2 > \chi^2_{DF(1-\alpha)}$).

For the random part, however, the applicability of the LRT is questionable. This is because the LRT tests whether the variances equal zero which is a value on the boundary of the parameter space $[0, \infty)$ for the variances. Therefore, this test will tend to accept the null hypothesis "*more often than it should*" as stated in Frees (2004, chapter 5). Snijders and Bosker (1999) still advocate the use of the LRT to test random effects bearing in mind that the test is, however, a one-sided test. For example, the hypotheses for testing the variance of the random intercepts are:

$$H_0 : \sigma^2_{u0} = 0$$

$$H_1 : \sigma^2_{u0} > 0 \ .$$

Therefore, the test could still be used and the p-value calculated based on the χ^2_1 distribution should be divided by two. This is because the test statistic is no longer χ^2_1 but a mixture of 0 and χ^2_1 distribution (Snijders and Bosker, 1999). Care must be taken, however, in applying the LRT to models estimated via REML. As mentioned before, REML is used to estimate the random part of the model, and the deviance from REML estimation describes only the random part of the model (Singer and Willett, 2003). In the case where the reduced model and the full model were both estimated via REML and the fixed part of each of the models are exactly the same, the LRT could be applied to test the extra random effects in the full model. Once again, the reduced model must be nested within the full model.

If two non-nested models need to be compared, two alternative goodness-of-fit criteria can be used instead of the LRT. Both criteria use the likelihood-ratio statistic of the fitted models only differing by a scale factor:

$$IC = L^2 + 2 \times (\text{Scale factor})(\text{Number of parameters in the model}) .$$

The **Akaike information criterion** (AIC) has a scale factor equal to one, and the **Bayesian information criterion** (BIC) has the scale factor equal to half of the log of the sample size (Singer and Willett, 2003), so:

$$AIC = L^2 + 2\varphi$$

and

$$BIC = L^2 + \log(m)\varphi ,$$

where L^2 is defined in equation 2.10, φ is the total number of parameters in the model (for both fixed and random parts) and m is the total sample size. Singer and Willett (2003) commented on the ambiguity of using BIC for the case of a longitudinal multilevel model as it is not clear whether the sample size m should be for the number of individuals in the data or the effective sample size that accounts for the repeated observations within individuals. It is worth mentioning that the models compared using AIC and BIC need not be nested but they should be fitted to the same sample, and smaller values for AIC or BIC indicate a better fit of the reduced model.

Model selection in the multilevel modelling framework also involves model checking. Section 2.1.1 presented the assumptions which the multilevel models are based upon. As for the classical regression models, the assumptions made for the multilevel model need to be checked after the fitting of the model. The failure of these assumptions compromises the interpretation of the estimated parameters. In addition, the conclusions regarding the relationship between the outcome and the covariates can be misleading. When the assumptions are not valid the hypothesis tests are invalid as well (Snijders and Bosker, 1999).

The model checking process of a multilevel model is very similar to that for the classical linear regression. The difference is that, because of the multiple levels and the multiple residual terms, each error component requires checking. A general graphical inspection is usually performed to assess the assumptions of linearity, homoscedasticity and normality. The linearity assumption is made for the relationship between outcome and explanatory variables. This assumption can be checked by directly plotting the outcome against the explanatory variables. For departures from this assumption additional terms for the explanatory variables, such as for example squared or cubic terms, can be included in the model. As stated before, the residuals are assumed to be normally distributed. This assumption can be checked by inspecting a normal probability plot. Last, but not least, the assumption of constant variance of the level one raw residuals must be checked. This assumption can be checked, for example, by plotting the residuals against the fitted values. Note that, when random slopes are added to the model, the composite residuals are no longer assumed to have constant variance.

The residuals $\boldsymbol{u}_j$, or the cluster specific effects, are random variables rather than parameters of the multilevel model (Snijders and Bosker, 1999). For a random slope model, such as the model in equation 2.6, each cluster has its own predicted line. If only the fixed part is considered these lines are all the same. The cluster specific effects need to be considered so that each cluster has their specific fitted line. These cluster specific effects need to be predicted (Frees, 2004; Skrondal and Rabe-Hesketh, 2004; Longford, 1993) from the model in order to check their assumptions. This prediction is usually performed through Empirical Bayes (EB) (Efron and Morris, 1975) estimation. This method combines information from the cluster of interest with the other clusters, accounting for the cluster size and the covariance matrix of the observations (Snijders and Bosker, 1999). If only the random intercept is considered, the EB estimate for the random intercepts is given by

$$\hat{u}_{0j} = \frac{n_j\hat{\sigma}^2_{u0}}{n_j\hat{\sigma}^2_{u0} + \hat{\sigma}^2_e}\tilde{y_j} = Sh \times \tilde{y_j} \ ,$$

where $\tilde{y_j}$ is the cluster mean of the raw residuals $y_{ij} - \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}$. The EB residuals are also called the shrinkage estimates because of the shrinkage factor $Sh = \frac{n_j\hat{\sigma}^2_{u0}}{n_j\hat{\sigma}^2_{u0}+\hat{\sigma}^2_e}$. This factor pushes the mean of the raw residuals for cluster j towards the general mean. In other words, the EB residuals bring the estimates of the random intercepts and random slopes closer to the mean. As the cluster size increases Sh approaches to one, and the EB residuals will be approximately the same as the mean of the raw residuals. The cluster specific effects can also be compared by means of a caterpillar plot (Goldstein, 2003). This plots, for each cluster, their predicted random effects with respective confidence intervals ordered according to their magnitude. The comparison is performed by assessing those confidence intervals that do or do not overlap with the others. This type of graph can also assist in grouping clusters according to their performance.

After the model has been selected and the assumptions checked the interpretation of the parameter estimates can proceed. In the analysis of multilevel linear models the interpretation of the fixed effects is the same as for the standard linear regression analysis. In other words, it can be said that for a unit increase in x_k, y would have an expected change of $\hat{\beta}_k$, keeping all other variables constant. If the model includes squared terms of some of the explanatory variables or interaction terms between any of them, these effects should be interpreted together. If a categorical variable is also included in the model, the interpretation of its effect compares the effect of each of the categories with the omitted category, the baseline.

Models presented in the subsequent chapters, however, consider the fit of a log-transformed outcome variable. In this case the interpretation of the parameters differs to that mentioned above. Instead of an expected change of $\hat{\beta}_k$ in the outcome for a unit change in x_k, the expected change in y is a $e^{\hat{\beta}_k}$ fold increase/decrease, depending on the sign of $\hat{\beta}_k$. In other words, there will be an expected

$$b_k\% = 100 \times (e^{\hat{\beta}_k} - 1)\% \tag{2.11}$$

change in y for a unit increase in x_k (Tufte, 1974). However, when x_k is also considered with a log-transformation in the model this is no longer the interpretation of $\hat{\beta}_k$. In this case $\hat{\beta}_k$ represents the elasticity of the outcome with respect to x_k. Dougherty (2002) defined elasticity as the proportional change in y for a given proportional increase in x_k. For cluster level variables the interpretation can follow as for the level one variables. In addition, the coefficients of the contextual variables that represent proportions can be multiplied by any constant a and the formula in equation 2.11 can be modified to

$$b_k\% = 100 \times (e^{\hat{\beta}_k \times a} - 1)\%, \tag{2.12}$$

where, for example, a can be equal to 0.1. This gives the interpretation that there will be an expected $b_k\%$ change in y for a 10 percentage point increase in contextual variable k.

The random part of the model can also be interpreted. To assist in the interpretation of the random part of the model, a plot with the cluster specific regression lines can be constructed by using the EB residuals. The presence of the random intercepts in the model means that each cluster has its own intercept u_j

that varies randomly across clusters. Therefore, a graphical representation of this model would show parallel regression lines, one for each cluster, showing how the outcome cluster mean varies randomly across clusters. However, if the number of clusters is relatively large, this plot can be constructed for a subset of clusters. The same plot for the random coefficients model will show non-parallel cluster specific lines, as each line will also depend on the values of $\boldsymbol{z}_{ij}$.

Interpretation can also be given for the variance components and the intra-cluster correlation can be calculated. The covariance term between the random intercepts and random slope, σ_{u01}, can also be interpreted. This parameter shows the relationship between the random slope and intercept, and it can be used to assess for example whether clusters with above average intercept have above average or below average slopes.

One very important point raised by all the authors cited so far is the need to centre the covariates around their means in order to improve the interpretation of the random effects. Centring is highly advisable for those explanatory variables where the value 0 (zero) has no substantive meaning, and it is good practice to centre or re-scale these variables. In a multilevel model (or in a longitudinal multilevel model) Singer and Willett (2003) discussed whether the centring should be around either the total mean or the group mean and advised the use of group centring only if it can be justified substantively.

2.1.4 Continuous versus Discrete Outcome

The models presented so far were formulated for a continuous outcome. However, multilevel models may also accommodate discrete outcomes. They are an extension of generalized linear models (GLM) where the relationship between the expected response and the linear predictor (η) is determined by a link function ($g(.)$) as shown in the model in equation 2.13:

$$E(y_i|x_i) = Pr(y_i = 1|x_i) = g^{-1}(\eta_i) \ . \tag{2.13}$$

In model 2.13, y_i is a binary outcome variable with usual values coded as 1 (success) and 0 (failure). The linear predictor η_i is defined as $\boldsymbol{x}_i^T\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is the vector of explanatory variables and $\boldsymbol{\beta}$ is its associated vector of fixed coefficients. The model in equation 2.13 is then a generalized linear one-level fixed effects model for non-clustered data. This model can also be written as:

$$g(\pi_i) = \eta_i \tag{2.14}$$

where $g(.)$ is the function that links the probability of success π_i with the covariates in the linear predictor. The choice of which link function to use will depend on the conditional distribution of the outcome variable. For binary outcomes, where the distributions are either Bernoulli or Binomial, the most usual link functions are the logit and the probit links defined respectively as:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

and

$$\phi^{-1}(\pi_i),$$

where ϕ^{-1} is the inverse of the cumulative standard normal distribution. Between these two link functions the logit is possibly the most used for its nice properties and interpretation (Rabe-Hesketh and Skrondal, 2005). A logistic model, i.e. a GLM model using the logit link, is a linear model for the log-odds, which is the log of the expected number of successes for each failure. The interpretation for the logistic regression model coefficients is as a multiplicative effect on these odds. Another nice way of interpreting the results of a logistic model is through the evaluation of the predicted probabilities that are given as

$$\hat{\pi}_i = \{1 + \exp(-\eta_i)\}^{-1} .$$

These predicted probabilities can be plotted for any combination of the covariates in the model and compared.

Now consider the model:

$$\text{logit}(E(y_{ij}|x_{ij}, u_j)) = \text{logit}(Pr(y_{ij} = 1|x_{ij}, u_j)) = \ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \eta_{ij} + u_j \ , \tag{2.15}$$

where

$$\pi_{ij} = \{1 + \exp(-\eta_{ij})\}^{-1}$$

and

$$y_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij}) \ .$$

This is a two-level random intercept logistic model, in which the success probabilities π_{ij} now depend not only on the individual but also on the clusters. As in the model in equation 2.1, the clusters are the level two units with subscript j and individuals are the level one units with subscript i. The linear predictor η_{ij} is now defined as $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$, where $\boldsymbol{x}_{ij}$ is the vector of explanatory variables for both of the levels of the data which is associated with the vector of fixed effects $\boldsymbol{\beta}$. The u_j are the cluster specific random intercepts and as before it is also assumed that:

$$u_j{\sim}N(0, \sigma_u^2).$$

It is then assumed that the log-odds are normally distributed in the population of clusters (Snijders and Bosker, 1999). It is worth noticing that the model in equation 2.15 does not contain the term e_{ij} for the level one residuals. Therefore, σ_e^2 is not estimated. This is because σ_e^2 is not a free parameter and is directly determined by the success probabilities (Snijders and Bosker, 1999). This model belongs to the class of generalized linear mixed models (GLMM).

GLMMs also allow for the inclusion of random coefficients at the cluster level. However, the estimation of the GLMMs needs approximation methods as there is no closed form for their marginal likelihood. Such methods are based on numerical integration and use iterative methods such the Newton-Raphson algorithm (Rabe-Hesketh and Skrondal, 2005). The main methods are those that approximate the marginal joint probability of the responses by the Gauss-Hermite adaptive quadrature (see Skrondal and Rabe-Hesketh (2004) for details) or those that use first and second order Taylor linearisation of the likelihood (see Goldstein (2003) and Snijders and Bosker (1999) for details). The methods based on the linearisation, called Marginal quasi-likelihood (MQL) and Penalized quasi-likelihood (PQL), allow for the use of IGLS or RIGLS. However, both MQL and PQL are not very stable methods. MQL produces biased estimates and PQL have less bias but are less precise than MQL. Another disadvantage of the linearisation methods is that the deviance produced with such methods cannot be used in testing the hypothesis of the model, while this can be done when using the adaptive quadrature methods. However, the accuracy of the adaptive quadrature methods depends on the number of integration points used in the estimation (STATA Press, 2005). All these methods differ mainly on how the random part of the model is estimated while the fixed part is estimated in roughly the same way (Snijders and Bosker, 1999). The choice between the estimation methods, however, will depend on the choice of statistical package available to be used in the analysis. `Stata` mainly uses adaptive quadrature methods while `MLwiN` uses MQL or PQL. Snijders and Bosker (1999) provides a list of references for more details on estimation methods of the GLMMs.

It is also worth mentioning that the fixed part of a GLMM can be tested, as in the linear multilevel model, through simple or multiple Wald test statistics. However, the EB residuals cannot be simply predicted as in the linear case as they do not have normal distribution, making the diagnostics of GLMMs a bit more complex to perform (Snijders and Bosker, 1999).

2.2 The Robust *Sandwich* Estimator for Multilevel Data

First consider the one-level cross-sectional linear model that assumes that the observations are independent and identically distributed (IID)

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \ ,$$

where it is assumed that

$$e_i \overset{IID}{\sim} N(0, \sigma_e^2).$$

This model is often estimated through OLS, which under this model is equivalent to ML estimation. ML estimators are unbiased but not fully efficient. A sufficient condition to guarantee consistent variance estimation of the ML estimators is that the assumed probability model, i.e. the assumption of normal distribution of the errors, is correctly specified. The asymptotically robust *sandwich* covariance estimator protects against the misspecification of the model assumptions in order to ensure variance consistency (Huber, 1967; White, 1982; Royall, 1986). It produces standard errors for the estimates of the regression coefficients which are robust to non-normal errors.

The principle behind the *sandwich* estimator is to estimate the covariance matrix of the parameter estimates directly from the observed sample. This method is then based on the observed sample Fisher information matrix and on a diagonal matrix of the cross-product of the estimated raw residuals. First, consider that the asymptotic model-based covariance matrix of the parameter estimates is given by:

$$Cov(\hat{\boldsymbol{\beta}}_{OLS}) = (X^T X)^{-1}(X^T \mathbf{V} X)(X^T X)^{-1} \ , \tag{2.16}$$

where $\mathbf{V} = E(\boldsymbol{e} \ \boldsymbol{e}^T)$, $\boldsymbol{e}$ is the vector of errors and X is the matrix with the explanatory variables $\boldsymbol{x}_i$. The idea of the *sandwich* estimator is then to substitute the unknown covariance matrix $\mathbf{V}$ by

$$\hat{V}_{sand} = \text{diag}(\hat{\boldsymbol{e}} \ \hat{\boldsymbol{e}}^T) \tag{2.17}$$

which is a diagonal matrix of the cross-product of the vector $\hat{\boldsymbol{e}}$ for the estimated raw residuals

$$\hat{e}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{OLS} \ .$$

The asymptotically robust *sandwich* estimator for the covariance matrix of the parameter estimates is then given by:

$$\widehat{Cov}_{sand}(\hat{\boldsymbol{\beta}}_{OLS}) = (X^T X)^{-1}(X^T \hat{V}_{sand} X)(X^T X)^{-1} \ . \tag{2.18}$$

Because the *sandwich* estimator is based on the OLS estimated raw residuals, which tend to be too small, it is biased downwards (MacKinnon and White, 1985). For this reason, statistical software implements some of the different corrections proposed in MacKinnon and White (1985) to adjust for the bias. These corrections include, for example, scaling the raw residuals $\hat{e}_i$ by $(1 - h_{ii})^{-1/2}$ or by the square of this adjustment, where h_{ii} are the main diagonal elements of the hat matrix $H = X(X^T X)^{-1} X^T$. These two corrections adjust each residual according to their influence, also protecting against residual heteroscedasticity. Another alternative to reduce bias is to use what MacKinnon and White (1985) called the degrees of freedom correction for $\hat{V}_{sand}$, as:

$$\hat{V}_{sand_{adj}} = \frac{m}{m-p} \hat{V}_{sand} \ ,$$

where m is the total sample size and p, here, is the number of regression parameters being estimated.

Now, consider the two-level random coefficients model:

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_j + e_{ij} \ . \tag{2.19}$$

This is the same model as in equation 2.6 in Chapter 2. The vector $\boldsymbol{x}_{ij}^T$ is the vector for the explanatory variables of both levels of the data hierarchy associated with the vector of fixed regression coefficients $\boldsymbol{\beta}$.

OLS estimation of a multilevel model provides unbiased but inefficient parameter estimates with negatively biased standard errors (Maas and Hox, 2004). For this reason, multilevel models are often estimated through IGLS which also assumes normality of the residuals in order to generate ML estimates (Goldstein, 1986). Nevertheless, even under violation of the normality assumption the IGLS parameter estimates for both the fixed and the random parts of the model are still asymptotically unbiased and consistent. On the other hand, the standard errors of these estimates are incorrect and cannot be used for hypothesis testing except in large samples (Goldstein, 2003). One way to correct the standard errors of the parameter estimates of a multilevel model is also to use the asymptotically robust *sandwich* covariance estimator, given by:

$$\widehat{Cov}_{rob}(\hat{\boldsymbol{\beta}}_{IGLS}) = (X^T\hat{\Sigma}_r^{-1}X)^{-1}(X^T\hat{V}_r^{-1}\hat{V}_{sand}\hat{V}_r^{-1}X)(X^T\hat{V}_r^{-1}X)^{-1} \ , \tag{2.20}$$

where $\hat{V}_r$ is defined as the estimated block-diagonal covariance matrix of the observations from the final iteration of the IGLS, with blocks corresponding to cluster j written as

$$\boldsymbol{z}_j^T\hat{\Sigma}_u\boldsymbol{z}_j + I_{n_j}\hat{\sigma}_e^2 \ ,$$

which is the same as in equation 2.8. The matrix $\hat{V}_{sand}$ in 2.20 is also a matrix with the cross-product of the estimated vector of raw residuals $\hat{\boldsymbol{e}}_j$. The difference is that for the multilevel case $\hat{V}_{sand}$ is a block-diagonal matrix where each of the blocks represent each of the n clusters (Goldstein, 2003). In a two-level multilevel model, each block of the matrix $\hat{V}_{sand}$ is calculated as

$$\hat{\boldsymbol{e}}_j \times \hat{\boldsymbol{e}}_j^T \ ,$$

where $\hat{\boldsymbol{e}}_j$ is the vector of level one estimated raw residuals,

$$\hat{e}_{ij} = y_{ij} - \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}_{OLS} \ ,$$

within cluster j. These blocks are not just a diagonal matrix as in the one-level case. They are the cross-product matrices once individuals nested within the same cluster are no longer thought to be independent. For a three-level model for example, with clusters K representing the highest level clusters, the cross-product matrix of the residuals is calculated at the third level. In this case, each of the blocks of the $\hat{V}_{sand}$ matrix represents one of the K clusters.

In the one-level case, if the assumed probability model is correctly or nearly correctly specified the use of the *sandwich* estimator is not necessary (Freedman, 2006). In this case the unrobustified asymptotic covariance estimation is more efficient and produces smaller standard errors (Maas and Hox, 2004). Furthermore, Maas and Hox (2004) stated that when level two residuals in multilevel models are non-normal, the use of the robust *sandwich* estimator is not advantageous. They showed that the violation of this assumption had almost no effect on the estimates of the fixed part of a multilevel model. Therefore, they confirmed the belief that the methods for the estimation of multilevel models are already quite robust to non-normality of the higher level residuals. However, this violation was shown to have an impact on the estimation of the random part of the model. Maas and Hox (2004) showed that the *sandwich* estimator performed better and was more accurate in this case. Hence, the importance of its use when the investigation of, for example, the level-two variation is of interest in the analysis. The comparison between asymptotic and robust standard errors for the estimation of a multilevel model is advisable. A large difference between these two estimators is an indication of model misspecification.

2.3 Multilevel Longitudinal Models

Multilevel longitudinal models are an extension of the cross-sectional multilevel models described until now. Therefore, all that was already discussed in the previous sub-sections can be applied to the longitudinal data models. To avoid repetition, this sub-section focuses on specific points that need to be addressed when fitting a multilevel longitudinal model.

The longitudinal models are an extension of cross-sectional models, but not only for the increased number of levels. For example, for a simple data set a longitudinal model can be represented by a two-level model. The difference is that the individuals are now the level two units and the level one units are their repeated measures. However, in practice, many longitudinal data sets originate from multistage sample surveys. Therefore, they should be represented by higher order multilevel models, such as by a three-level model where the clusters represent the higher level units (Longford, 1993).

In a longitudinal multilevel model most of the variability of the data is expected to be at the level two units, the individual level (Goldstein, 2003). This is in contrast to cross-sectional multilevel models where individuals are the level one units and most of the variability is found at this level. Therefore, in a longitudinal analysis, one should consider a well specified level two model, which should include individual characteristics represented by main effects and interaction effects. It is worth mentioning that random slopes for individual and higher level units may also be considered under this framework. When only the random intercept is considered in a longitudinal multilevel model, it assumes that the covariance structure of the observations is exchangeable. This assumption is not always valid, particularly when the repeated measures are taken within short periods of time. One way to relax this assumption is to include, for example, a random slope for the time variable (Snijders and Bosker, 1999; Goldstein, 2003). Care must be taken, however, to ensure that the time point related to zero has a meaningful interpretation as it represents the initial status. Therefore, the centring of the time variable or its re-scaling is highly advisable.

This type of random coefficients longitudinal multilevel model is also known as the growth curve model (Rabe-Hesketh and Skrondal, 2005). It is a two-level random coefficients model, where level two are the individuals and level one are the measurements taken on various occasions that are nested within individuals. In the set of explanatory variables at the occasion level there is usually a variable to represent time, often represented by the variable age or some related measurement (Frees, 2004). The variable for time receives a random effect as growth is expected to vary across individuals (Bryk and Raudenbush, 1992). It is also common to represent growth as a polynomial function of the variable for time, generally including squared, cubic or even quartic terms for age for example. The growth curve model then fits separate trajectories for each individual and these are calculated by using the EB estimates, mentioned in section 2.1.3, for the random part of the model.

One of the major advantages of fitting a longitudinal model in the multilevel framework is the possibility of assessing cross-level effects (Frees, 2004). For example, with the inclusion of cross-level interactions between an individual variable and the time variable. This allows the direct investigation of whether the effect of the specific variable on the outcome varies over time. The cross-level interactions need not only be between level two and level one, but maybe between any levels of the data which are of interest to the analyst.

Now consider the following model representation:

$$y_{tij} = \boldsymbol{x}^T_{(1)tij}\boldsymbol{\beta}_{(1)} + \boldsymbol{x}^T_{(2)ij}\boldsymbol{\beta}_{(2)} + \boldsymbol{x}^T_{(3)j}\boldsymbol{\beta}_{(3)} + v_j + \boldsymbol{z}^T_{tij}\boldsymbol{u}_{ij} + e_{tij} \ . \tag{2.21}$$

This is a three-level multilevel model for change (Singer and Willett, 2003) or a longitudinal multilevel model. It is an extension of the model in equation 2.6. The three levels are represented by, respectively from the higher to the lower, the subscript j for the n clusters, the subscript i for the n_j individuals nested within clusters and the subscript t for the T_{ij} time points or occasions for each individual. According to what was already mentioned, neither the T_{ij}s or the n_js need to be the same for every individual or cluster.

One of the main differences of the two models is that model in equation 2.21 includes three sets of explanatory variables, one for each of the levels. The vector $\boldsymbol{x}_{(1)tij}$ contains the time-varying covariates or the occasion level variables. In this set, at the occasions level, there might be a variable representing time (Frees, 2004). The vector with all explanatory variables of the model in equation 2.21 is defined as:

$$\boldsymbol{x}^T_{tij} = \left(\boldsymbol{x}^T_{(1)tij},\ \boldsymbol{x}^T_{(2)ij},\ \boldsymbol{x}^T_{(3)j}\right) .$$

In this vector there are variables for each of the three levels representing main effects of continuous or categorical variables, same-level interaction terms or cross-level interaction terms. The vector $\boldsymbol{x}_{tij}$ is associated with the vector of fixed regression coefficients:

$$\boldsymbol{\beta}^T = \left(\boldsymbol{\beta}^T_{(1)},\ \boldsymbol{\beta}^T_{(2)},\ \boldsymbol{\beta}^T_{(3)}\right) ,$$

where the number between parentheses indicates to which level the regression coefficient refers.

The other components of the model in equation 2.21 are the error terms and the vector $\boldsymbol{z}^T_{tij}$ that in matrix form for individual i in cluster j is defined as:

$$Z_{ij} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_{T_{ij}} \end{pmatrix} . \tag{2.22}$$

Here it is also assumed that:

$$v_j \sim N(0, \sigma_v^2) \ ,$$

$$\boldsymbol{u}_{ij} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}\right)$$

and

$$e_{tij} \sim N(0, \sigma_e^2) \ .$$

As in the cross-sectional case, these three residual terms are assumed to be mutually independent and uncorrelated with the covariates in the model.

The design matrix in equation 2.22 contains the vector of ones for the random intercept and the random effect for the time variable, here called z. In this case, the model in equation 2.21 has only the random slope for the time variable. In the growth curve model, such as the model in equation 2.21, the random intercept is interpreted as the initial status, and the random slope for the variable that represents time is interpreted as the rate of growth. The covariance term σ_{u01} then measures the relationship between the initial status and the rate of growth, measuring for example if those that had a high initial value will grow faster, or not, than those with lower initial values.

Consider the composite residuals

$$r_{tij} = v_j + u_{ij} + e_{tij} \ ,$$

from the estimation of a three-level random intercept model of the form

$$y_{tij} = \boldsymbol{x}_{tij}^T \boldsymbol{\beta} + v_k + u_{ij} + e_{tij} \ .$$

The total variability for this model can then be expressed as:

$$Var(y_{tij} | \boldsymbol{x}_{tij}) = \sigma_v^2 + \sigma_u^2 + \sigma_e^2 = \sigma_r^2 \ .$$

For a three-level random intercept model authors such as Longford (1993) and Rabe-Hesketh and Skrondal (2005) presented different intra-cluster correlation coefficients that can be calculated. Here, only the correlation between two measurements of the same individual in the same cluster is presented, which can be calculated as:

$$Cor(y_{tij}, y_{t'ij}|\boldsymbol{x}_{tij}, \boldsymbol{x}_{t'ij}) = \frac{\sigma_v^2 + \sigma_u^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2} \qquad \forall t \neq t'.$$

However, for the model in equation 2.21 which includes both random intercept and random slope, the composite residuals have actual form given as

$$r_{tij} = v_j + \boldsymbol{z}_{tij}^T \boldsymbol{u}_{ij} + e_{tij} \ .$$

The diagonal elements of the covariance matrix of the observations are given as:

$$\sigma_v^2 + \boldsymbol{z}_{tij}^T \Sigma_u \boldsymbol{z}_{tij} + \sigma_e^2 \ ,$$

and the off-diagonal elements are given as:

$$\sigma_v^2 + \boldsymbol{z}_{tij}^T \Sigma_u \boldsymbol{z}_{t'ij} \qquad \forall t \neq t'.$$

If $\boldsymbol{z}_{tij}$ contains time, this covariance matrix depends on the value of the time variable. The variance of the observations will have a quadratic relationship with the time variable. The elements of the correlation matrix for the observations can be calculated in a similar way to equation 2.9, as:

$$\rho_{tt'ij} = \frac{\sigma_{r_{tt'ij}}}{\sqrt{\sigma_{r_{tij}}^2 \sigma_{r_{t'ij}}^2}} \qquad \forall t \neq t'.$$

which shows the dependence on the time lag between the observations. Hence, the correlation structure is no longer exchangeable.

Modelling longitudinal data within the multilevel framework has an important advantage as this approach automatically handles unbalanced data sets. Because the observations are nested within individuals, models of this type can easily handle unbalanced and unequally spaced data sets. In unbalanced data sets, individuals can have measurements taken at different numbers of time points, for reasons of panel attrition or determined by design. In unequally spaced data sets, the time lag between each measure for all individuals does not need to be the same. Therefore, a data set which is unbalanced by design can be handled with this approach. The multilevel longitudinal models presented so far, although allowing for the presence of a random slope for time, still assume that the occasion level residuals are uncorrelated. As already mentioned, this assumption is not usually plausible particularly for observations measured close in time (Hox, 2000). Bryk and Raudenbush (1992) mentioned that correlation is bound to exist between the level one residuals for each individual and that the level one residuals can be correlated with some occasion level variables. However, Bryk and Raudenbush (1992) also stated that to assume the level one residuals are not correlated in longitudinal data sets with few time points is a practical assumption and that under this assumption significance tests are unlikely to be affected. If it is necessary to impose a covariance structure on the residuals, multilevel models are flexible enough to allow for this through a slight change of approach that treats each response as a component of a multivariate normal distribution. This is the multivariate multilevel modelling approach that is briefly described in sub-section 2.3.2. Before introducing the multivariate multilevel model, alternative error covariance structures which can be imposed on the error components are briefly listed in the next sub-section.

2.3.1 Covariance Structures

The error covariance structure of the model in equation 2.21 allowed for heteroscedasticity but not for autocorrelation between the composite residuals. This section reviews the most common types of covariance matrices that can be imposed on the error components.

Suppose that the number of time points T_{ij} is equal to 3 for every individual i and that a two-level longitudinal model is considered. The following structures can be imposed (Singer and Willett, 2003; Diggle et al., 2002; Fitzmaurice et al., 2004):

- **Unconstrained**: It imposes no specific structure on the error covariance matrix Σ_r. This covariance matrix is of the form:

$$\Sigma_r^{unc} = \begin{pmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}.$$

  Each diagonal term, the variances, and off diagonal term, the covariances, of Σ_r^{unc} has their own value and they are parameters of the model to be estimated. Models estimated imposing the unconstrained, or unstructured, error covariance will have smallest deviance due to the larger number of estimated parameters. With T occasions there will be $\frac{T\times(T+1)}{2}$ extra parameters.

Fitzmaurice et al. (2004) stated this structure is not advisable for highly unbalanced data sets or data sets with relatively few individuals. Singer and Willett (2003) advised comparing models fitted with imposed unstructured correlation using the AIC and BIC criteria.

- **Compound Symmetry**: Also known as the exchangeable structure or uniform. This is the usual structure assumed in the longitudinal multilevel model when only the random intercept is considered. In other words, it is assumed that the variance at any time point is the same, as well as that the covariance between any pair of time points is the same. For the data set being considered in this section, this covariance matrix is of the form:

$$\Sigma_r^{exch} = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 & & \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{pmatrix}.$$

  With this structure the correlation between any pair of residuals will also be the same and equal to ρ as in equation 2.5. In addition, if a random slope is fitted but $\hat{\sigma}_{u01}$ and $\hat{\sigma}_{u1}^2$ are small the exchangeable structure may hold.

- **Heterogeneous Compound Symmetry**: This is an extension of compound symmetry but not assuming homoscedasticity along the diagonal terms of Σ_r. In addition the assumption of equal covariance between pairs of residuals is also relaxed.

$$\Sigma_r^{hexch} = \begin{pmatrix} \sigma_1^2 & & \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 \end{pmatrix}.$$

  This structure has a constant autocorrelation parameter ρ also estimated by the model.

- **Autoregressive**: This is the first-order autoregressive correlation structure, also called exponential for continuous time data (Schabenberger and Pierce, 2001). The variances are assumed constant across time and equally spaced pairs of responses have the same covariance (Fitzmaurice et al., 2004) which depends on the lag between them. It causes the “band-diagonals” of Σ_r to be the same. The main diagonal expresses a constant variance term and the other diagonals are determined by:

$$Cov(r_t, r_{t'}) = \sigma^2\rho^{\text{lag}} \qquad \text{lag} = 1, 2... \quad .$$

The Σ_r matrix under this structure has the form:

$$\Sigma_r^{ar} = \begin{pmatrix} \sigma^2 & & \\ \sigma^2\rho & \sigma^2 & \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{pmatrix}.$$

This structure assumes that the correlation between pairs of residuals diminishes for larger lags. The model estimates only two variance components. However, the degree to which the correlation diminishes is determined by a constant ρ.

- **Heterogeneous Autoregressive**: This is an extension of autoregressive structure but not assuming homoscedasticity along the diagonal terms of Σ_r, just as with the heterogeneous exchangeable structure. In addition the terms of the off diagonals are determined by:

$$Cov(r_t, r_{t'}) = \sigma_{r_t}\sigma_{r_{t'}}\rho^{\text{lag}} \qquad \text{lag} = 1, 2... \quad .$$

Under this structure the covariance matrix of the residuals has the form:

$$\Sigma_r^{har} = \begin{pmatrix} \sigma_1^2 & & \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 \end{pmatrix}.$$

This is more flexible than Σ_r^{ar}.

- **Toeplitz**: This structure has similar characteristics to the Σ_r^{ar}. However, the elements of the band-diagonals are not forced to reduce by a fixed fraction (Singer and Willett, 2003). This structure still considers the main diagonal to be constant and the covariance matrix under this structure has the form:

$$\Sigma_r^{toep} = \begin{pmatrix} \sigma^2 & & \\ \sigma_1 & \sigma^2 & \\ \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}.$$

This imposes that pairs equally separated in time have the same correlation (Fitzmaurice et al., 2004) and is only appropriate for equally spaced data. The different variance components are parameters of the model to be estimated. Compared to those with Σ_r^{ar}, models with the Toeplitz[2] structure will have less residual degrees of freedom.

[2] Named after Otto Toeplitz, also defined as an ARMA process.

Two more general covariance structures that deserve mentioning are the following.

- **Spatial Power**: Also known as Markov Structure (Khattree and Naik, 1999). This is a reparameterisation of the exponential correlation structure, which, as mentioned earlier, is equivalent to a continuous time autoregressive structure. The exponential correlation structure can be written as:

$$Cov(r_t, r_{t'}) = \sigma^2 \exp\left(\frac{|t-t'|}{-\phi}\right) \qquad \forall t \neq t' \ .$$

  This structure, like the AR(1), imposes that the correlation between any pairs of residuals will be smaller if measured further apart (Diggle et al., 2002). Furthermore, the larger the value of $1/\phi$ the faster the correlation decays towards zero as the distance between the pairs of residuals increases. The reparameterisation for the Spatial Power structure involves setting

$$\rho = \exp\left(\frac{-1}{\phi}\right)$$

  and expressing the covariance terms as:

$$Cov(r_t, r_{t'}) = \sigma^2 \rho^{|t-t'|}.$$

  This is a direct generalization of AR(1) for unequally spaced data that takes into account the distance between the T occasions by powering ρ by $|t-t'|$. The name for this structure, spatial power, is justified as it is usually applied to studies of spatial processes (Khattree and Naik, 1999). For the data set considered in this section, the Spatial Power covariance matrix is also of the form:

$$\Sigma_r^{pow} = \begin{pmatrix} \sigma^2 & & \\ \sigma^2\rho^1 & \sigma^2 & \\ \sigma^2\rho^2 & \sigma^2\rho^1 & \sigma^2 \end{pmatrix}.$$

- **General Linear**: Assuming that the residual covariance matrix can be expressed as a linear function of $\boldsymbol{\theta}$, the general linear covariance structure (SAS Institute Inc,Version 8, 1999; Khattree and Naik, 1999; Jennrich and

Schluchter, 1986) and (Pourahmadi, 2007) is of the form:

$$\Sigma_r^{gen} = \theta_0 A_0 + \theta_1 A_1 + ... + \theta_k A_k,$$

where the matrices A_k are known symmetric matrices and the parameters θ_k are unknown and unrelated covariance parameters to be estimated by the model (Khattree and Naik, 1999). The known matrices A_k can be set to represent any of the known structures or any desirable structure with the requirement that Σ_r^{gen} must be a positive definite matrix. For example, the compound symmetry structure could be expressed as:

$$\Sigma_r^{exch} = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 & & \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{pmatrix}$$

$$= \sigma_u^2 \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2.3.2 Multivariate Multilevel Models

Longitudinal multilevel models, as described above, assume conditional independence between the repeated outcomes, after controlling for the explanatory variables and the random effects. However, in longitudinal panel data, it is expected that successive measurements within the same individual are correlated. For example, the income of a head of household measured in month $t + 1$ is expected to be correlated with their income measured in month t conditioned on the head of the household's characteristics. This conditional correlation imposes a structure in the error covariance matrix, which is often of interest in the analysis of a longitudinal data set.

Multivariate multilevel models provide the appropriate tools for the analysis of a longitudinal data set where the error covariance matrix has no restrictions and is also of interest. This extends the general multivariate regression analysis, where a balanced data set with no missing observations is required (Longford, 1993). Multivariate multilevel models are also seen as an extension of the two-level growth curve model. However, time is now treated as a discrete variable. Therefore, a categorical variable for time is included in the model and each occasion is represented by a "dummy" variable. These occasion dummies are treated as having both fixed and random coefficients. Extra levels of the data hierarchy, in addition to the occasion and the individual levels, can also be considered in the model. The occasion dummies can be set to vary randomly at these higher levels as well. However, this is not a necessary set up. A potential disadvantage of the multivariate multilevel model is that, as it is equivalent to setting up one equation for each occasion response, it may not be suitable for panel data with a large number of occasions.

There are some advantages of fitting models to longitudinal data within the multivariate multilevel framework. For example, a balanced data set is no longer a requirement. The data can include individuals with different numbers of time points. Therefore, these models can be used to analyse rotating panel data and can be further constrained to accommodate the planned missing data as performed in Yang et al. (2002). However, they are not as flexible as the growth curve models when dealing with unequally spaced data (Fraine et al., 2005) as some error structures can only be considered for equally spaced data. The multivariate multilevel approach also handles missing data, as long as the assumption of missing at random holds[3].

As already mentioned, this approach does not assume that the repeated outcomes are conditionally independent (Griffiths et al., 2004). The error variance components are parameters of the model to be estimated. Moreover, constraints on these parameters can be made in order to impose different error covariance structures, such as those in subsection 2.3.1. However, here lies another disadvantage of the multivariate approach. Although the variance components are now estimated and they can also be constrained, they are no longer interpreted as cluster specific effects or individual specific effects as before (Snijders and Bosker, 1999).

The multivariate multilevel approach extends the random slope model in equation 2.21 so that each individual response in a given occasion t is considered as a component of a multivariate normally distributed random vector $\boldsymbol{y}_{ij}$. These $\boldsymbol{y}_{ij}$ are simultaneously modelled under the multivariate multilevel model (Goldstein, 2003).

To make it clear, consider the same hierarchy as before, where occasions are nested within individuals which are nested within clusters. The occasion level (subscript t, varying from 0 to T) defines the multivariate structure (Goldstein, 2003). The individuals are the level one (subscript i, varying from 1 to n_j) and

[3] This assumption is described in section 2.4 of Chapter 6.

clusters the level two (subscript j, varying from 1 to n). A multivariate multilevel model with only the time variable as covariate, treated as having both fixed and random effects at the individual level, can be written as:

$$y_{tij} = \boldsymbol{d}_{tij}^T \boldsymbol{\beta} + \boldsymbol{d}_{tij}^T \boldsymbol{v}_j + \boldsymbol{d}_{tij}^T \boldsymbol{u}_{ij}. \tag{2.23}$$

In this model $\boldsymbol{d}_{tij}$ is the vector with the T occasion dummies. They are defined to indicate whether the row in the data set refers to the response at occasion t, being equal to 1 or equal to zero (Snijders and Bosker, 1999). Note that this model can handle intermittent missing response by setting all the elements of the dummy for the missing occasion to zero. The occasion dummies are associated with the vector of fixed regression coefficients

$$\boldsymbol{\beta}^T = (\beta_1,\ \beta_2, \ldots,\ \beta_T)\,,$$

and are also associated with the vectors of random effects at both the individual $\boldsymbol{u}_{ij}$ and the cluster level $\boldsymbol{v}_j$. Note that all the T dummies are included in the model. Therefore the model in equation 2.23 does not contain the intercept. This is a two-level multivariate model.

Now consider a balanced data set where the total number of occasions per individuals is fixed and equal to four and that the time points are labelled from 0 to 3. Here it is also assumed that:

$$\boldsymbol{v}_j {\sim} MN(\boldsymbol{0}, \Sigma_v) \text{ and } \boldsymbol{u}_{ij} {\sim} MN(\boldsymbol{0}, \Sigma_u),$$

where

$$\Sigma_v = \begin{pmatrix} \sigma_{v0}^2 & & & \\ \sigma_{v10} & \sigma_{v1}^2 & & \\ \sigma_{v20} & \sigma_{v21} & \sigma_{v2}^2 & \\ \sigma_{v30} & \sigma_{v31} & \sigma_{v32} & \sigma_{v3}^2 \end{pmatrix} \text{ and } \Sigma_u = \begin{pmatrix} \sigma_{u0}^2 & & & \\ \sigma_{u10} & \sigma_{u1}^2 & & \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 & \\ \sigma_{u30} & \sigma_{u31} & \sigma_{u32} & \sigma_{u3}^2 \end{pmatrix}.$$

Defining $\Sigma_r = \Sigma_v + \Sigma_u$, the multivariate vector of responses for an individual is:

$$\boldsymbol{y}_{ij} {\sim} MN(D_{ij}\boldsymbol{\beta}, \Sigma_r)\ ,$$

where D_{ij} is a matrix containing the vectors $\boldsymbol{d}_{tij}^T$. The model in equation 2.23 is a fully multivariate multilevel model with saturated covariance structure (Snijders and Bosker, 1999). Therefore, this model has no occasion level variance estimated. This is an important assumption of this model, that is that there are no measurement errors in the repeated outcomes (Fraine et al., 2005). However, it is a necessary assumption to ensure model identification. This model also assumes that the effect of time varies randomly across clusters. An alternative model formulation would be to consider a common random intercept at the cluster level for all the occasions. This model can be written as:

$$y_{tij} = \boldsymbol{d}_{tij}^T \boldsymbol{\beta} + v_j + \boldsymbol{d}_{tij}^T \boldsymbol{u}_{ij}. \tag{2.24}$$

Here it is assumed, that

$$v_j \sim N(0, \sigma_v^2) \text{ and } \boldsymbol{u}_{ij} \sim MN(\boldsymbol{0}, \Sigma_u),$$

where σ_v^2 is a scalar and Σ_u is as defined above. Linear and non-linear constraints can be applied to the elements of Σ_u in order to express the different forms of correlation structures. Multivariate multilevel models can be fitted imposing these different structures. The IGLS and RIGLS methods, described in section 2.1.2, can also be used to estimate the multivariate multilevel model. These methods provide, for the fixed part of the model, statistically efficient parameter estimates and accurate standard errors. They also provide efficient estimates of the error variance components (Goldstein, 2003).

The multivariate model in equation 2.24 has no explanatory variables other than the time dummies but the inclusion of such variables is straightforward. This model still allows for the inclusion of different sets of covariates for the different levels of the data, including the occasion level. These variables can be further considered as having common or separate coefficients for each of the time points. Separate coefficients are produced by including interaction terms between the time dummies and the covariates, and they can be jointly tested for their significance in the model.

The same steps for model checking as those for univariate multilevel models are applied to the multivariate multilevel models. In addition, hypothesis testing for the significance of parameters of a given explanatory variable can be performed for each of the outcomes. In other words, some explanatory variables may be statistically significant at one occasion but not at others.

This approach has not been vastly explored in the statistical literature. For example, Yang et al. (2000) presented the multivariate approach for the analysis of a longitudinal data set on voting attitudes. They modelled a discrete response comparing the general multilevel longitudinal model with the multivariate model.

The models from both approaches were estimated via penalized quasi-likelihood estimation. Their multivariate model, a two-level model, had time dummies treated as random at both the cluster level and the individual level. Multivariate Wald tests were used to decide on the inclusion or exclusion of variables for each of the responses. Furthermore, the models tested were based upon a data set with a fixed number of occasions for each individual.

Barbosa and Goldstein (2000) presented a multilevel longitudinal model for discrete response assuming the responses within the same individuals were positively correlated. Barbosa and Goldstein (2000) used the same data example as Yang et al. (2000) trying to extend their models to accommodate unequal time points but noted that in this case the multivariate multilevel approach could no longer be applied. Instead, Barbosa and Goldstein (2000) fitted one three-level longitudinal model and two time-series multilevel models as defined in Goldstein et al. (1994). In this definition, the time-series model is a multilevel longitudinal model where the level one variance is considered as a function of time through an autocorrelation function (first or second order). This allows for rather complex dependency structures of the level one residuals. In their analysis of the time-series models, Barbosa and Goldstein (2000) considered different autocorrelation functions and aimed to compare their results with the multivariate model in Yang et al. (2000).

In another article Yang et al. (2002) applied the multivariate longitudinal framework to data under a non-random missing mechanism. This was allowed by setting up constraints in the covariance matrices of both levels considered in the analysis. Fraine et al. (2005) compared the longitudinal growth curve model with the multivariate multilevel model. Their models were applied to data on student well being. They advocate the use of multivariate models when the number of time points is small. Their multivariate model considered an unstructured error covariance and the test for different specifications for the covariance matrix was supported. Plewis (2005) also compared the multivariate multilevel model formulation with the growth curve models, for both continuous and binary outcomes. His findings under the different specifications were consistent. His multivariate models imposed restriction to the error covariance matrix and tests for other structures were also suggested.

2.4 Non-response in Longitudinal Data

As already mentioned in the previous sections, the methods for the analysis of longitudinal data sets under the multilevel modelling framework allow for unbalanced, or incomplete, longitudinal data sets. Lack of balance in a longitudinal data set can be planned or unplanned: planned when the data were generated from a rotating sampling scheme for example and unplanned due to uncontrolled non-response.

2.4.1 Planned Non-response

Longitudinal, or panel, data are commonly collected through rotating sampling schemes such as that described in Chapter 3. These schemes consist of the substitution of part of the sample in successive occasions, or waves, of a panel survey. For example, in every wave a group of individuals, or a panel, is excluded from the sample (rotated out) and it is substituted by another panel of individuals (that are rotated in). In some rotating designs, panels once rotated out are allowed to be rotated in again later on. However, in others, once a panel is rotated out it is definitively excluded from the sample.

The reasons for using rotation schemes were stated in Hsiao (2003). The first was related to optimal sampling. As mentioned in Steel (1997), the portion of the sample that overlaps between consecutive months (waves of the panel) of a rotating panel survey has an effect on the precision of the estimates of change. The second reason is more practical, as it relates to the concern that individuals participating in longitudinal surveys may change their behaviours influenced by the time they are in the sample.

These schemes generate unbalanced data sets that were planned by the sampling design. The unbalance arises from the fact that not all individuals are observed on every occasion. In addition, for those panels which were rotated out and rotated back into the sample, planned non-response will occur for the out of sample period. This type of planned non-response can be assumed as being missing completely at random (Fitzmaurice et al., 2004). Therefore, methods for balanced data can be extended for application to planned unbalanced data sets (Hsiao, 2003). The use of a multilevel framework is also recommended in the presence of rotating samples (Goldstein, 2003). This issue is dealt with in Chapter 5.

2.4.2 Unplanned Non-response

Unplanned non-response in a longitudinal data set can be of different forms, due to:

- delayed entry, causing **initial non-response**, where the initial wave has missing values;
- intermittent or **wave non-response**, causing missing values to appear in the middle of the panel;
- early-exit, or **drop-out** or **attrition**, when individuals leave the panel prematurely;
- individuals refusal to participate, causing **unit non-response**, and
- **item non-response** to various questions in the survey.

Because of these different forms, and also due to the different occasions, the problem of unplanned non-response is more serious in longitudinal data sets than in cross-sectional data sets (Fitzmaurice et al., 2004). Non-response in a longitudinal data set is usually not homogeneous across all the occasions. A particular concern is whether the non-response process happens at random or whether non-respondents are systematically different compared to respondents. This potential self-selective mechanism might be a source of bias and in smaller sample sizes might cause an effect on the efficiency of the estimates (Winkels and Davies, 2000). Moreover, ignoring the non-response mechanism when missingness is related to the outcome causes the analysis to be seriously biased (Fitzmaurice et al., 2004). Therefore, the investigation of the mechanism that generates the non-response, mainly in the sense of how this mechanism is related to the outcome variable, is suggested by Little (1995); Diggle et al. (2002) and Frees (2004).

Consider a two-level longitudinal model where subscript t represents the occasion level and subscript i the individual level. In addition, consider a response indicator R_{ti} that takes the value 1 when the data are observed and 0 otherwise. The vector of responses for individual i $\boldsymbol{y}_i$ can be partitioned as $\boldsymbol{y}_i = \{\boldsymbol{y}_i^O, \boldsymbol{y}_i^M\}$, where $\boldsymbol{y}_i^O$ are the observed responses and $\boldsymbol{y}_i^M$ are the missing responses. Three main types of non-response mechanisms are identified and they are listed below.

MCAR (missing completely at random). This is the most restrictive mechanism. The assumption of MCAR holds if

$$Pr(R_{ti}|\boldsymbol{y}_i^O,\ \boldsymbol{y}_i^M,\ X_i) = Pr(R_{ti})\ ,$$

meaning that the response probability does not depend on the outcome values or on the matrix of explanatory variables X_i. In other words, the outcome is not related to the response mechanism. However, it can happen that

$$Pr(R_{ti}|\boldsymbol{y}_i^O,\ \boldsymbol{y}_i^M,\ X_i) = Pr(R_{ti}|X_i)\ .$$

This still means that the distribution of response does not depend on the outcome variable. However, it does depend on the matrix of explanatory variables X_i. This is called covariate-dependent MCAR. When missingness is dependent on the covariates, the model should include the set of variables that are the predictors of the response (Fitzmaurice et al., 2004).

A similar idea can be applied to the drop-out mechanism as defined by Little (1995). In this sense, the drop-out is completely at random when the probability of individuals dropping out in a given occasion does not depend on past, current or future values of their outcomes. If the drop-out mechanism is covariate-dependent the response model should also include the drop-out predictors.

In the presence of MCAR, the data can be treated as if they were unbalanced by design. The usual methods for balanced data can then be applied. Alternatively, the set of completers can be used in the analysis. Note that, under the MCAR assumption, the set of completers can be considered to be a random sample from the target population. Analysis based on a complete-case data set is not biased under MCAR (Little, 1995). However, a complete-case analysis is less efficient and may discard substantial parts of the data.

MAR (missing at random). This is a less restrictive assumption than MCAR, and occurs when:

$$Pr(R_{ti}|\boldsymbol{y}_i^O,\ \boldsymbol{y}_i^M,\ X_i) = Pr(R_{ti}|\ \boldsymbol{y}_i^O,\ X_i)\ .$$

That is, the response mechanism depends on the observed outcomes and on the covariates but not on the missing outcomes.

A similar idea can also be applied to the drop-out mechanism. Drop-out is at random if the probability of dropping out depends on the observed response, i.e. on the past outcomes.

Care must be taken, however, when analysing data based on this mechanism. Under MAR, the set of completers can no longer be considered as a random sample from the target population (Fitzmaurice et al., 2004). Therefore, an analysis based on the set of completers will be biased. In addition, estimation methods based on ML are recommended in this situation when all the available data are used in the analysis.

NMAR (non-missing at random). This is also called non-ignorable missing. This occurs when the response mechanism depends on both the set of observed and missing outcomes and it cannot be ignored in the analysis.

In a similar fashion, Little (1995) defined non-ignorable outcome-based drop-out and random coefficient based drop-out. In the former, drop-out depends on the missing components but not on future values of the outcome. The random coefficient based drop-out, in turn, depends on the current, past and future values of the outcome while depending on the values of the random coefficients of the individuals in the analysis.

For any type of non-ignorable missingness both a model for the response and a model for the drop-out mechanism should be evaluated simultaneously. Failing to do so generates seriously biased results. Diggle et al. (2002, Chapter 13) presented a review of models to predict non-response that can be used in the analyses of longitudinal data. This review includes the two main classes of models for drop-out mechanism also presented in Little (1995). These are the selection models and pattern mixture models. Diggle et al. (2007) stated that the literature for drop-out models in longitudinal analysis is quite extensive. These authors also presented the most common approaches for the analysis of longitudinal data with drop-out and proposed a new method that takes the histories of the subjects into account. Most of the proposed methods to model the drop-out mechanism only accommodate time-constant covariates (such as those in Little (1995), Hawkes and Plewis (2006) and Diggle et al. (2007)). Roy and Lin (2002) described a method that accommodates time-varying covariates, in which a transitional model for the missing time-varying covariates is estimated in addition to a multivariate random coefficient model for the response and a random coefficient selection model for the drop-out mechanism. All these methods, however, ignore intermittent non-response.

Under any of these mechanisms, Frees (2004) suggested that when possible some follow-up procedures should be performed to attenuate non-response bias. Multiple imputation methods can also be applied in longitudinal data sets with non-response. Fitzmaurice et al. (2004) discussed some imputation methods for longitudinal data but this is beyond of the scope of this thesis.

2.4.3 Longitudinal Weights

One other alternative approach to attenuate the non-response bias in survey data is to use adjusted sampling weights based on the response distribution. These adjusted weights compensate for the unequal probabilities of selection and for the unplanned non-response that occurs in the surveys (Kalton and Bryk, 2000). In panel surveys, these weights are usually in the form of cross-sectional and longitudinal weights, which depend on the patterns of wave non-response and on the attrition patterns. However, due to the varied number of measurement occasions, the weighting process in panel surveys is inevitably more complex than in cross-sectional surveys.

When the set of longitudinal weights is not available with the released panel data, they can be constructed. There are different methods to calculate these weights and the choice between the type of longitudinal weights to be used depends on the objective of the analysis. They are usually calculated to: adjust for wave non-response and attrition patterns or to adjust for the attrition patterns only (Lepkowski, 1989). The objective of using longitudinal weights is to compensate for the data loss (Kalton and Bryk, 2000) in each occasion. Therefore, there is a different set of weights for every occasion which adjusts the responding patterns to compensate for the non-responding patterns (Lepkowski, 1989). The longitudinal weights are usually calculated for the set of respondents, and the non-response cases are eliminated from the data set or are assigned weight zero.

The sets of longitudinal weights that are calculated to account for both the wave non-response and the attrition patterns are the most complex type. There will be up to 2^T patterns of non-response in a panel data set with T occasions. This would require the construction of up to $2^T - 1$ sets of longitudinal weights to allow the analysis of data for all possible combinations of occasions, T. The set of longitudinal weights that accounts only for the attrition patterns, in turn, is the simplest. For example, in a longitudinal analysis that includes data from the first occasion to occasion t, only the set of weights at occasion t is needed which is adjusted to compensate for sample losses in all previous occasions (Kalton and

Bryk, 2000). For that, only the set of individuals present from the first occasion to occasion t needs to be considered in the analysis. However, this results in the elimination of valid data. Lepkowski (1989) suggested modifying the wave non-response patterns so that they are expressed as attrition patterns. This modification results in fewer data being eliminated when the attrition weights are calculated. However, it ignores the possibility that wave non-respondents might be fundamentally different to those who leave the survey prematurely.

Panel surveys, such as the BHPS (Taylor et al., 2009), LFS (ONS, Office for National Statistics, 2009), Panel Study of Income Dynamic (PSID) (Gouskova et al., 2008), Survey of Income and Program Participation (SIPP) (Fuller and An, 1996; Kobilarcik and Singh, 1996; Allen and Petroni, 1994) and Survey of Labour and Income Dynamics (SLID) (LaRoche, 2003; Hunter et al., 1992), use different methods to calculate the longitudinal weight adjustments. These methods vary in complexity but all have as a first step the definition of a base weight which is usually the cross-sectional weight for the first occasion already adjusted to account for initial wave non-response.

The simplest method involves classifying respondents and non-respondents in weighting cells or classes according to the information available for all. The non-response adjustment factor is calculated as the inverse of the response rate in each of the classes (LaRoche, 2003). This response rate is calculated as the weighted sum of the respondents sample over the weighted sum for the eligible sample in that class. Respondents have their base weights adjusted by this factor and non-respondents receive weight zero. This method often uses decision trees to define the different cells and is usually followed by some kind of calibration method (Kalton and Bryk, 2000).

Another method commonly used involves fitting logistic regression models for the propensity of being a respondent (or non-respondent) (Hunter et al., 1992; Rizzo et al., 1996). The outcome variable for the logistic model is the response indicator, like the R_{ti} from the previous section. The covariates are usually categorical variables taken from the data of the previous occasions. The model can include main effects and interactions between these variables and also sampling design variables (Lepkowski, 1989). The non-response adjustment factor is calculated as the inverse of the predicted probabilities for the respondents and applied to their base weights. Non-response bias is expected to be reduced once the model controls for the covariates that are related to the response propensity (Kalton and Bryk, 2000). When only categorical variables are used this method works in a similar way to the adjustment cells methods.

The use of longitudinal weights is advisable in order to compensate for sample losses between sequences of occasions. This practice ensures the sample is representative of the population at the time the sample was selected (LaRoche, 2003). Ignoring the non-response in a longitudinal analysis might yield biased estimates, as this implies the assumption of equal distributions of the outcome variable for respondents and non-respondents (Pfeffermann and Sikov, 2008). Non-response reduces the sample size and in longitudinal surveys it might have an effect on the availability of the longitudinal component. The use of methods that adjust for non-response reduces the bias in the estimation of population parameters while preserving the relationships between the survey variables, provided that the elimination of the available cases is not significant (Kalton and Bryk, 2000).

2.5 Summary

This chapter focussed on a review of the models for the analysis of longitudinal data within the multilevel modelling framework. Random intercepts, random slopes and multivariate multilevel models were described in detail. This chapter also presented a review of the alternative error correlation structures that can be imposed on the estimation of multivariate multilevel models; the robust sandwich covariance estimator for multilevel models and the issue of non-response in longitudinal data. Chapter 6 presents a review of methods for the analysis of a longitudinal data under the multilevel modelling framework accounting for the sampling weights, including methods for calculating longitudinal weights.

Chapter 3

The Brazilian Monthly Employment Survey

3.1 Introduction

In the introductory chapter of this thesis the Brazilian labour force survey was mentioned as the survey providing the data to be used. This chapter presents some details of the methodology of this survey whose official name *Pesquisa Mensal de Emprego* translates to Brazilian Monthly Employment Survey, hereafter denoted as PME.

The PME is a probabilistic household sample survey conducted by the Brazilian national statistics office, the IBGE. This survey has a complex multistage sampling design. It is conducted every month and the households selected in the sample are rotated in and out following a rotation scheme. Its main aim is to produce monthly indicators on the Brazilian labour market and it currently covers the urban areas of the six main metropolitan regions of Brazil. The next section presents some history of the PME survey and details of its design. The information in this section was taken from official documents provided by the IBGE, mostly from the PME methodological notes in IBGE (2002).

IBGE makes the PME data available in monthly files. These files need to be merged in order to create a longitudinal data set. However, care must be taken when merging these data files to ensure that the households and individuals within households are correctly matched. This is an important issue of this survey which is dealt with in Section 3.3. This chapter concludes with a discussion on the selection of two working data sets from the PME to be used in the following chapters: a cross-sectional and a longitudinal data set.

3.2 The PME

3.2.1 History

Since the 60s, the IBGE have recognized the need to conduct a regular household survey to be able to better understand the changes in the country's social, demographic and economic development. IBGE's main national household survey PNAD[1], which started in 1967 as a quarterly survey, became annual in the late 70s. However, PNAD did not provide information on short term changes in the labour market. Hence, IBGE developed its labour force survey, first implementing it in 1980 (IBGE, 2001a). The main objective of the PME is to investigate the characteristics of the Brazilian labour force, producing monthly indicators that aid the evaluation and planning of the country's socio-economic development (IBGE, 2009). The PME is a labour force survey comparable to the UK Labour Force Survey (UK-LFS) and the U.S. Current Population Survey (CPS).

Brazil is the largest country in South-America, and according to the IBGE the estimated population in the year 2005 was about 186 million inhabitants. Geographically, the country is divided into 5,560 municipalities. These municipalities are grouped into 26 Federation Units (or States), and one Federal District. These states are grouped into five great regions: North, North-East, Centre-West, South-East and South. Some municipalities are also grouped to form the 28 Brazilian metropolitan regions.

Recognizing that the core of the Brazilian labour market is located in the metropolitan areas, the PME was initially designed to cover the nine metropolitan areas that existed at the time the survey was introduced, and the Federal District. However, due to budget constraints, the survey was first implemented in two of these regions and two years later it was covering six metropolitan regions: Recife and Salvador both located in the North-East Region; Belo Horizonte, Rio de Janeiro and São Paulo located in the South-East and Porto Alegre located in the South of Brazil. These are still the six metropolitan regions currently covered by the survey.

PME has been one of the main sources of labour market indicators in Brazil. Since it was first implemented, the PME survey methodology required some revisions. The first revision was in 1982. It was concerned with changes in the conceptualization of work and aimed to relate to the questionnaire of the national household survey PNAD. Later, in 1988, the survey had its second revision, where

[1] Pesquisa Nacional por Amostra de Domicílios.

only the sample size was reduced (Silva and Moura, 1988). This second revision was justified as it improved the quality of the field work. Before this revision, a listing process was carried out every two years with a sample of new households found in selected enumeration areas being added to the original sample. However, the number of interviewers remained the same or even reduced over the years. This resulted in an increasing number of incomplete interviews. Silva and Moura (1988) showed that this survey revision reduced the rate of non-response from around 24% to 5%. Later, in 1993, a similar revision was performed again in order to reduce the non-response rate.

In 1996, besides the PME, there were other surveys investigating the labour force being conducted by other institutes of research in Brazil. In an effort to improve the utility of the indicators provided by the different surveys, the Ministry of Labour proposed the unification of such surveys. For this purpose IBGE started the latest revision of PME. No consensus was reached on the unification of such surveys which differed mostly on the definition of the main labour market indicators and design. However, IBGE continued with the revision process recognizing that the PME survey should be updated to be able to capture the latest changes in the Brazilian economy (IBGE, 2002). The revised survey was only fully implemented in 2002. The main modifications concerned the definition of certain variables following recommendations from the International Labour Organization (ILO) to be applied to general labour force surveys. In the PME this involved modifications of the sampling design, the questionnaire ordering and the regional coverage. The revised PME was reduced to cover only the urban areas of the six already included metropolitan regions. Due to important methodological changes in this latter revision, the historic series initiated with the first implementation of PME, was discontinued. However, IBGE still makes available both series.

IBGE is the first user of the PME data: it uses the PME mainly to produce monthly reports comparing the monthly employment indicators and the gross flows in and out of employment. It also produces technical reports such as IBGE (2001a,b, 2002, 2003, 2009). Some recent studies conducted by IBGE using the PME data have been: on the increasing participation of women in the labour market and their allocation as the head of the household (IBGE, 2006b); on the profile of domestic workers and their increasing participation in the labour force (IBGE, 2006c) and also on the increasing participation of the population aged over 50 in the labour market (IBGE, 2006a).

Other users besides the IBGE are those that implement more elaborate analysis of the PME data. However, the majority of the studies conduct cross-sectional or pooled cross-sectional time series analyses, while very few conduct longitudinal analyses (Corseuil and Santos, 2002). One example of the use of the PME exploring its longitudinal component can be found in Sedlacek et al. (1989). This study used data from the PME before the latest revision to investigate labour market mobility in a short period of time. The problem of matching individual records was raised in this study. Ferrao (2002) is also another example of the use of the PME data before its latest revision. This study used the multilevel modelling approach to investigate the participation in the labour market. Other studies, also utilizing the PME data before the latest revision, are: Lemos (2002, 2006), both used old PME with matching difficulties mentioned in the former; Schwartzman (1999) that mentioned the reformulation and implementation of the question for skin colour; and Barros et al. (2000) which used pooled time series analysis. The lack of studies exploring the longitudinal component of the PME might be justified by the difficulty in matching individual records to form a proper longitudinal data set. This issue is discussed later on in this chapter.

3.2.2 Sampling Design and Rotating Panel Design

The PME sample of households was designed to be representative of the urban population for each of the six metropolitan regions covered by the survey. The sampling scheme adopted by the PME is characterised as a stratified two stage cluster design with unequal probabilities of selection. Samples of households are selected separately from each metropolitan region.

In each of the metropolitan regions groups of municipalities or pseudo-municipalities[2] form the independent strata. The census enumeration sectors, which are the primary sampling units (PSU), are selected independently from each stratum in each of the metropolitan regions. The secondary sampling units (SSU) are the households which are selected from each PSU. All residents from the selected households are surveyed. The motivation for this design is to ensure the spread of the sample within each metropolitan region.

The PSUs are selected by systematic sampling with probability proportional to their total number of private occupied households, as listed in the 2000 Demographic Census. From the selected PSUs, the households are selected by simple systematic sampling, using a random start and a fixed interval of selection with

[2] The definition of pseudo-municipality is that when more than one municipality of smaller size, according to the 2000 Census, are joined together to represent one stratum with sufficient size to allow for the minimum number of PSUs to be selected from it (IBGE, 2002).

the initial goal of selecting 16 households per PSU. Each PSU remains in the sample for a decade. It will only be replaced earlier by a similar one if there exists a shortage of households to be selected or if the whole selection process is renewed.

Table 3.1 presents the PME sample composition. It shows the total number of selected PSUs and selected households when the sample for the last revision was designed, March 2001. The table also shows the equivalent information for December 2005, for comparison. The change over time is due to the rotation of the sample and the re-listing process carried out every year. Information for 2005 in this table was constructed from the monthly micro-data.

Table 3.1: The PME Sample Composition

	Number of Munici-palities*	Number of PSUs in the Population and Selected in the Sample			Number of Selected HHs		1/ (Sampling Fraction)*
		Population Mar-01*	Sample Mar-01*	Sample Dec-05	Mar-01*	Dec-05	
Recife	14	3,068	261	283	4,715	5,610	200
Salvador	10	4,604	243	272	4,684	5,549	200
Belo Horizonte	33	14,710	359	389	6,644	7,544	200
Rio de Janeiro	19	20,612	406	441	7,576	8,309	500
São Paulo	39	3,023	431	471	7,820	9,119	700
Porto Alegre	30	4,982	329	378	5,773	6,763	200
Total	145	50,999	2,029	2,234	37,212	42,894	

Notes: * Taken from IBGE (2002).

Another important feature of the PME sampling design is its rotation scheme. By design, each selected household is interviewed in four consecutive months. They are left out of the sample for eight consecutive months and return after this period to be interviewed again in four consecutive months, after which they are excluded from the sample. This characterises a rotating panel design known as *4-8-4*. According to this rotation scheme, 25% of the sample is substituted every month. In addition, surveys one year apart have 50% of households in common. This rotating design allows following selected households over time for a period of 16 months, with an 8 month gap between the fourth and fifth interviews.

Table 3.2 shows the representation of the rotation scheme for the year 2004. The cells of this table show the interview time for each panel and month. The super columns in Table 3.2 represent the selection groups[3]. In 2004 these selection groups were called D, E, F and G. Selection group D is not shown in the table. Each selection group is divided into eight sub-samples with approximately the same size, called rotation groups (RG). These are the eight columns (labelled from 1 to 8) in Table 3.2. The RGs are composed by a set of PSUs from any of the six metropolitan regions. The panels of households, which are rotated in and out, are represented by the combination of selection groups and RG. For example,

[3] Selection groups are an administrative make up. It is formed by a set of households.

the panel of households F4 was first introduced in the survey in January 2004. In addition, once allocated to a RG, a PSU can only belong to this RG. In this sense, panels G4 and F4, for example, are composed by the same set of PSUs. However, the households allocated in selection group G are not the same as those allocated in F. In other words, each selected PSU has households allocated to different panels over the time that PSU is considered in the PME.

Table 3.2: The PME Rotation Scheme 4-8-4

	Selection E								Selection F								Selection G						
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7
Jan									4	3	2	1											
Feb	5									4	3	2	1										
Mar	6	5									4	3	2	1									
Apr	7	6	5									4	3	2	1								
May	8	7	6	5									4	3	2	1							
Jun		8	7	6	5									4	3	2	1						
Jul			8	7	6	5									4	3	2	1					
Aug				8	7	6	5									4	3	2	1				
Sep					8	7	6	5									4	3	2	1			
Oct						8	7	6	5									4	3	2	1		
Nov							8	7	6	5									4	3	2	1	
Dec								8	7	6	5									4	3	2	1

According to this rotation scheme, samples in consecutive months of the same year have 6 panels in common, and samples one year apart have 4 panels in common. This means that for every 12 months of data there is 50% of the sample that overlaps. Moreover, in every month there are households being interviewed from the first until the eighth time. This rotation scheme is the same as the one adopted by the CPS (Bureau of Labour Statistics, 2002). The use of rotating panel designs, like the one for the PME, over permanent samples is justified as it reduces the respondent's burden. It also allows for longitudinal comparisons when a substantial number of household units are matched in consecutive months and years (Sedlacek et al., 1989).

The PME sample is designed such that households once excluded from the sample after the eighth interview, do not return to the survey in the following years. However, the PME, like most surveys of this type, is not a longitudinal survey per se. It is designed to follow household spaces (dwellings) but not individuals over time. If residents from a selected household move out from it, while the household is in the PME sample, no effort is made to follow these residents. The new occupants of this household will be those to be interviewed for the rest of the period that this household remains in the sample. This represents a drawback of this data set as one cannot guarantee individual records to be accurately linked across the distinct monthly surveys.

One other important aspect of the PME sampling design is that, unlike similar labour force surveys, in the Brazilian one, there is no change in the questionnaire for the different months that the households are surveyed. Therefore, the selected household will respond to the exact same questionnaire in all eight times it is interviewed.

3.2.3 Sampling Weights and Unit Non-response Corrections

By design, the PME sampling scheme is initially self-weighted within each metropolitan region. This means that the selected households and individuals from the same metropolitan region would have the same sampling weight. The sampling weights are the inverse of the sampling fraction, the inclusion probabilities, presented in Table 3.1 earlier in this chapter. However, the last methodological revision of the survey also included the correction of the sampling weights to account for unit non-response. This causes the loss of the self-weighted characteristic of the design.

Without the treatment for non-response the design weights in a given metropolitan region g, are written as:

$$w_g = \frac{1}{f_g} = \frac{1}{n_{hg} p_{jhg} \frac{m_{jhg}}{M_{jhg}}},$$

where:

f_g is the sampling fraction in the metropolitan region g;

n_{hg} is the number of selected PSUs in the h^{th} stratum in the metropolitan region g;

p_{jhg} is the relative size of the jth PSU in the hth stratum (according to the 2000 Census) in the metropolitan region g;

M_{jhg} is the number of listed households in the j^{th} PSU in the h^{th} stratum, and

m_{jhg} is the number of selected households in the j^{th} PSU in the h^{th} stratum in the metropolitan region g.

For simplicity of notation the subscript for the metropolitan region g is omitted hereafter. The treatment involves adjusting m_{jh} to represent m^*_{jh}, the number of respondent households (with interview undertaken) and n_h to represent n^*_h, the number of PSUs within each stratum with at least one household interviewed. The corrected sampling weights are written as:

$$w^*_{ijh} = \frac{1}{n^*_h p_{jh} \frac{m^*_{jh}}{M_{jh}}} \ , \tag{3.1}$$

which are the same for each household i in PSU j but may vary across PSUs and strata for the same metropolitan region.

Two types of sampling weights are available in the survey micro-data. The corrected weights w^*_{ijh} (3.1) and the calibrated weights w'^*_{ijh} (3.2), that are adjusted by the population projection of each of the metropolitan regions[4]. The calibrated weights are written as:

$$w'^*_{ijh} = w^*_{ijh} \frac{P}{\hat{P}^*} \ , \tag{3.2}$$

where P is the population projection of a given metropolitan region and $\hat{P}^*$ is the estimated population, calculated from the sample using the non-response adjusted weights w^*_{ijh}. These sampling weights are provided in each of the monthly data sets. However, the survey does not provide longitudinal weights.

3.2.4 Variables in the Survey and Main Concepts

The field work of the PME is divided into two operations: annual update of the sampling frame, where households in the selected sectors are listed and classified, and monthly face-to-face interviews, which are spread out in the four weeks of the month following a pre-defined schedule. Data are collected using an electronic questionnaire divided into four parts. Part 1 contains the variables that serve to identify the statistical units. Part 2 collects demographic data on all residents of the selected households. Residents aged 10 years or over answer Parts 3 and 4 of the questionnaire. Part 3 collects educational data and part 4 work characteristics. Table 3.3 presents a list of the main characteristics investigated in the PME.

The statistical unit of investigation of the PME are people within the selected households. Households, or dwellings, are defined as “the site structurally separated and independent, designated to serve as habitation for one or more individuals, or which has been used as such”. The PME covers all people residing in the household except people residing in embassies or consulates, hospitals,

[4] This projection is according to the population growth between the 1991 and 2000 census, and based on the fertility, mortality and migration rates, as explained in IBGE (2002).

Table 3.3: Characteristics Investigated by the PME

Units / Sub-Sets	Characteristics
Households	location and density.
Resident population	socio-demographic (gender, date of birth, race/colour, household dependency and family dependency).
Resident population aged 10 or over	education (literacy, school attendance and details on level of education, professional courses) and employment status.
Employed population away from work	reason and duration for being away from work.
Employed population at work	number of jobs, occupation, time in the job, main activity, number of employed people, usual and actual earnings and working hours, contribution to public pension, private or public sector, informality, fixed contract, action taken and search for job and availability to work extra hours.
Unemployed population	for those who have had a job before: time since last job, if in the last year, investigates the characteristics of previous work and reasons for leaving it; looking for work, actions to find a work, time since last looked for work, time searching, availability to work and number of hours wished to be working.
Economically inactive population	identifies those marginally attached to the economically active population and availability to work and discouragement.

boarding schools, orphanages, prison service establishments, military/defence establishments or religious houses. The main concepts adopted by PME follow the recommendations of the ILO made for the investigation of the labour force characteristics, and some of them are listed below:

Population of Working Age: Those aged 10 or over in the reference week which is the week from Sunday to Saturday prior to the week of interview.

Employment/Work: Paid or unpaid activity in the production of goods or services. Excluded from this concept are all unpaid activities for charity or for religious institutions, or in the production for auto-consumption. Those who undertook some paid or unpaid work for at least one hour in the reference week are classified as the **employed population**, those without a job and looking for one in this same week are the **unemployed population** and **inactive population** are those otherwise (students and those not looking for jobs).

Earnings: Money, benefit in kind, products or goods received in return for work done. It is sub-divided into usual (usual earnings for a complete month, the reference month, excluding any benefit received in that month) and actual (payment actually received including any extra earnings in the reference month) monthly earnings, from the main and second jobs.

Main Job: Work had in the reference week. This is also the work which the respondent worked for more hours, for a longer period of time and which paid the highest salary.

Occupational Strata: The employed population is divided into: employees (work for an employer, for a given work pattern in payment of money, products, goods or benefit), domestic workers (paid domestic services done in one or more households), self employed (employees at their own establishment or working on their own without any other employee, with or without a partner), employers (establishment owners with at least one employee under their supervision) and unpaid workers.

Employment Strata: This classifies the employed population into formal, informal, military worker and workers in the public sector (government posts). In Brazil, formal workers have an official document called *carteira de trabalho* (labour card), and this should be signed by the employer, otherwise the worker is classified as informal.

3.3 Matching of Households and Individuals

Around the world, LFS surveys have been used in a similar fashion to the PME. Rowe and Nguyen (2004) used the individual longitudinal data of the Canadian LFS to study unemployment transitions. They recognized the potential flaws of the data, such as problems in matching individual records, non-existence of retrospective data on unemployment and data censoring, which are also observed in the PME. Problems in the matching of individual records were also identified in Madrian and Lefgren (1999) for the CPS. In this report they proposed a combination of variables to form a key variable to assist in the matching and mentioned that when trying to match data of such surveys non-response should be considered.

This section presents the issue of matching the monthly PME data. As mentioned earlier, the main goal of the PME is to provide monthly indicators of the Brazilian labour market. Although it presents a rotating panel design that follows households over time, it was not originally designed to serve as a longitudinal survey. For this reason, like some equivalent labour force surveys such as the one conducted in Canada (Rowe and Nguyen, 2002), it does not provide on its monthly micro-data files a unique individual identifier that allows individuals within selected households to be matched over time. Instead, a set of variables must be used to uniquely identify each of the households and individuals. IBGE ensures that the values for these variables do not change in any of the months the household is in the sample. These variables are those that represent the metropolitan region, the stratum, the PSU and the panel that the households were in and two other variables that are called “control” and “series”. All these variables must be taken together to identify each one of the households. The omission of any of these variables causes the mismatching of household records. For matching individuals, however, IBGE does not provide any explicit recommendation of the set of variables to be used in any of the official documentation of the survey. It recommends to the user, who intends to apply longitudinal methods on the PME data, linking the individuals within matched households by using some of their demographic variables such as date of birth and gender.

Some studies mentioned in the previous section identified this problem when using PME data collected prior to the latest methodological revision. Antonaci and Silva (2007) explored this issue in more current PME data. In this study, the observed matching rates, for both households and individuals, were compared to the matching rate expected by design. Table 3.4 presents this expected rate of sample that overlaps between pairs of months. For example, 75% of the sample should be matched when considering samples one occasion apart, and 50% for samples taken one year apart. It is important to notice that, due to these rotating panels, samples from months separated by intervals of 4 to 8 months will not overlap, as well as samples separated by 16 months or more.

Table 3.4: Percentage of Sample Overlap between Pairs of Months

Month Interval	Sample Overlapping Determined by Design (%)
1	75.0
2	50.0
3	25.0
4-8	0.0
9	12.5
10	25.0
11	37.5
12	50.0
13	37.5
14	25.0
15	12.5
16 or more	0.0

Antonaci and Silva (2007), using PME data from March 2002 till December 2005, concluded that the observed household matching rates were not too different from those expected by design. Using the same set of variables as those mentioned above they found, for example, that the observed rate of household matching in samples one month apart was 71%; for samples nine months apart this rate was 11% compared to the expected 12.5%; and for samples one year apart this rate was 44%. Individuals were matched within the matched households considering their gender and date of birth. Due to the clustered design, used in the PME sample, the same expected rates are valid when matching individuals. Antonaci and Silva (2007) concluded that the rates of individual record matching were less than half of what was expected by the design (as in Table 3.4) for lags greater than three. They indicated this could be due to inconsistencies on the variable for date of birth.

The matching exercise presented in Antonaci and Silva (2007) compared month-by-month samples. Considering the set of PME data available to be used for years 2004 and 2005, this section presents further matching exercises. These exercises were performed considering the matching between data sets representing the different interview times. The monthly data for both years were pooled together and files for each of the eight interviews were created. Each one of these files could contain samples from every month of both years. By design, and if there were no unit non-response, the expected matching rate for any panel from one interview to the other is 100%. The different rates indicate the absence of households, for either non-matching or non-response reasons.

Table 3.5 displays the matching rates for nine of the panels. These panels are those that by design should have data for the eight interviews completed during the observation period. These panels are labelled from F4 to F8 and G1 to G4, and their starting date in the survey is presented in the table between brackets. The columns of Table 3.5 show four selected pairs of comparisons: comparing the 1^{st} interview with the 2^{nd}, the 4^{th} with the 5^{th}, the 7^{th} and the 8^{th} and also the 1^{st} with the 5^{th}. It can be noticed that around 96% of the households in the 1^{st} interview are matched with the households of the 2^{nd} in every panel. The matching rates when comparing the last two interviews are also around 96 to 97%. However, the matching between the 4^{th} and the 5^{th} interviews, considering the gap of eight months between them, and the 1^{st} and the 5^{th} are those showing the lower rates. As mentioned by Antonaci and Silva (2007) the main problem is with matching households in the first four interviews with any of the last four. For example, 84% of the households from the panel starting in January 2004 were successfully re-interviewed on the 5^{th} occasion. These lower rates might be an indication of mismatching and can be understood as panel non-response.

Table 3.5: Household Matching Rates of Selected Pairs of Interview Time and Panel

Panel	Interview Times (%)			
	1 and 2	4 and 5	7 and 8	1 and 5
F4 (01/04)	96.4	84.1	97.5	83.6
F5 (02/04)	97.9	89.0	97.7	88.7
F6 (03/04)	96.8	89.1	97.5	88.4
F7 (04/04)	96.3	89.3	96.9	89.0
F8 (05/04)	96.8	87.8	95.7	88.1
G1 (06/04)	96.6	89.5	97.2	89.6
G2 (07/04)	97.0	87.2	97.2	86.8
G3 (08/04)	97.2	87.2	97.7	86.9
G4 (09/04)	96.4	87.4	96.1	87.5

Madrian and Lefgren (1999) evaluated the matching of individual records in the CPS monthly data. Although expecting similar rates to those for the household matching, they recognized that the rates for individuals would be systematically smaller due to individual non-response, mortality and residential mobility. The same can be expected to happen in the PME case. Furthermore, measurement errors in the variables for date of birth, for example, also contribute to the lower individual matching rates. Madrian and Lefgren (1999), therefore, performed different matching exercises trying different combinations of key variables.

Similar exercises to those performed in Madrian and Lefgren (1999) were performed for the PME data available. The results for all these exercises are not shown here as the thorough investigation of the matching process goes beyond the scope of this thesis. The exercises were performed to assist in the selection of the final data set for the analysis. The exercises involved testing different combinations of key individual variables added to the set of those that identify the household. The variables considered were: the number of order of the individuals in the household; the number of household members; date of birth; and characteristics of the individuals such as gender, education level, skin colour and age. Table 3.6 presents the results for three of the exercises. The cells in this table represent the first off-diagonal of the comparison between interviews, i.e. the rate comparing the interviews taken one month apart. It is worth mentioning that these are values calculated for the same nine panels as in Table 3.5.

Table 3.6: Individuals Matching Rates of Selected Pairs of Interview Time and Panel

Pair of occasions	Number of Order	Gender and DOB	Other Variables
1 - 2	96.0	93.8	87.5
2 - 3	96.2	94.3	89.5
3 - 4	96.3	94.7	90.6
4 - 5	81.5	47.6	34.3
5 - 6	96.2	94.8	89.9
6 - 7	96.5	95.2	91.7
7 - 8	96.5	95.3	92.2

The first column in Table 3.6 considered the matching of individuals adding the variable for the number of order of individuals (ORD) within the households. At the first glance this variable appears to be the best candidate to be used combined with the household identifiers. However, this variable is not fixed throughout the different interviews: any change in the household structure causes the members to be re-ordered thereby receiving different ordering from those of previous interviews. Furthermore, if a new family moves to a household already in the survey, the same ordering is assigned to the new members. Hence, basing the individual identification solely on the addition of the ORD variable may cause the perfect match of records from completely different individuals or may cause the records of the same individual not to be matched at all. For this exercise Table 3.6 shows that the lowest rate is for the match between the 4^{th} and 5^{th} interviews.

The second column in Table 3.6 shows the matching rates when considering the gender and date of birth (DOB) variables added to the set of household identifiers. DOB can be constructed for every individual from the variables provided for day, month and year of birth. However, all these three variables present missing observations, which generates DOB missing. Furthermore, this variable needs to be re-calculated in every occasion, some occasions might present missing values and others not. Therefore, matching the individual records solely on this variable causes individuals who are observed on different occasions with valid data for all other variables not to be matched due to a missing DOB. Notice from Table 3.6 that a lower average matching rate is found when compared to the previous exercise. These lower rates might reflect the problems with the DOB variable, as also identified in Antonaci and Silva (2007), or might be an indication of a more accurate match than the previous one. Once again, the matching rate between the 4^{th} and the 5^{th} interviews is the lowest.

The last column in Table 3.6 shows the matching rates when considering the addition of the variables for total number of members in the household, gender, skin colour and year of birth (YOB). The variable for total number of members in the household is a control for residential mobility. However, the addition of this variable might prevent the successful matching of households that had a change in the number of members over the time of the survey. The addition of YOB was dictated by the difficulty in merging using DOB, as YOB contains less missing values. Table 3.6 shows that this last matching exercise contains the lowest rates of matching. This is a more restrictive set of key variables which might cause lower rates of mismatched individuals. These results are in accordance with the results presented by Madrian and Lefgren (1999).

From all the matching exercises performed, some not shown here, it was noted that the matching of individual records from different PME data files is not as effective as matching households themselves. It was also possible to conclude that the initial matching problems identified when only households were being considered are magnified when performing the matching for the individual records, particularly when the lag between the interviews is greater than three. The exercises indicated that individual records are better matched within either the first four interviews or the last four, but not among these two sets of interviews. This match can be done, but at a lower rate of accuracy and without confidence that actual matches are not being discarded.

3.3.1 Non-response Patterns for Household Units

The results for the matching between household records showed some indication of panel non-response (Little and Rubin, 2002). This was also identified in Antonaci and Silva (2007). This sub-section aims to provide a brief investigation of the patterns of non-response for the PME data set. Chapter 7 deals with this issue of non-response a bit further. Due to the difficulty in matching individual records this sub-section deals only with non-response at the household level.

A longitudinal data set for the PME survey can provide up to eight measurement occasions for each household. If every time the household is interviewed it is classified as "in" and each time that the household is not interviewed it is classified as "out", this gives a total of 256 (2^8) possible patterns of non-response. Note that, since the PME micro-data provided are only for those households interviewed in each month, no information on those households that were selected but never interviewed is available. Therefore the pattern for "out" in every occasion is not observed.

Consider the nine panels that should have the eight interviews completed by design in the 2004/2005 pooled data. These panels are those that have their first interview from January to September 2004. Therefore, if a household from any of these panels is classified as "out" this is evidence of panel non-response. Table 3.7 shows a summary for the classification as either "in" or "out" for the household units in these nine panels. The aim of this exercise is to illustrate the possible different sizes of samples of household units if different patterns of non-response were to be allowed. The last column of Table 3.7 represents the summation over the nine panels. It shows that if only the units with eight interviews completed (Completers in the table) were to be considered this number would be equal to 26,274. This represents 66% of the total number of households observed in these nine panels. In the table, "Wave NR" stands for intermittent non-response from a given interview time and it means that the unit was observed until one time before that. For example intermittent non-response from interview 6 indicates that the household unit was "in" in the first five interviews, "out" at the 6^{th} but returns at some point after that. If the household does not return to be "in" it is classified as "Drop-out" in the table.

Note, from Table 3.7, that the drop-out from interview time 5 is the most

Table 3.7: Non-response Patterns for Households in Panels with the Complete Set of Interviews Determined by Design

Interview	Panel									
Time	F4	F5	F6	F7	F8	G1	G2	G3	G4	Sum
Wave 8										
Completers	2,699	3,131	3,045	2,943	2,702	2,991	2,974	2,989	2,800	26,274
Wave NR	35	46	47	52	81	56	47	41	68	473
Wave 7										
Wave NR	26	17	33	20	42	28	22	25	20	233
Drop-out	25	29	24	20	43	30	24	16	17	228
Wave 6										
Wave NR	41	28	33	27	46	44	59	49	40	367
Drop-out	15	21	18	24	22	27	22	24	19	192
Wave 5										
Wave NR	266	171	159	144	157	144	196	192	169	1,598
Drop-out	198	176	184	163	162	149	196	202	185	1,615
Wave 4										
Wave NR	77	77	62	78	61	66	54	49	86	610
Drop-out	19	14	12	19	12	6	19	16	20	137
Wave 3										
Wave NR	78	64	96	78	109	67	64	77	89	722
Drop-out	24	15	9	7	16	16	19	14	7	127
Wave 2										
Wave NR	110	67	101	117	105	106	97	90	116	909
Drop-out	20	13	24	21	11	21	18	18	15	161
Wave 1										
Wave NR	759	611	577	644	736	652	652	636	638	5,905
Total	4,392	4,480	4,424	4,357	4,305	4,403	4,463	4,438	4,289	39,551

frequent. It is worth remembering that the set of households being considered in the table are those that should be present for the full set of interviews and the higher frequency of drop-outs from interview 5 indicates the failure to re-interview the households after one year from the first interview. There is also, for interview 5, a high level of intermittent non-response, indicating that although initially failing to interview the household at the 5^{th} interview, the survey manages to re-include some of those units in subsequent interviews. The highest number of intermittent non-response is for those from interview time 1, indicating an initial difficulty in interviewing certain households. It is this category, which includes the patterns with "out" for the first four interviews. These patterns represent 27% of the total intermittent non-response from the first interview, considering the summation over the nine panels in this exercise.

In summary, Table 3.7 shows that, of all household units in the nine panels designed to have the full set of interviews in 2004/2005, 7% drop-out of the sample prematurely, 26% have intermittent non-response (wave non-response) and 66% complete the full set.

3.4 The Selected Working Data Sets

To be able to work with the PME data, two working data sets are defined: one cross-sectional data set described in detail in Section 4.3.1 and a longitudinal data set described in this section.

At the time this research began, micro-data from the years 2004 and 2005 were available to use. The aim is to be able to form a longitudinal data set from the available data. Table 3.2 already showed that some of the panels in the available data included households being interviewed from their fifth interview onwards only. So no information is available for their first four interviews. The idea was to restrict the available data to consider only those households that had data for their starting interview in 2004 or 2005. In this way, the longitudinal data set only contains the households that entered the sample from January 2004. These households are followed over time until the end of 2005.

Table 3.8 gives a representation of PME panels after this first restriction. This forms the general working data set from which a cross-sectional and a longitudinal data set can be selected. Recall that by design in every month there are households being interviewed from the first time to the eighth. Therefore, each column represents an occasion and under them there are the monthly samples that the data will be selected from. For example, from the whole January 2004 sample, only those households being interviewed for the first time are to be considered, whereas from the February 2004 sample, households being interviewed for their first and second times are considered, and so on. Selecting the households that had their first interview from January 2004 onwards allows the construction of a longitudinal series that starts from their first interview.

According to the representation in Table 3.8, different longitudinal data sets could be formed. There is one, for example, where households have the full set of interviews (eight interviews), and another where households have up to the first four interviews. It is worth reinforcing that households which have their last four interviews in the year 2004 were not considered in the working data set. The working longitudinal data set will be formed by only those panels that by design should have the full set of eight interviews, which are the first nine panels in the table.

As discussed in Section 3.3, the matching of individual records is rather difficult. Although the matching of all eight time points is difficult, it is not impossible. In order to fulfil the methodological motivation of analysing a data set containing the full set of interviews from the PME survey it was decided to

Table 3.8: Representation of the Working Data-Set

	Occasions							
Panel	1	2	3	4	5	6	7	8
F4	Jan-04	Feb-04	Mar-04	Apr-04	Jan-05	Feb-05	Mar-05	Apr-05
F5	Feb-04	Mar-04	Apr-04	May-04	Feb-05	Mar-05	Apr-05	May-05
F6	Mar-04	Apr-04	May-04	Jun-04	Mar-05	Apr-05	May-05	Jun-05
F7	Apr-04	May-04	Jun-04	Jul-04	Apr-05	May-05	Jun-05	Jul-05
F8	May-04	Jun-04	Jul-04	Aug-04	May-05	Jun-05	Jul-05	Aug-05
G1	Jun-04	Jul-04	Aug-04	Sep-04	Jun-05	Jul-05	Aug-05	Sep-05
G2	Jul-04	Aug-04	Sep-04	Oct-04	Jul-05	Aug-05	Sep-05	Oct-05
G3	Aug-04	Sep-04	Oct-04	Nov-04	Aug-05	Sep-05	Oct-05	Nov-05
G4	Sep-04	Oct-04	Nov-04	Dec-04	Sep-05	Oct-05	Nov-05	Dec-05
G5	Oct-04	Nov-04	Dec-04	Jan-05	Oct-05	Nov-05	Dec-05	*Jan-06*
G6	Nov-04	Dec-04	Jan-05	Feb-05	Nov-05	Dec-05	*Jan-06*	*Feb-06*
G7	Dec-04	Jan-05	Feb-05	Mar-05	Dec-05	*Jan-06*	*Feb-06*	*Mar-06*
G8	Jan-05	Feb-05	Mar-05	Apr-05	*Jan-06*	*Feb-06*	*Mar-06*	*Apr-06*
H1	Feb-05	Mar-05	Apr-05	May-05	*Feb-06*	*Mar-06*	*Apr-06*	*May-06*
H2	Mar-05	Apr-05	May-05	Jun-05	*Mar-06*	*Apr-06*	*May-06*	*Jun-06*
H3	Apr-05	May-05	Jun-05	Jul-05	*Apr-06*	*May-06*	*Jun-06*	*Jul-06*
H4	May-05	Jun-05	Jul-05	Aug-05	*May-06*	*Jun-06*	*Jul-06*	*Aug-06*
H5	Jun-05	Jul-05	Aug-05	Sep-05	*Jun-06*	*Jul-06*	*Aug-06*	*Sep-06*
H6	Jul-05	Aug-05	Sep-05	Oct-05	*Jul-06*	*Aug-06*	*Sep-06*	*Oct-06*
H7	Aug-05	Sep-05	Oct-05	Nov-05	*Aug-06*	*Sep-06*	*Oct-06*	*Nov-06*
H8	Sep-05	Oct-05	Nov-05	Dec-05	*Sep-06*	*Oct-06*	*Nov-06*	*Dec-06*
I1	Oct-05	Nov-05	Dec-05	*Sep-06*	*Oct-06*	*Nov-06*	*Dec-06*	*Jan-07*
I2	Nov-05	Dec-05	*Sep-06*	*Oct-06*	*Nov-06*	*Dec-06*	*Jan-07*	*Feb-07*
I3	Dec-05	*Sep-06*	*Oct-06*	*Nov-06*	*Dec-06*	*Jan-07*	*Feb- 07*	*Mar-07*

consider a longitudinal data set, which included all eight time points, for as much sample as possible given some criteria for validating the matching. One alternative, however, to be able to work with individual data from the PME is to consider only the heads of household. These are the household reference units and therefore the most important member of the household. In addition, these individuals are easier to identify from the data set. Each household contains only one individual identified as the head of the household and for that similar matching rates as those found for the households are expected. Furthermore, other matching criteria using some of their individual characteristics, such as gender and age, would still be necessary to guarantee a more accurate matching.

With the objective of modelling job earnings, the variable for the usual employment earnings in the main job was chosen as the variable of interest from the other three job earnings variables available in the data. As defined earlier in this chapter, the usual job earnings is the contractual pay received in the reference month and excludes any benefits received. This income component was also used in Jenkins (2000), and for simplicity is adopted here and hereafter referred to as labour income. As also presented earlier in this chapter, IBGE defines the employed population as those who undertook some kind of paid or unpaid work for at least one hour in the reference week. It also includes those who had a job but were temporarily absent from this occupation. Job earnings are defined as the benefit in return for work done and are measured as monthly earnings from the main and second jobs. Therefore, only those classified as employed have data for job earnings. For this reason, the analysis sample is reduced to include only employed heads of household.

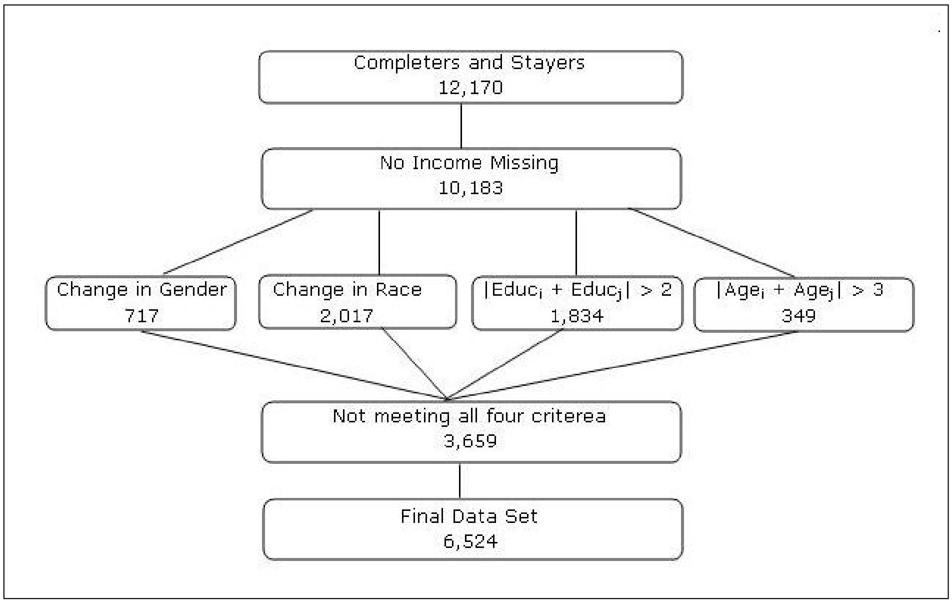
The aim is to select a balanced data set of employed heads of household who were employed at all eight time points starting from the first interview. There are different strategies to select this longitudinal data set and different sets of criteria for validating the matching of individuals. The choice of a complete-case balanced data set is adopted for simplicity since the methods used to analyse this data set are flexible enough to accommodate an unbalanced data set. To guarantee a balanced data set, the heads of household that do not have, by design, all 8 time points were not considered. Those that dropped out of the panel or with intermittent non-response were also not considered. This leaves a total of 12,170 heads of household. Furthermore, only those with valid data for all variables in the analysis (see Tables 4.2 and 5.1) were selected, which leaves 10,183 heads of household.

The set of criteria to validate the matching across the eight time points was chosen based on the exercises performed earlier in this chapter. To ensure that the same head of household is being followed over time, the data set was further reduced to consider only those heads that, from one interview to the next, had:

- no change on the variable for gender;
- no change on the variable for skin colour;
- a change in the categorical variable for education of up to two (ordered) categories ($|Educ_r - Educ_s| \leq 2$);
- and a change in the declared age of up to three years ($|Age_r - Age_s| \leq 3$).

Figure 3.1 presents the sample sizes for the different steps of the selection of the data, including the number of heads of household that did not satisfy the validation criteria. The final sub-set in Figure 3.1, of 6,524 employed heads of household, composes the balanced working longitudinal data set to be used in Chapters 5 and 7. It is worth mentioning that this is a quite restricted data set that was formed to fulfil the methodological motivations of this thesis. This working longitudinal data set might not be appropriate to draw important substantive conclusions.

Figure 3.1: Sample Sizes at Wave 1



3.5 Summary and Discussion

In summary, created in 1980, the PME survey was last revised and implemented with a new methodology from September 2001. It covers the urban areas of the six main metropolitan regions in Brazil and is conducted every month through face-to-face interviews using an electronic questionnaire. It is a probabilistic sample survey with complex design that includes household rotating within a period of 16 months. It has the same rotating panel design as the equivalent U.S. survey, the CPS. As other similar labour force surveys, it follows the recommendations of the ILO in its concepts and definitions and its main objective is to investigate issues related to work in the target population. Data available for this study are from the 2004 and 2005 PME and a selected sub-sample from those years is to be used as the working data set.

Problems with the matching of the monthly PME data are well known. The matching rates were assessed for both household and individual units. The matching rates for the households were similar to those expected by design. Although the comparisons indicated that households in the first four or in the last four consecutive monthly interviews are more easily matched than when households in any of the first four are matched to any of the last four interviews. The lower household matching rates can be partly justified by the pattern of unit non-response that showed that some households interviewed in the first four interviews did not return for the last four or the opposite, as some households were just interviewed in the last four. Exercises performed showed that the matching of individual records is not as accurate as for the household units. When the matching rates of individual records were assessed even lower rates were found. The low rates for individuals are not only influenced by the household unit non-response but also by other factors that include residential mobility and other changes on household structure. Measurement errors on the potential variables that serve to identify the individuals are also cause for the lower rates. The exercises performed for the individual matching identified greater difficulty in matching individuals that were interviewed 9 months or more apart. These exercises indicate that if matching of individuals needs to be performed a considered amount of effort is required to guarantee a small portion of data matched over time. The suggested alternative was to make use of the data of heads of household. This alternative was taken forward and the working data set described in this chapter was restricted to include only the heads of the households.

Chapter 4

Models for Income: Review and Preliminary Analysis

4.1 Introduction

This chapter starts with a review of the main aspects of the Brazilian economy and labour market. A detailed discussion of the studies of job earnings determinants is given. One of the aims of this chapter is to select a model that will serve as the basis for the analysis presented in the following chapters. This chapter then proceeds with a preliminary analysis of the selected data set from the Brazilian labour force survey (the PME) introduced in Chapter 3. With the objective of selecting a model for the determinants of the log of the job earnings, a cross-sectional multilevel model is fitted, only considering the heads of household and the PSU level. An assessment on the significance of the variability between the PSUs is also provided.

It is worth mentioning that throughout the preliminary analysis the PME sampling design is taken into account, whereas, in the multilevel modelling, the sampling weights are not used. The inclusion of the sampling weights and the sampling design variables in the calculation of means and proportions was performed by using the `Stata svy` commands (STATA Press, 2005). The issue of including the sampling weights in a multilevel modelling analysis is raised in Chapters 6 and 7 of this thesis.

4.2 The Brazilian Economy and Labour Market: Brief Review

Brazil is the 5^{th} most populated country in the world. The Brazilian population is very diverse, being formed by native Americans, Africans and Europeans (Smith and Vinhosa, 2002; World Bank, 2003). It is distributed within a vast territory, which differs greatly in living conditions. Brazil is not a poor country: it was ranked the 10^{th} world economy in 2006 according to the total GDP[1]. With GNI[2] per capita of US$5,910 in 2007, Brazil was classified as an *upper-middle-income* economy according to the World Bank (2008). However, Brazil has one of the most unequal economies in the world, with income distribution very skewed to the right (Ferreira et al., 2006; Ferreira, 2000).

Brazil is more unequal than the average Latin American and African countries, behind South Africa only. According to the World Bank (2003), in Brazil the average income of the richest 20% is 33 times the average income of the poorest 20%. Income inequality reduces economic growth and induces high poverty levels which lead to weak social mobility (World Bank, 2003). The Brazilian poverty headcount index in 2004 was 22%[3]. This means that 22% of the Brazilian population had monthly household income per capita below the R$ 100.00 (one hundred *Reais*) threshold.

Poverty and income inequality are closely connected to the labour market and the job earnings (Barros and Mendonça, 1995). Ferreira (2000) stated that the labour market works as a generator of inequalities as it transforms the differences in the innate (such as skin colour, gender and intellectual capacity) and acquired (such as education and experience) characteristics of workers into differences in the wage premium. The Brazilian labour market is segmented, imperfect and inequitable but flexible and efficient (World Bank, 2002). The main segments are the formal and the informal sectors (Barros et al., 2000). Moreover, the Brazilian labour force suffers from discrimination mainly due to differences in skin colour and gender.

The Brazilian income inequality presented a stable distribution over the 80s and the 90s, only showing signs of a decreasing trend in the current decade. This stability was mainly due to the wage distribution which was also stable throughout

[1] GDP: Gross Domestic Product.

[2] GNI: Gross National Income.

[3] The poverty line was of R$ 100.00 which represents the threshold value for the targeting of the social programme *Bolsa Familia* (Ferreira et al., 2006).

the last three decades (Menezes-Filho et al., 2000; Nascimento and Souza, 2005). In an analysis of the determinants of income inequality, Barros et al. (2006) drew attention to its recent fall, in particular, between the years 2001 and 2004. They commented that the current income inequality index is the lowest of the last 30 years although it is still very high. A smaller wage inequality was also identified in Barros et al. (2006). They concluded that this decrease of the income inequality was due to an increase in the income from other sources, mainly from public policies in the form of cash-based programmes (O.E.C.D., 2006; Ferreira et al., 2006), and recent changes in the labour market with the development of social protection networks.

Over the last 30 years, the Brazilian economy went through important macroeconomic changes. Two parallel scenarios were identified in Ferreira and Barros (1999): the first, characterized by the economic growth with stable income and wage distributions but with a decline on income inequality; and the second, characterized by the deterioration of the labour market with small effect on worsening poverty levels. Due to the high inflation rates, there were four currency reforms in this period, and the Brazilian currency changed its name each time. The Brazilian economy changed from being a closed economy (Hay, 2001) to being an open economy (Gonzaga et al., 2005). It is no longer dominated by the agriculture and public manufacturing sectors. Instead, the Brazilian economy is dominated by the private services sector (Ferreira, 2000; Ramos, 2007). The last three decades were also a period of structural changes in the Brazilian population. It initially showed signs of population growth (Smith and Vinhosa, 2002), particularly in the urban areas, but soon it underwent changes in the demographic structure with the fall of the fertility rate. During this period, there was a reduction in the dependency ratio and in the size of families (Barros et al., 2006). An ageing of the population was also observed (Ferreira et al., 2006). There was an increase in the educational level of the population followed by an increase in the returns to higher education level signalling that the Brazilian labour market became more selective (IPEA, 2005). An increase in the informal sector, particularly in the metropolitan areas, was also observed (Passos et al., 2005).

4.2.1 Job Earnings

This section reviews some studies that investigated income determinants based on individual level data using the labour earnings as the income variable. Labour earnings are often used in studies of income determination in the labour market literature (Jenkins, 2000). In Brazil, job earnings are the largest share of the household income per capita. Therefore, it is the most important income component (Ferreira et al., 2006).

Models for labour income are usually estimated based on the Mincer equation, created by Mincer and Polachek (1974). This equation is based on the human capital theory, which measures the quality difference between workers with emphasis on the individual investments, as individuals accumulate human capital during their life cycle (Fernandes, 2002). The human capital theory evaluates the income differential for the different levels of investment. Education is an example of such investment. Although seen as a costly resource, individuals invest in education at an young age expecting pay rises later in life (Heckman and Singer, 1985) and income returns might vary with the individual level of education. The human capital theory is then based upon the productivity capacity of the individuals, and their abilities, innate characteristics and acquired knowledge generate product and income flow (Hanushek, 2006; Scully, 1981).

In a recent study of the Mincer equation, Lemieux (2006) discussed how this equation could be modified. Mincer and Polachek (1974) initially proposed modelling job earnings, expressed in the natural logarithmic scale, as a function of the individual's education and their experience in the labour market. Recognizing that experience in the labour market is not usually observed in the most common empirical labour market data, age is used as a proxy of experience. Lemieux (2006) discussed the use of squared terms for both experience and education. The inclusion of the squared terms are justified as the relationship between education and wages, for example, is not expected to be linear but to be approximated by a squared function and therefore the squared term should be included and tested in the model. While education has an expected convex ("J-shaped") relationship with the response of log-earnings, experience is expected to have a concave (inverted "U-shaped") relationship with the response.

Ferreira and Barros (1999) modelled the household income, instead of job earnings, based on the Mincer equation. Using data from PNAD they performed cross-sectional analyses of four different years. They observed that education (with a squared term) had a convex effect on income while experience had a concave effect with a maximum at 40 years of experience. Barros et al. (2000) performed counterfactual simulations for the log-earnings also using data from PNAD and few more explanatory variables than those suggested by the Mincer equation. They presented what was called *traditional results*:

> *"manufacturing industry and productive services are the economic activities associated with higher wages while agriculture and personal services are those associated with lower wages... employers tend to have a higher income than employees and self-employed. ... Higher wages are for formal ... and informal are those with the lowest wages. ... Labour income tends to be lower in the north-east."*

In a study of wage differentials in the industrial sector, Arbache (2001) used human capital variables expressed by measured and unmeasured abilities, applying different economic theories of wage determination. Education was used as a proxy for workers abilities. Experience was represented in the model by two types of experience variables, one as the experience in the labour market and another as experience in the firm. These two variables had a positive relationship with wage differentials.

Econometric issues to be considered when estimating job earnings equations were raised in Menezes-Filho (2002). Some of these were, for example, endogeneity and causal relationships which can be tackled by using fixed effect models or instrumental variables approach (Hsiao, 2003). This study also had the Mincer equation as the basis using the log-earnings as the response variable. The inclusion of squared terms for education and age were justified: as it is not expected that the returns to education will be the same in all the educational levels; and as returns to experience vary in the life cycle.

Coelho and Corseuil (2002) also reviewed some studies on income differentials. They provided a list of the most used covariates in these studies. These are: individual variables such as educational level, age, skin colour, gender, position in the family or household and region of residency; job variables such as sector of activity, number of working hours, type of worker, job tenure or experience in the firm, use of bonus hours and if workers receive benefits in the wage. They concluded that in most of the studies conducted during the 80s and 90s, education is the main factor for income differential and that the more educated the workforce, the higher is their productivity and hence their income. Experience is also another important factor. The more experience in the firm the workers acquire, the greater is their capacity to implement the job well which increases their efficiency and productivity. They also suggested that living conditions in different regions, as used in Azzoni and Servo (2002), could be controlled for and variables for family background, when possible, should include parental information and information on the other members of the family. They also mentioned that it is important to control for non-productive characteristics such as sex and skin colour, and the studies reviewed showed that the gender wage gap is decreasing and that there are still some wage differentials by skin colour which is sometimes mistaken to be regional differences. In addition, Silveira-Neto and Azzoni (2006) presented some evidence of regional income disparity in Brazil and the existence of two well-defined regional clusters. One formed by the rich south and south-east states and the other, poorer, formed by the north and north-east states. They stated that although the average per capita income in Brazil is relatively high, the regional differential between these two clusters is significant and that these differences have existed since colonization.

Few studies in the literature of the Brazilian labour market investigate income mobility. Nascimento and Souza (2005) related the less variable wage distribution to the stability of income inequality in a study on income mobility using data from PME. In this study, Nascimento and Souza (2005) used fixed effects models and stated that studies including the dynamic analysis of income mobility are still very much incipient in the Brazilian economic literature. In the UK literature, Jenkins (2000) studied income mobility, also reviewing the multivariate modelling framework. He raised some important points about the study of income mobility as it is closely related to movements in and out of poverty. Also recognizing that labour market status is connected to economic welfare, he highlighted the importance of conducting longitudinal studies to assist in the developing of public policies. In Brazil, longitudinal analysis is still incipient, as also mentioned in Corseuil and Santos (2002), and the majority of the studies conduct cross-sectional or pooled cross-sectional analyses.

4.3 Preliminary Analysis

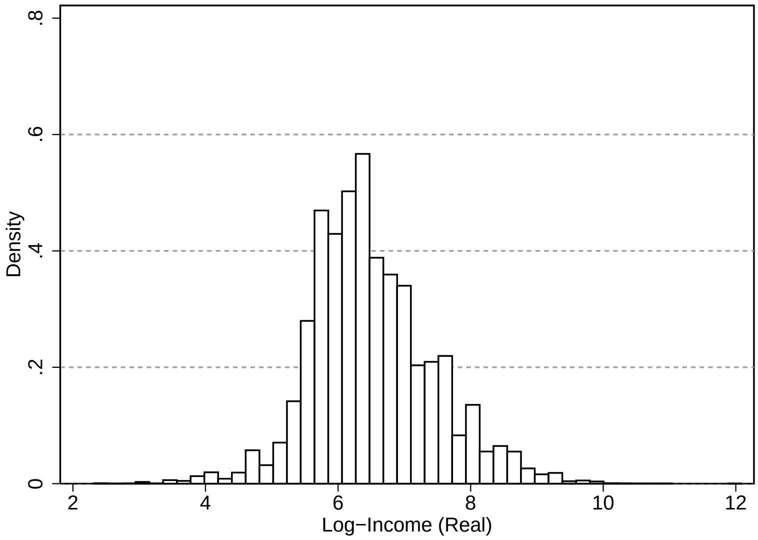
Following the brief review on models for labour income, the next sub-sections present an application of the cross-sectional multilevel model, as described in Section 2.1. The main aim of this preliminary analysis is to select a cross-sectional model for the job earnings that will serve as the basis for further applications in the following chapters of this thesis. A two-level random intercept model is fitted and the significance of the PSU level is assessed. It is worth reinforcing that the longitudinal component of the PME data is not being taken into account in this chapter. The next sub-section describes the PME sample being analysed.

4.3.1 Data Selected for Cross-sectional Analysis

Chapter 3 presented a description of the PME survey and its drawbacks in respect to the matching of households and individual records throughout the survey months. One alternative found in that chapter was to make use of the subset of the heads of household in what was defined as the general working data set. This working data set is formed by pooling data from the 2004 and 2005 PME surveys. In order to conduct a cross-sectional analysis of the PME data, where the number of measurement occasions equals one, the cross-sectional working data set to be used in this chapter only considers the first interview of these heads of households. In this sense, only the PSU and the heads of household levels are accounted for.

The analysis sample in this chapter, therefore, contains the first occasion data of the 57,412 employed heads of household. The sample distribution of the labour income, as expected, was shown to be very skewed to the right. Therefore, the logarithm transformation is applied and Figure 4.1 presents the sample histogram of the log-transformed income. It is worth mentioning that this labour income is expressed in real terms, that is, it is inflated to prices of September 2006[4].

Figure 4.1: Sample Distribution of the Log-transformed Real Labour Income



[4] This was the price index available at the time this data set was constructed. The nominal income was adjusted for inflation relative to the base month of September 2006.

The PME data set has limited choices of individual and household variables. Table 4.1 presents the distribution of the heads of household in the analysis sample according to some of their characteristics and their jobs characteristics. A better explanation of some of the variables in Table 4.1 is presented in Table 4.2 at the end of this sub-section. The results in this table were generated taking the PME sampling weights and the design variables into account. It was performed using `Stata svy` commands (STATA Press, 2005). Note that, due to the way the cross-sectional data was selected, the survey weights available for each head of household are from different survey periods. As the records selected in the working cross-sectional data set are from January 2004 to December 2005, the percent of the population and the average real income, as presented in Table 4.1, are over a 24 month period.

The PME covers the urban areas of the six main metropolitan areas of Brazil. São Paulo is the largest metropolitan area. In the PME, the employment status of the population is investigated for those aged 10 or older. Table 4.1 shows the age distribution of the selected heads of household. They are more concentrated in the age group between 40 and 49 years of age. In addition, employed heads of household are in the majority well educated; 43% have 11 or more years of schooling, which represents the completion of the equivalent of high school or over. The second most frequent education level is for those that have from 4 to 7 years of schooling. This represents attending basic education, but not completing it.

Table 4.1: Distribution of Employed Heads of Households and Average Labour Income by Selected Covariates

	Distribution of employed HoHH (%)	Average Income
Metropolitan Regions		
Recife	6.07	853
Salvador	6.85	994
Belo Horizonte	9.65	1,165
Rio de Janeiro	26.57	1,126
São Paulo	42.07	1,438
Porto Alegre	8.79	1,172
Age Group		
10 to 19	0.37	491
20 to 29	13.47	812
30 to 39	28.68	1,137
40 to 49	30.92	1,368
50 to 59	18.96	1,452
60 or over	7.60	1,439
Gender		
Men	73.30	1,375
Women	26.70	878
Skin Colour		
White	56.47	1,604
Black	9.09	663
Yellow	0.78	2,933
Parda (Mixed)	33.55	765
Indigenous	0.11	749
Collapsed Skin Colour		
Others	43.53	783
White	56.47	1,604
Education in Years of Schooling		
Less than 1	3.56	415
1 to 3 years	7.25	496
4 to 7 years	27.92	613
8 to 10 years	17.97	757
11 years or over	43.09	2,089
Not defined	0.21	501
Occupation		
Employee	68.30	1,167
Self-Employed	24.15	911
Employer	7.43	3,095
Unpaid Workers	0.11	-
Informality		
Formal	71.73	1,189
Informal	28.27	786
Employees in the		
Private Sector	82.33	1,144
Public Sector	17.67	1,781
Sector & Informality		
Formal Private Sector	62.92	1,216
Informal Private Sector	19.42	910
Formal Public Sector	3.23	1,540
Informal Public Sector	1.61	1,242
Military Services	12.83	1,918

Table 4.1 – continued from previous page

	Distribution of employed HoHH	Average Income
Activities		
Manufacturing	18.60	1,328
Building	9.93	824
Commerce	18.71	1,055
Financial	13.84	1,822
Social Services	13.88	1,771
Domestic Services	6.31	359
Other Services	18.00	1,082
Other Activities	0.73	913
Duration of Employment		
Average duration in months	98	-
Hours Worked		
Average hours worked per week	44	-
Average Number of Household Members	3	-
Total	-	1,243

A recent study from IBGE pointed to an increase in the selection of women as the reference person in the household, with this proportion being almost 30% in August 2006 (IBGE, 2006b). Table 4.1 shows that the proportion of female heads of household in the analysis sample is 26.7%. The variable that measures skin colour in the PME is self-declared and classifies the population into five sub-groups: White, Black, Yellow (Asian descendants), Mixed and Indigenous. Table 4.1 shows that more than half of the heads of household classify themselves as white. A common practice adopted in studies of the Brazilian population is to create a dichotomous variable that represents white against others, collapsing all other categories. This is also presented in Table 4.1.

On job characteristics, Table 4.1 shows that these heads of household are in the majority employees (68%) or self-employed (24%). There is a quite small share of unpaid workers and only 7% are employers. The share of the employee heads of household can be further classified as formal (72%) and informal (28%) employees. Furthermore, Table 4.1 shows that few employees are in the public sector. The majority are formal employees in the private sector and around 13% are in military service. These characteristics are, however, only for employees.

Looking again at all employed heads of households, Table 4.1 shows that they are mostly engaged in the commerce and manufacturing activities as well as other services. Table 4.1 also presents average values for some continuous characteristics for this sub-set of employed heads of household, such as: their average duration of employment is of 98 months and they work on average 44 hours per week. The last variable presented in Table 4.1 is the total number of household members. This variable is discrete with an average of 3 members in the household.

Table 4.1 also presents the average labour income according to these same characteristics. It can be observed from this table that employed heads of household in the Southern metropolitan regions - Belo Horizonte, Rio de Janeiro, São Paulo and Porto Alegre - have higher average income than those in the northeast regions. São Paulo is the metropolitan region with higher average income. There is income differential for gender and skin colour: female heads of household have lower average income than males and whites higher than others. Older heads of household have higher average income as well as employers, and informal employees earn less than formal ones. Heads of households employed in the public sector have higher average earnings than those in the private sector; both formal and informal heads of household earn more in the public sector than in the private sector. However, those in the military services have the highest earnings on average. Heads of household engaged in financial activities earn on average more than on any other activity. To conclude, Table 4.1 shows that education returns are as expected, with higher average income for the best educated.

Table 4.2 presents the set of variables considered as explanatory variables in the models to follow. These variables were selected based on the review presented earlier in this chapter. These are almost all the variables from Table 4.1, excluding those that were applied to employee heads of household only. Furthermore, the variable for type of worker now breaks the employee category into two: formal and informal heads of household. Table 4.2 provides further explanation for each of the variables. In summary they are: interview month, dummies for male and white, age, education, number of members in the household, type of worker, type of activity, metropolitan region, duration of employment and working hours. There is also a dummy indicating whether the heads of household had their questionnaire completed by another member of the household and not themselves. Furthermore, squared terms for age and education are also considered. It is worth mentioning that, in this analysis sample, the income variable contains 4.8% of missing values. Therefore, heads of household with missing income are not included in the analysis that follows. Also note that the small sub-set of unpaid workers, which have zero labour income, is not included either. This brings the final data set to a total of 54,663 heads of households.

Table 4.2: Variables Included in the Analyses

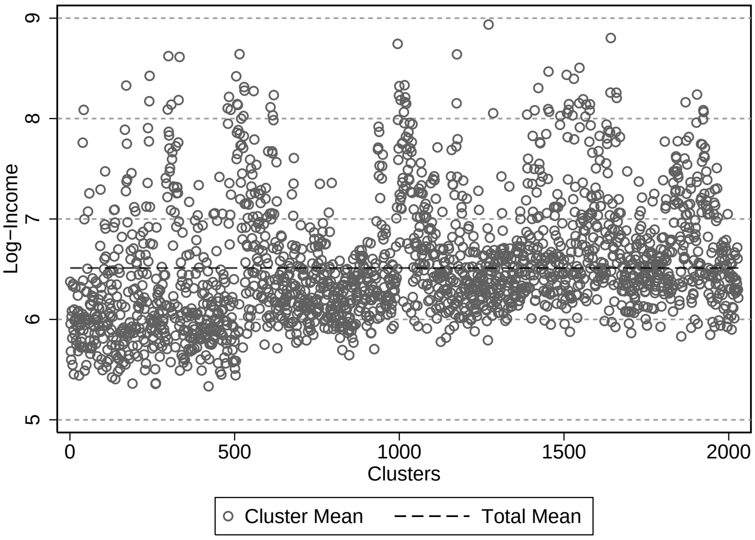
Variables	Categories	Definition
Metropolitan Regions	Recife Salvador Belo Horizonte Rio de Janeiro São Paulo Porto Alegre	This is a cluster level variable. It is the stratifying variable used in the design of the PME. This variable is included in the model as a control variable. Recife is the baseline category.
Month of First Interview		Represented in the model as a continuous variable from 0 to 23 starting with January 2004 till December 2005.
Gender	Females Males	An indicator for male heads of household is considered in the model.
Age Group		It is used in a continuous form. The squared term is also considered. Under the mincer equation age is a proxy for experience. In the multilevel model age is centred around 40.
Skin Colour	White Black Yellow Parda (Mixed) Indigenous	This categorical variable is later collapsed to represent an indicator for whites against all other categories, referred to as others.
Education in Years of Schooling		It is used in a continuous form from 0 to 17 years of schooling. The squared term is also considered.
Type of Worker	Employer Informal Employee Formal Employee Military Service Self-Employed	Employer is the baseline category. Employees are now separated into formal and informal employees.
Type of Activity	Manufacturing Building Commerce Financial Services Social Services Domestic Services Other Services Other Activities	Manufacturing is the baseline category.
Duration of Employment		Continuous variable in months. Serves as a proxy for experience in the firm.
Working Hours		Working hours expressed on the natural logarithm scale.
Proxy Respondent		This variable is used as a control for hard to count heads of household.
Number of Household Members		This variable is used as a control for change in the household structure.

4.3.2 Methodology

4.3.2.1 Cross-sectional Multilevel Modelling

The multistage nature of the PME sample determines that heads of household are nested within PSUs. The multilevel modelling approach (Goldstein, 2003) accounts for such data complexity, modelling the different sampling stages as different sources of variability with potential random influences. Figure 4.2 presents the average of the log-income across PSUs. Notice that the PSU averages vary considerably across clusters. A random intercept model, as described in Chapter 2, would be a reasonable starting model formulation to account for the variation between PSUs.

Figure 4.2: Average of Log-Income across PSUs



Consider the two-level random intercept model where heads of household are the level one units (subscript i), which are nested within PSUs, the level two units (subscript j):

$$y_{ij} = \boldsymbol{x}_{(1)ij}^T \boldsymbol{\beta}_{(1)} + \boldsymbol{x}_{(2)j}^T \boldsymbol{\beta}_{(2)} + u_j + e_{ij}. \tag{4.1}$$

In model 4.1 the outcome variable y_{ij} is the continuous variable for the logarithm of real labour income of the employed heads of household. This is modelled as a function of covariates for each of the two levels considered in this analysis. The vector $\boldsymbol{x}_{(1)ij}$ contains the explanatory variables at the heads of household level. This includes categorical and continuous variables as well as interaction terms. The vector $\boldsymbol{x}_{(2)j}$ contains the explanatory variables at the PSU level. Each of these vectors of covariates is associated with the respective vector of fixed regression coefficients $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$. As this is a two-level random intercept model, the random part of the model contains two mutually independent residuals terms: u_j, at the PSU level (or PSU random intercepts) and e_{ij}, at the heads of household level. As also described in Chapter 2, here it is assumed that both residual terms are normally distributed with mean zero and respective variances σ_u^2 and σ_e^2. This model further assumes that heads of household within the same PSU are conditionally correlated. This conditional correlation is expressed by the intra-cluster correlation coefficient, see equation 2.5 in Chapter 2.

4.3.3 Computational Aspects

Random intercept models can be fitted using different statistical software, which, when the right options are specified, can all produce the same results. For example, in `Stata` there are the commands `xtreg` and `xtmixed` that produce Maximum Likelihood Estimation when the option MLE is specified (Rabe-Hesketh and Skrondal, 2005). However, the `xtreg` command does not allow more than two levels of data hierarchy. The `Stata` command `GLLAMM` (Generalized Linear Latent and Mixed Model) (Rabe-Hesketh et al., 2004) can also produce similar results while using numerical integration methods but it needs longer computational time than other `Stata` commands. This model can also be estimated using the MLwiN software (Rasbash et al., 2001), which uses IGLS to estimate the random intercept model also producing ML estimates under the assumption of normality for the distribution of the residuals.

4.3.4 Results

4.3.4.1 Model Selection

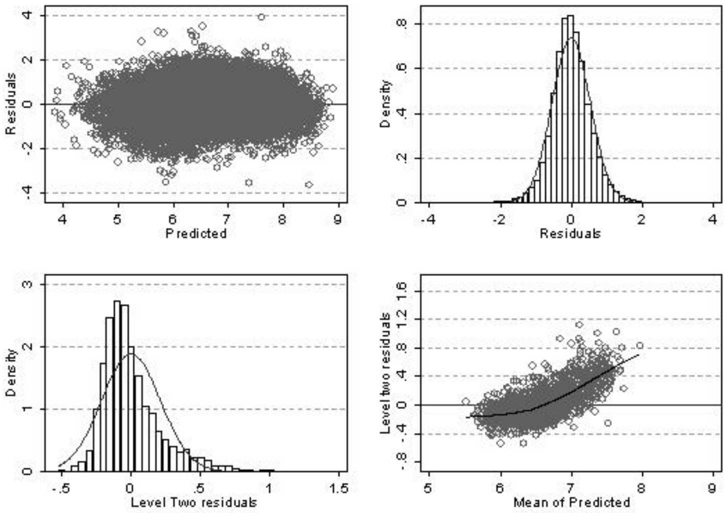
The first step of the cross-sectional multilevel analysis was to estimate the model in equation 4.1 considering only the set of fixed main effects as explanatory variables. The covariates included in the model were those described in Table 4.1 and their effects were assessed in terms of their significance. It is worth mentioning that the variable for metropolitan region is the only cluster level covariate included at this stage as it is an important variable for the estimation of labour income in Brazil. All other variables considered at this stage were level one covariates.

The first column of Table 4.3 presents the results for this first model estimated using the Stata command `xtmixed`. Notice that the variables for the number of household members and proxy respondent variables were not significant at the 5% level, and therefore not included in the model. It is worth remembering that the variables for age and duration of employment are now centred respectively around the age of 40 and the average employment duration.

Table 4.3 also presents the estimated variance components for this model. Both variance terms are statistically significant. The between PSU variance $\hat{\sigma}_u^2$ of 0.05 is relatively small compared to the within PSU variance $\hat{\sigma}_e^2$ of 0.30, but still significantly different to zero. This gives an estimated intra-cluster correlation coefficient $\hat{\rho}$ of around 15%. This indicates the conditional correlation within cluster and that the PSU level should be considered in this analysis.

Before attempting to interpret this initial model, the residual diagnostics is performed. Figure 4.3 displays the level one residual in the first row and level two residual in the second row.

Figure 4.3: Residual Diagnostic - Main Effects Model



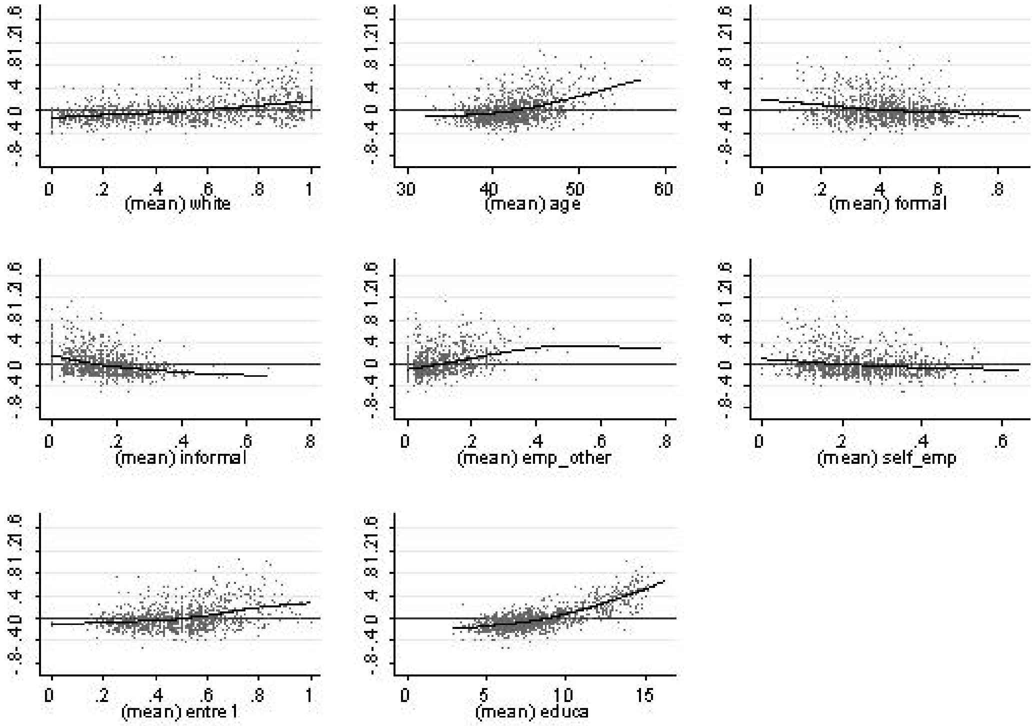
Observe that the residuals at the heads of household level seem to be normally distributed. However, level two residuals show the presence of some extreme positive values. This may indicate the violation of the normality assumption. The fourth plot shows some evidence of non-constant variance for the PSU level residuals. This might also indicate the presence of unmeasured cluster effects or that the level two residuals are correlated with explanatory variables. The inclusion of additional PSU level variables or contextual effects may address this problem.

As identified earlier in this chapter, there is some evidence of gender and race discrimination in the Brazilian labour market. Before the inclusion of contextual variables, the model selection proceeded with the inclusion of interaction terms between level one variables. For this purpose, only the interaction terms between the dummy variables for males and whites and all other variables were considered and tested in the model. The column (2) of Table 4.3 presents the results for the model fitted with only the significant interaction terms. The inclusion of such terms improved the model fit, as indicated by the LRT ($L^2 = 556.43$ with 19 degrees of freedom). However, the residual diagnostic plots for this model presented the same patterns as those of the previous model.

To try to improve the fit of this model the next step was to include contextual effects. Unfortunately, PSU level variables, other than the metropolitan region, were not available in the data set. Due to confidentiality protection such variables are not immediately available in any of the official surveys. One alternative found was to construct PSU level variables from the monthly PME data. This was performed by pooling data from years 2004 and 2005 for all interviewed individuals. Population means and proportions for specific variables were calculated taking the sampling design and sampling weights into account for each PSU. For simplicity, the contextual variables were calculated for the variables initially considered as covariates in the model. Before deciding which of these variables to include in the analysis, the level two residuals were plotted against the average PSU values of the explanatory variables in the cross-sectional data set. Some of these plots are presented in Figure 4.4 and similar behaviour to that in the fourth plot on Figure 4.3 was observed. Further model selection was performed and Figure 4.4 presents the plots for the significant contextual variables in the column labelled (3) of Table 4.3.

The column labelled (3) of Table 4.3 presents the final cross-sectional multilevel model for the log of real labour income of employed heads of household. Figure 4.5 presents the residual diagnostics plots for this model. Notice that the inclusion of PSU level variables improved the shape of distribution of the level two residuals. This model can then be interpreted, starting from the estimates of the fixed part of the model as presented in the following sub-section.

Figure 4.4: Level Two Residuals by Significant Contextual Effects



Note: Level two residuals on the vertical axis.

Figure 4.5: Residual Diagnostics - Contextual Effects Model

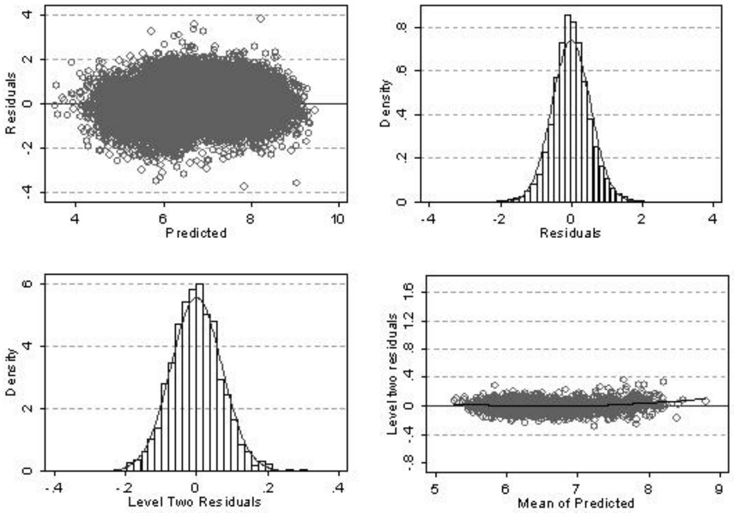


Table 4.3: Cross-sectional Multilevel Modelling: Two-level Variance Components Model

	(1)		(2)		(3)	
	Coeff	SE	Coeff	SE	Coeff	SE
Constant (intercept)	4.334	0.038	4.319	0.056	4.753	0.162
Month	-0.004	0.001	-0.004	0.001	-0.004	0.001
squared term	$0.213^{\dagger}$	$0.056^{\dagger}$	$0.205^{\dagger}$	$0.056^{\dagger}$	$0.216^{\dagger}$	$0.055^{\dagger}$
Males	0.379	0.006	0.598	0.070	0.594	0.070
White	0.109	0.006	0.092	0.010	0.071	0.010
Age	$5.333^{\dagger}$	$0.272^{\dagger}$	$2.293^{\dagger}$	$0.513^{\dagger}$	$1.427^{\dagger}$	$0.509^{\dagger}$
squared term	$-0.351^{\dagger}$	$0.015^{\dagger}$	$-0.293^{\dagger}$	$0.028^{\dagger}$	$-0.311^{\dagger}$	$0.027^{\dagger}$
Education	-0.011	0.002	-0.029	0.004	-0.030	0.004
squared term	$5.711^{\dagger}$	$0.133^{\dagger}$	$6.454^{\dagger}$	$0.231^{\dagger}$	$6.055^{\dagger}$	$0.229^{\dagger}$
Type of Worker						
(Employer as baseline)						
Informal	-0.545	0.012	-0.615	0.027	-0.593	0.027
Formal	-0.360	0.010	-0.435	0.025	-0.412	0.025
Military service	-0.247	0.015	-0.369	0.030	-0.339	0.029
Self-Employed	-0.635	0.011	-0.806	0.026	-0.783	0.026
Type of Activity						
(Manufacturing as baseline)						
Building	-0.069	0.010	0.181	0.052	0.180	0.052
Commerce	-0.114	0.008	-0.001	0.017	-0.001	0.017
Financial	-0.024	0.009	0.154	0.019	0.147	0.019
Social Services	-0.044	0.011	0.082	0.018	0.081	0.018
Domestic Services	-0.135	0.012	-0.067	0.017	-0.059	0.017
Other Services(a)	-0.025	0.008	0.035	0.017	0.031	0.016
Other Activities(b)	-0.272	0.027	-0.080	0.082	-0.088	0.081
Metropolitan Region						
(Recife as baseline)						
Salvador	0.116	0.023	0.117	0.023	0.077	0.016
Belo Horizonte	0.310	0.021	0.311	0.021	0.241	0.014
Rio de Janeiro	0.311	0.021	0.313	0.021	0.218	0.015
São Paulo	0.461	0.021	0.460	0.021	0.342	0.016
Porto Alegre	0.362	0.022	0.360	0.022	0.224	0.021
Duration of Employment (×120)	0.216	0.005	0.239	0.009	0.234	0.009
Squared term (×120)	-0.049	0.002	-0.068	0.005	-0.068	0.005
Working Hours (in Log)	0.459	0.008	0.494	0.011	0.498	0.011
Interaction Terms of Male and:						
White	-	-	0.024	0.011	0.024	0.011
Age	-	-	0.004	0.001	0.004	0.001
Squared term	-	-	0.000	0.000	0.000	0.000
Education	-	-	0.026	0.005	0.026	0.005
Squared term	-	-	-0.001	0.000	-0.001	0.000
Type of Worker						
(Employer as baseline)						
Informal	-	-	0.074	0.030	0.072	0.030
Formal	-	-	0.079	0.027	0.080	0.027
Military service	-	-	0.167	0.035	0.166	0.034
Self-Employed	-	-	0.217	0.028	0.215	0.028
Type of Activity						
(Manufacturing as baseline)						
Building	-	-	-0.298	0.053	-0.286	0.053
Commerce	-	-	-0.152	0.019	-0.156	0.019
Financial	-	-	-0.239	0.022	-0.240	0.021
Social Services	-	-	-0.196	0.023	-0.199	0.023
Domestic Services	-	-	-0.312	0.037	-0.312	0.037

Table 4.3 – continued from previous page

	(1)		(2)		(3)	
	Coeff	SE	Coeff	SE	Coeff	SE
Other Services	-	-	-0.086	0.019	-0.085	0.019
Other Activities	-	-	-0.257	0.087	-0.241	0.086
Duration of Employment (×120)	-	-	-0.034	0.010	-0.034	0.010
squared term (×120)	-	-	0.027	0.005	0.027	0.005
Working Hours (in Log)	-	-	-0.094	0.016	-0.092	0.016
Contextual Effects						
Proportion of White	-	-	-	-	0.104	0.029
Average Age	-	-	-	-	0.009	0.002
Proportion of Formal Workers	-	-	-	-	-1.324	0.111
Proportion of Informal Workers	-	-	-	-	-1.606	0.131
Proportion of Military Workers	-	-	-	-	-1.866	0.124
Proportion of Self-Employed Workers	-	-	-	-	-1.574	0.125
Proportion with Proxy Respondent	-	-	-	-	0.157	0.040
Average Education	-	-	-	-	0.073	0.004
$\hat{\sigma}_u^2$	0.055	0.002	0.055	0.002	0.011	0.001
$\hat{\sigma}_e^2$	0.301	0.002	0.298	0.002	0.297	0.002
$\hat{\rho}$	0.154		0.156		0.035	
Number of Observations	54,663		54,663		54,663	
-2× Log-Likelihood	93,056		92,500		90,084	

(1) Model with level one main effects and metropolitan region variable.
(2) Model adding interaction terms.
(3) Model adding other contextual variables.
(a) Other Services include services as post offices, housing, food, personal, urban cleaning and aerial transportation.
(b) Other Activities include all other activities not yet classified, such agriculture, fishing, forestry, international organizations and non specified activities.
† Values at 10^{-3}.

4.3.4.2 Model Interpretation

Note from Table 4.3 that the indicator for male heads of household interacts with almost all level one explanatory variables. Hence, the interpretation of the effects of these variables in the interaction terms accounts for the differences between male and female heads of household and are displayed in Table 4.4. The effects for age, education and duration of employment on income, accounting for the interaction with the indicator for male heads of household, are displayed in Figures 4.6(b) to 4.6(d). Figure 4.6(a) presents the effect of month of the interview on income. Based on Table 4.4 and Figures 4.6(a) to (d) the following is observed from the level one fixed effects.

The effect of the month of the first interview on the real labour income is displayed in Figure 4.6(a). This effect starts as a negative effect on income, having the lowest point just before month 10, which then starts increasing and becoming a positive effect in later months of 2005. The relationship presented in Figure 4.6(a) may be indicative of an increase in the job earnings in real terms of employed heads of household over the period of analysis.

The interaction effects between the indicator for males with the other variables indicate that there are income differentials between male and female heads of household. At the baseline the income differential is 81% in favour of male heads of household. White heads of household earn more and this effect is greater for males (9.99%) than females (7.37%), with all else held constant.

Table 4.4: Percentage Impact on Average Real Labour Income - Interaction Terms

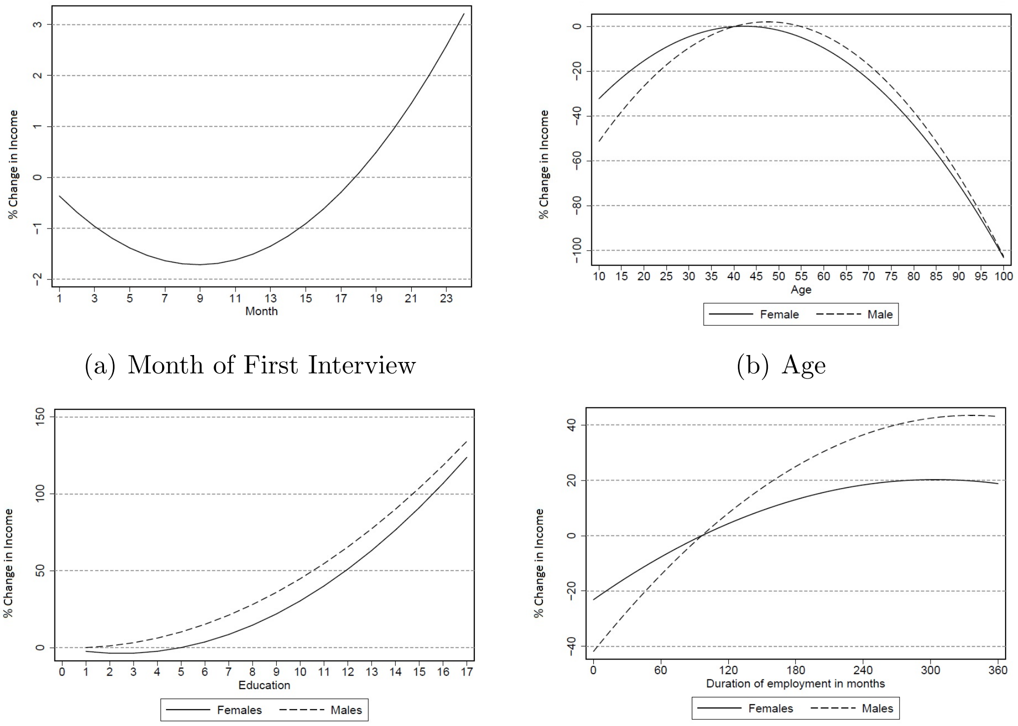
	Females	Males
Sex (at the baseline)	-	81.06
White	7.37	9.99
Type of Worker (Employer as baseline)		
Informal	(44.71)	(40.57)
Formal	(33.77)	(28.26)
Military service	(28.77)	(15.95)
Self-Employed	(54.28)	(43.30)
Type of Activity (Manufacturing as baseline)		
Building	19.73	(10.05)
Commerce	(0.06)	(14.45)
Financial	15.79	(8.95)
Social Services	8.46	(11.13)
Domestic Services	(5.69)	(30.93)
Other Services	3.18	(5.19)
Other Activities	(8.44)	(28.05)
Working Hours (in Log)	64.53	50.11

Note: Values in parentheses indicate % decrease.

Following suggestions observed in the reviewed literature on models for job earnings presented earlier in this chapter, squared terms for age and education variables were also considered. These effects were tested and showed to be significant, and, therefore, kept in the model. This confirms that the effects of age and education on income are not linear, as presented in Figures 4.6(b) and (c) respectively. As described in Table 4.2, age is centred around age 40 years and the effect is zero at this age. Figure 4.6(b) shows that the impact of age on income has an inverted U-shape. Notice the different curves for males and females showing that there is a higher impact on the income for male heads of household older than 40, than for female heads of household. Figure 4.6(c) shows the impact of education. Education returns are as expected: the more educated the head of the household the higher the income, showing a J-shape relationship. Furthermore, the impact of education on income for male heads of household is higher than for female heads of household.

Table 4.4 shows that employers earn more than any other type of worker as already identified in Barros et al. (2000). The income differential between

Figure 4.6: Impact of Month of First Interview, Age, Education and Duration of Employment on Income



(a) Month of First Interview

(b) Age

(c) Education

(d) Duration of Employment

males and females varies for each type of worker. Compared to employers, female informal workers earn around 45% less with all other variables held constant, while male informal workers earn around 41% less. There is always a greater income differential for female heads of household in any other category for type of worker compared to employers.

The gap between female and male heads of household engaged in certain activities tends to get narrower, as shown in Table 4.4. For example, female heads of household engaged in activities in the financial sectors, building, social services or other services earn, respectively, around 16%, 20%, 8% and 3% more than those in the manufacturing sector (holding all else constant). Male heads of household, in turn, earn more in the manufacturing sector than any other. This result, for male heads of household, is also in agreement with the *traditional result* as presented earlier in this chapter. In addition, those working in other activities or in the domestic services are the less well paid activities for the heads of household.

Duration of employment also has a non-linear relationship with income as shown in Figure 4.6(d). It is also worth recalling that duration is centred around its sample mean (around 96 months for the data set being considered). The effect of the duration of employment on income for those working more than the average duration is positive with males having a greater increase in income than females.

Working hours is considered in the model on the logarithmic scale. Table 4.4 shows that the impact on income of a unit change in the log hours is greater for female heads of household than for male heads of household. Because this variable is on the log scale it could also be interpreted as the income elasticity with respect to working hours. The real labour income increases by 0.5% for female heads of household when the working hours increase by 1% according to the results in Table 4.3. This effect for male heads of household is slightly lower at around 0.41%.

The variable for metropolitan region is a level two variable, as every head of household in the same PSU belongs to the same metropolitan region. This variable is actually the first stratification variable in the PME sampling design. Using equation 2.11 in Chapter 2 and the values displayed in Table 4.3 it is possible to conclude, for example, that heads of household in the metropolitan region of Salvador earn 8% more than those in the Recife (holding all other variables constant). Furthermore, notice that heads of household in the South-east metropolitan regions earn around 25% more, while the highest income differential is for those living in the metropolitan region of São Paulo, at around 40%. This is in accordance with Silveira-Neto and Azzoni (2006).

Besides metropolitan region there are another eight significant PSU variables. Table 4.5 presents these estimated effects using equation 2.12 from Chapter 2. It is worth mentioning that the proportion of different types of workers is being compared to the proportion of employers, hence, the negative effects. For example, it can be observed that, there will be an increase of about 1.05% in the average income for a 10% increase in the proportion of white heads of household. Another possible interpretation is that heads of household in PSUs with a greater proportion of white heads of household have higher average real income than those with lower proportions. The variables that represent averages do not have the same interpretation. Instead, for example, a higher mean PSU education level has a positive effect on the income of heads of household. As for age, heads of household in PSUs with older heads of household (more experienced) have slightly higher average labour income.

Table 4.3 also presents the results for the random part of the model, given by the variance effects for level one and level two. The estimated intra-cluster correlation coefficient $\hat{\rho}$, interpreted as the correlation between two randomly selected heads of household in the same PSU, is 3.45%. This is smaller than those calculated for the previous models, mainly because this last model included more

Table 4.5: Percentage Impact on Average Real Labour Income - Contextual Variables

Contextual Effects	$\hat{\beta}$	$\hat{\beta}a^1$	b%
Proportion of White	0.10	0.01	1.05
Average Age	0.01	0.01	0.88
Proportion of			
Formal Workers	(1.32)	(0.13)	(12.40)
Informal Workers	(1.61)	(0.16)	(14.83)
Military Workers	(1.87)	(0.19)	(17.02)
Self-Employed Workers	(1.57)	(0.16)	(14.57)
Proportion with Proxy Respondent	0.16	0.02	1.58
Average Education	0.07	0.07	7.51

[1] $a = 1$ for means and 10% for proportions

PSU level covariates. These estimates for the variance terms also indicate that for these data there is a greater within PSU variability, with $\hat{\sigma}_e^2$ equal to 0.297, than between PSUs variability, where $\hat{\sigma}_u^2$ equals to 0.01. Although still relatively small, the between PSU variance term is statistically significant indicating the importance of accounting for the PSU level.

4.4 Summary

This chapter presented a brief review of the recent trends in the Brazilian economy, focussing attention on some recent studies on wage determination. Most of the reviewed studies had the Mincer equation model as a basis. This chapter also presented an analysis of a cross-sectional data set from the PME survey. For that, the first interviews for employed heads of household were considered. A preliminary analysis of this sub-set was presented followed by a multilevel model analysis. A more elaborate model than that suggested by the Mincer equation was selected for the log of the labour income. This cross-sectional multilevel analysis accounted for the PSU and the heads of household level. It was shown that in this analysis sample, the between PSU variability is small relatively to the within PSU variability, however, still significant. This indicates the importance of accounting for the PSU level in the longitudinal analyses to follow. The final model presented here served as basis for the modelling exercise in the following chapters.

Chapter 5

Longitudinal Multilevel Modelling Accounting for the Rotating Design

5.1 Introduction

As discussed in Section 3.2 (Subsection 3.2.2) the Brazilian labour force survey (the PME) has a non-consecutive, but symmetric, rotating panel design characterized as 4-8-4 (*a*-*b*-*a* as described in the introductory chapter of this thesis). This means that the selected sample units stay in the sample for four consecutive months, are rotated out for eight months and return to the sample for another four consecutive months. Therefore, due to this rotation pattern there will be a gap of eight months between the fourth and fifth interviews for every head of household in a longitudinal data set. This chapter aims to illustrate how to incorporate the rotation pattern in the analysis of a longitudinal data set under the multilevel modelling framework.

To fulfil the main objective of this chapter the longitudinal working data set is considered. This data set includes all eight interviews for the employed heads of household and was described in Section 3.4. This longitudinal data set is analysed by means of growth curve models and multivariate multilevel models as presented in Section 5.2. The analysis of the longitudinal data set should take the gap between the fourth and fifth interviews into account. This can be seen as a similar problem to modelling unequally spaced time data. Although all the measurements are taken monthly, the time distance between the fourth and the fifth interviews, for example, is not of one but eight months. One way to account for this larger interval between these two interviews, when treating time as a continuous variable, as in the random slope models, is to consider it varying as $(0, 1, 2, 3, 12, 13, 14, 15)$ instead of from 0 to 7. However, this is irrelevant when time is treated as a discrete variable as in the multivariate multilevel models. When this is the case, one approach that accounts for the gap is to constrain the error covariance matrix to depend on the temporal distance, i.e. the lag, between the different measures as in Yang et al. (2002). Different structures for the error covariance matrix are discussed in this chapter and extended to a general rotation scheme, and models with the different structures are fitted and compared. This chapter concludes with further discussion of alternative correlation structures and extensions to this analysis to be considered.

5.2 Methodology

This section presents how the growth curve model and the multivariate multilevel models can be used to analyse data provided by rotating panel designs. The growth curve model accounts for this design by treating time as a continuous variable. The multivariate multilevel model, which treats time as discrete, accounts for this gap by constraining the error covariance matrix to depend on the temporal distance between the interviews.

5.2.1 Growth Curve Models

Consider the three-level longitudinal data for the PME survey. Also consider the non-consecutive PME rotating design of 4-8-4. One way to express the variable for time would be to represent the interview times as $(1, 2, 3, 4, 5, 6, 7, 8)$; or better $(0, 1, 2, 3, 4, 5, 6, 7)$ so as to improve interpretation. However, this does not account for the gap of eight months that exists between the fourth and fifth interviews. Consider then that the variable for time should be expressed to vary from 0, as for the first measurement occasion, to $d - 1$, that is the total time span minus 1. For the PME case time varies from 0 to 15, but due to the gap the vector for time can be expressed as $(0, 1, 2, 3, 12, 13, 14, 15)$.

Consider the growth curve model with only the time variable as a covariate:

$$y_{tij} = \boldsymbol{x}_{(1)tij}^{T}\boldsymbol{\beta}_{(1)} + v_j + \boldsymbol{z}_{tij}^{T}\boldsymbol{u}_{ij} + e_{tij}\ . \tag{5.1}$$

This is a three-level random slope model and time has both fixed and random effects. It is worth remembering that the three levels of the data hierarchy are the occasions (level one units, subscript t) which are nested within heads of household (level two units, subscript i) which are nested within PSUs (level three units, subscript j). The model in equation 5.1 has a similar formulation to the one described by equation 2.21. As it is a three-level model it has three error terms and the assumptions about their distributions, made previously, hold here.

With the objective of accounting for the rotating panel design, the matrix of explanatory variables, which is associated with the vector of random effects, is

$$Z_{ij} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 12 \\ 1 & 13 \\ 1 & 14 \\ 1 & 15 \end{pmatrix}.$$

The first column of Z_{ij} represents the random intercept and the second column the values for the time variable. The gap between the fourth and fifth interviews is accounted for in the model by declaring the time variable as $(0, 1, 2, 3, 12, 13, 14, 15)$, as mentioned before. One characteristic of the random slope model is that by treating time as a continuous variable it accommodates unequally spaced longitudinal data. Therefore, it allows the gap to be incorporated in the analysis in this way, once the model depends on the measurements of the time variable.

Under this model formulation, the blocks V_j of the covariance matrix of the composite residuals V no longer have an exchangeable structure, as when only the random intercept is considered. The matrices V_j can be expressed as the sum of the between PSU variation, σ_v^2, the between heads of household variation, $\boldsymbol{z}_{tij}^T \Sigma_u \boldsymbol{z}_{tij}$ and the between occasions variation, σ_e^2, where

$$\Sigma_u = \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix}.$$

The main diagonal of V_j are the variance terms, which are given by

$$Var(y_{tij}|\boldsymbol{z}_{tij}) = \sigma_v^2 + \boldsymbol{z}_{tij}^T \Sigma_u \boldsymbol{z}_{tij} + \sigma_e^2$$

and the off-diagonals are the covariance terms given by

$$Cov(y_{rij}, y_{sij} | \boldsymbol{z}_{rij}, \boldsymbol{z}_{sij}) = \sigma_v^2 + \boldsymbol{z}_{rij}^T \Sigma_u \boldsymbol{z}_{sij}.$$

This imposed form of the residual covariance matrix is a quadratic function of time which is determined by the matrix of between heads of household variation $\boldsymbol{z}_{tij}^T \Sigma_u \boldsymbol{z}_{tij}$. In general form it can be given as

$$\gamma_{rs} = \sigma_{u0}^2 + (t_r + t_s)\sigma_{u01} + t_r t_s \sigma_{u1}^2,$$

where r and s are different occasions so that t_r and t_s represent values for the time variable (Singer and Willett, 2003).

Another more flexible extension of this growth curve model is the one which includes polynomial functions for the time variable. For example, the model could include quadratic or cubic terms for time with both fixed and random effects. However, this generates an even more difficult to interpret structure for the residual covariance matrix. If the squared term for time is also considered in the model, the general form for the elements of the between heads of household variation matrix is given as

$$\gamma_{rs} = \sigma_{u0}^2 + t_r t_s \sigma_{u1}^2 + t_r^2 t_s^2 \sigma_{u2}^2 + (t_r + t_s)\sigma_{u10} + (t_r^2 + t_s^2)\sigma_{u20} + t_r t_s (t_r + t_s)\sigma_{u21},$$

where Σ_u is now given as:

$$\begin{pmatrix} \sigma_{u0}^2 & & \\ \sigma_{u10} & \sigma_{u1}^2 & \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 \end{pmatrix}.$$

Models including the squared term for time will be tested in this chapter.

Because the structure of the residual covariance matrix depends on the values for the time variable, an analysis of this same data but treating time as varying from 0 to 7, would provide different results. This is illustrated later in this chapter. Treating time as varying from 0 to 7 in the case of the PME data is the same as ignoring the gap in the rotating design. This is the same as assuming that each interview was taken one month apart from each other including the fourth and the fifth interviews, which is not true for the PME design.

5.2.2 Multivariate Multilevel Models

Consider the multivariate multilevel model with only the time variable as a covariate:

$$y_{tij} = \boldsymbol{d}_{tij}^T \boldsymbol{\beta} + v_j + \boldsymbol{d}_{tij}^T \boldsymbol{u}_{ij}. \tag{5.2}$$

This is a two-level multivariate model for the three level hierarchical structure of PME data set. Heads of household are the level one units that are nested within the PSUs, the level two units. The occasion level is represented in the model to define the multivariate structure (Goldstein, 2003). The multivariate model in equation 5.2 is similar to the one in equation 2.24.

For the model in equation 5.2, the vector with the occasion dummies $\boldsymbol{d}_{tij}$ now contains a total of $T = a + a$ dummies. For the PME data, this means that $\boldsymbol{d}_{tij}$ contains a total of eight dummies, one for each occasion, with both fixed effects and random effects, at the heads of household level, in the model. Hence, it is assumed that:

$$v_j \sim N(0, \sigma_v^2) \text{ and } \boldsymbol{u}_{ij} \sim MN(\boldsymbol{0}, \Sigma_u),$$

where

$$\Sigma_u = \begin{pmatrix} \sigma_{u0}^2 & & & & & & & \\ \sigma_{u1,0} & \sigma_{u1}^2 & & & & & & \\ \sigma_{u2,0} & \sigma_{u2,1} & \sigma_{u2}^2 & & & & & \\ \sigma_{u3,0} & \sigma_{u3,1} & \sigma_{u3,2} & \sigma_{u3}^2 & & & & \\ \sigma_{u12,0} & \sigma_{u12,1} & \sigma_{u12,2} & \sigma_{u12,3} & \sigma_{u12}^2 & & & \\ \sigma_{u13,0} & \sigma_{u13,1} & \sigma_{u13,2} & \sigma_{u13,3} & \sigma_{u13,12} & \sigma_{u13}^2 & & \\ \sigma_{u14,0} & \sigma_{u14,1} & \sigma_{u14,2} & \sigma_{u14,3} & \sigma_{u14,12} & \sigma_{u14,13} & \sigma_{u14}^2 & \\ \sigma_{u15,0} & \sigma_{u15,1} & \sigma_{u15,2} & \sigma_{u15,3} & \sigma_{u15,12} & \sigma_{u15,13} & \sigma_{u15,14} & \sigma_{u15}^2 \end{pmatrix}.$$

Therefore,

$$\Sigma_r = \Sigma_u + J\sigma_v^2,$$

where J is a $T \times T$ matrix with 1 in every entry. The multivariate vector of responses $\boldsymbol{y}_{ij}$ can be written as

$$\boldsymbol{y}_{ij} \sim MN(D_{ij}\boldsymbol{\beta}, \Sigma_r) .$$

The multivariate model 5.2 treats the time variable as discrete. Therefore, it does not matter how the occasion dummies are labelled. In the above formulation it was assumed that they were labelled as $(0, 1, 2, 3, 12, 13, 14, 15)$. However, they could have been alternatively labelled from 0 to 7. This model does not incorporate the gap in such a straightforward way as the random slope model. Instead, this can be achieved by constraining the parameters for the error covariance matrix to depend on the temporal distance between the different occasions. Therefore, a lag-dependent structure of the residual correlation can be imposed. For this, the following matrix serves as a basis:

$$\begin{pmatrix} \mathbf{0} & & & & & & & \\ 1 & \mathbf{0} & & & & & & \\ 2 & 1 & \mathbf{0} & & & & & \\ 3 & 2 & 1 & \mathbf{0} & & & & \\ \mathbf{12} & 11 & \mathbf{10} & 9 & \mathbf{0} & & & \\ 13 & \mathbf{12} & 11 & \mathbf{10} & 1 & \mathbf{0} & & \\ \mathbf{14} & 13 & \mathbf{12} & 11 & 2 & 1 & \mathbf{0} & \\ 15 & \mathbf{14} & 13 & \mathbf{12} & 3 & 2 & 1 & \mathbf{0} \end{pmatrix}. \tag{5.3}$$

The entries of this matrix represent the time distance between the pairs of occasions, or the lag, for the non-consecutive rotating design of 4-8-4, such as the one adopted on the PME survey. Due to the symmetric rotation pattern this lag matrix has the form:

$$\begin{bmatrix} A_{a \times a} & B^T_{a \times a} \\ B_{a \times a} & A_{a \times a} \end{bmatrix}.$$

The sub-matrix A has lags varying from 0 to 3 and the top right entry of sub-matrix B incorporates the gap of eight months between the fourth and the fifth interviews and is 9. Sub-matrix B is a banded matrix with $T - 1$ bands, where T is the total number of occasions. The bottom left entry of sub-matrix B has lag $d - 1$, the total time span minus one. The next subsection presents a lag matrix for a general, not necessarily symmetric, rotation patten *a-b-c(d)*.

The resulting covariance matrix accounting for the gap and following the pattern of the lag-matrix 5.3, can be written as

$$\Sigma_r = \Sigma_u + J\sigma_v^2 \tag{5.4}$$

$$= \begin{pmatrix} \sigma_0^2 &&&&&&& \\ \gamma_1 & \sigma_0^2 &&&&&& \\ \gamma_2 & \gamma_1 & \sigma_0^2 &&&&& \\ \gamma_3 & \gamma_2 & \gamma_1 & \sigma_0^2 &&&& \\ \gamma_{12} & \gamma_{11} & \gamma_{10} & \gamma_9 & \sigma_0^2 &&& \\ \gamma_{13} & \gamma_{12} & \gamma_{11} & \gamma_{10} & \gamma_1 & \sigma_0^2 && \\ \gamma_{14} & \gamma_{13} & \gamma_{12} & \gamma_{11} & \gamma_2 & \gamma_1 & \sigma_0^2 & \\ \gamma_{15} & \gamma_{14} & \gamma_{13} & \gamma_{12} & \gamma_3 & \gamma_2 & \gamma_1 & \sigma_0^2 \end{pmatrix} + J\sigma_v^2. \tag{5.5}$$

Because σ_v^2 is added to every entry of Σ_u, the resulting structure of Σ_r is the same as the one imposed on Σ_u.

5.2.3 Lag Matrix for a General Rotating Design

This subsection presents a generalization of the lag-matrix in 5.3 for a general, not necessarily symmetric, rotating panel design *a-b-c(d)*. The general lag matrix has the form:

$$\begin{bmatrix} A_{a\times a} & B^T_{a\times c} \\ B_{c\times a} & C_{c\times c} \end{bmatrix}. \tag{5.6}$$

Matrix A is defined as

$$A = \begin{pmatrix} a-a &&&& \\ \vdots & a-a &&& \\ a-3 & \vdots & a-a && \\ a-2 & a-3 & \vdots & \ddots & \\ a-1 & a-2 & a-3 & \dots & a-a \end{pmatrix}$$

and following the same pattern, matrix C is defined as

$$\begin{pmatrix} c-c & & & & \\ \vdots & c-c & & & \\ c-3 & \vdots & c-c & & \\ c-2 & c-3 & \vdots & \ddots & \\ c-1 & c-2 & c-3 & \dots & c-c \end{pmatrix}.$$

Matrix B is slightly more complicated than A and C as it depends on a, b and c. Matrix B is a $c \times a$ matrix formed by $T-1$ diagonals of banded lags. Here, T is the number of occasions $a+c$. The top right entry of matrix B has lag equal to $b+1$ and the top left $b+a$, the bottom left entry has lag equal to $b+a+c-1 = d-1$ and the bottom right is $b+c$. Matrix B is then defined as

$$\begin{pmatrix} b+a & b+a-1 & \dots & b+2 & b+1 \\ b+a+1 & b+a & \dots & b+3 & b+2 \\ \dots & \dots & \dots & \dots & \dots \\ a+b+c-2 & \dots & \dots & b+c & \vdots \\ a+b+c-1 & a+b+c-2 & \dots & b+c+1 & b+c \end{pmatrix}.$$

5.3 Variables in the Longitudinal Analysis

Table 5.1 presents a similar table to the one presented in Chapter 4 with the covariates to be included in the models that follow. The variables in this table are organized according to which level of the data hierarchy they belong to. Note that some of the variables are defined here differently to the previous chapter. The first difference is that some of the variables previously defined at the heads of household level are here defined at the occasion level. Note that heads of household were restricted to be employed at all times but not restricted to the same job at all times. Therefore, job characteristics change over time. Some other variables were forced to be defined at the heads of household level. This was achieved by repeating their values observed at the first occasion over the other three. These variables were: age, education and the indicator for white. It is worth noticing that the variables for age and duration of employment had their effects centred on, respectively, the age of 40 and the occasion mean durations. The variable for occasions, called *wave*, and month of the first interview were reduced by one (-1) now varying respectively from 0 to 3 and 0 to 20. Squared terms were also considered; see Table 5.1. Note that the variable representing time incorporates the gap between the fourth and the fifth interview.

Table 5.1: Variables Included in the Longitudinal Analyses

Variables	Definition
	Occasion Level Covariates
Wave	Time variable with values (0, 1, 2, 3, 12, 13, 14 and 15).
Type of Worker	Employer, Informal, Formal, Military Service and Self-Employed.
Type of Activity	Manufacturing, Building, Commerce, Financial, Social Services, Domestic Services, Other Services and Other Activities.
Duration of Employment	In months with squared term - a proxy for experience in the firm.
Working Hours	Working hours on log scale.
Proxy Respondent	An indicator for the use of proxy respondent.
Number of Household Members	Considered as a continuous variable.
	Individual Level Covariates
Month of First Interview	From 0 to 20.
Gender	Females as the baseline.
Age	Age at the first interview and squared term.
Race	An indicator for whites against all other categories collapsed.
Education in Years of Schooling	From 0 to 17 years of schooling at the first interview.
	Cluster Level Covariates
Metropolitan Regions	Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo and Porto Alegre.

5.4 Alternative Covariance Structures

In Chapter 2, different residual covariance structures were presented. However, some of these structures are not completely adequate for the analysis presented in this chapter. This section examines these structures and proposes alternative ones that seem more adequate to the objective of this chapter.

The first structure to be examined is the unconstrained denoted as Σ_r^{unc} as presented in Chapter 2. This structure, as the name suggests, sets no constraints on the parameters of the error covariance matrix. These parameters are freely estimated and determined by the data (Singer and Willett, 2003). Therefore, the potential effect of the gap between the fourth and fifth interviews can also be determined by the data. Furthermore, a multivariate model for the longitudinal data set that imposes an unstructured error covariance would have 36 additional parameters to be estimated. As mentioned in Chapter 2, this model would have the smallest deviance when compared to others imposing different structures for Σ_r. This is not the most parsimonious structure and it could have a detrimental effect on the efficiency of the inferences about the mean structure of such a model (Pourahmadi, 2007).

The AR(1), denoted as Σ_r^{ar} in Chapter 2, is a lag-dependent structure which assumes that the variances are constant overtime and that pairs of equally spaced residuals have a constant covariance determined as a function of ρ as

$$(\Sigma_u)_{rs} = \gamma_{rs} = \sigma^2 \rho^{|s-r|} \qquad \forall r \neq s \ .$$

This means that the AR(1) structure imposes that the residual correlation at the heads of household level has a fixed decay determined by a fraction of ρ. However the same decay is not observed for Σ_r due to the addition of σ_v^2 to every entry of Σ_u. Although this structure is lag-dependent, it assumes that the occasions are equally spaced over time. This is not the case for the PME rotating design. Therefore, the AR(1) structure is not an appropriate covariance structure for this longitudinal data set when accounting for the PME rotating design. It would be adequate, however, if the gap between the mid-point occasions was to be ignored.

The Toeplitz error covariance, Σ_r^{toep} as presented in Chapter 2 is also a lag-dependent structure assuming equal variance over time and equal covariance between equally spaced residuals. Although the Toplitz has a banded-diagonal structure, it is not as tightly constrained as the AR(1). The covariances between two occasions for the Toeplitz, do not depend on a fraction of ρ, they only depend on the temporal distance between them (Rochon and Helms, 1989). Because of this stationary structure, the Toeplitz error covariance is only appropriate for equally spaced data. Like the three previous structures, it could only be imposed on the multivariate models for this longitudinal data set if the gap was to be ignored.

None of the above structures fully accounts for the PME rotation pattern. The AR(1) and the Toeplitz are both lag-dependent but they assume that the measurements were taken at equal time intervals. More appropriate structures would be those that do account for the gap between the fourth and fifth interviews and allow for irregularly spaced data, such as those in the longitudinal data set. Two alternative structures are proposed here: the temporal power and the general linear lag-dependent.

The structure hereafter referred to as the **temporal power** structure is the Spatial Power (SAS Institute Inc,Version 8, 1999) structure presented in Chapter 2. As mentioned in that chapter, this structure is a reparameterization of the exponential correlation which involves setting

$$\rho = \exp\left(\frac{-1}{\phi}\right)$$

and expressing the covariance as

$$\gamma_{rs} = \sigma^2 \rho^{|t_s - t_r|}.$$

Therefore, the temporal power structure takes into account the temporal distance between the occasions by powering ρ by $|t_s - t_r|$. This structure is generally applied to spatial data. However, when dealing with temporal data instead, the supposed spatial process is thought of as a one dimensional process (Khattree and Naik, 1999). Hence, only one coordinate variable, containing the time variable that incorporates the rotating design, should be considered. The temporal power covariance structure can be imposed on any data that are generated from a sample including rotating panel schemes, as long as the coordinate variable incorporates the rotation pattern. Based on the lag matrix 5.3 and given the time variable defined as $(0, 1, 2, 3, 12, 13, 14, 15)$, the temporal power covariance structure for this longitudinal data set can be written as:

$$\Sigma_r^{temp} = \begin{pmatrix} \sigma^2 & & & & & & & \\ \sigma^2\rho^1 & \sigma^2 & & & & & & \\ \sigma^2\rho^2 & \sigma^2\rho^1 & \sigma^2 & & & & & \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho^1 & \sigma^2 & & & & \\ \sigma^2\rho^{12} & \sigma^2\rho^{11} & \sigma^2\rho^{10} & \sigma^2\rho^9 & \sigma^2 & & & \\ \sigma^2\rho^{13} & \sigma^2\rho^{12} & \sigma^2\rho^{11} & \sigma^2\rho^{10} & \sigma^2\rho^1 & \sigma^2 & & \\ \sigma^2\rho^{14} & \sigma^2\rho^{13} & \sigma^2\rho^{12} & \sigma^2\rho^{11} & \sigma^2\rho^2 & \sigma^2\rho^1 & \sigma^2 & \\ \sigma^2\rho^{15} & \sigma^2\rho^{14} & \sigma^2\rho^{13} & \sigma^2\rho^{12} & \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho^1 & \sigma^2 \end{pmatrix} + \mathrm{J}\sigma_v^2.$$

The second alternative structure is hereafter referred to as the **general linear lag-dependent** covariance structure. This is the general linear covariance structure as presented in Chapter 2 which expresses Σ_u as a linear combination such as

$$\Sigma_u = \theta_0 A_0 + \theta_1 A_1 + ... + \theta_q A_q, \tag{5.7}$$

where the matrices A_q are known symmetric matrices and θ_q are unknown parameters to be estimated and these parameters are unrelated to each other (Khattree and Naik, 1999). For the current problem, the A_q matrices are set in a way to represent the temporal distance between the occasions. In other words they are set to be lag-dependent. Motivated by the Toeplitz structure and having the lag matrix 5.3 as the basis for the 4-8-4 rotation pattern, a general linear lag-dependent covariance structure of the following form is considered:

$$\begin{pmatrix}
\gamma_0 & & & & & & & \\
\gamma_1 & \gamma_0 & & & & & & \\
\gamma_2 & \gamma_1 & \gamma_0 & & & & & \\
\gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 & & & & \\
\gamma_{12} & \gamma_{11} & \gamma_{10} & \gamma_9 & \gamma_0 & & & \\
\gamma_{13} & \gamma_{12} & \gamma_{11} & \gamma_{10} & \gamma_1 & \gamma_0 & & \\
\gamma_{14} & \gamma_{13} & \gamma_{12} & \gamma_{11} & \gamma_2 & \gamma_1 & \gamma_0 & \\
\gamma_{15} & \gamma_{14} & \gamma_{13} & \gamma_{12} & \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0
\end{pmatrix}. \tag{5.8}$$

Like the temporal power structure, the general linear lag-dependent structure can be imposed on data from any pattern of rotating design such as those in the general lag matrix 5.6. However, it requires that the matrices A_q can be used to represent the specific lag matrix and that Σ_u is still a positive definite matrix.

For the longitudinal data set, the matrix 5.8 can be re-written and decomposed as follows:

$$\Sigma_u = \gamma_0 A_0 + \gamma_1 A_1 + \gamma_2 A_2 + \gamma_3 A_3 + \gamma_9 A_9 + \gamma_{10} A_{10}$$
$$+\gamma_{11} A_{11} + \gamma_{12} A_{12} + \gamma_{13} A_{13} + \gamma_{14} A_{14} + \gamma_{15} A_{15}.$$

$$\Sigma_u = \gamma_0 \begin{pmatrix}
1 & & & & & & & \\
0 & 1 & & & & & & \\
0 & 0 & 1 & & & & & \\
0 & 0 & 0 & 1 & & & & \\
0 & 0 & 0 & 0 & 1 & & & \\
0 & 0 & 0 & 0 & 0 & 1 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix} + \gamma_1 \begin{pmatrix}
0 & & & & & & & \\
1 & 0 & & & & & & \\
0 & 1 & 0 & & & & & \\
0 & 0 & 1 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 1 & 0 & & \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}$$
$$+\gamma_2 \begin{pmatrix}
0 & & & & & & & \\
0 & 0 & & & & & & \\
1 & 0 & 0 & & & & & \\
0 & 1 & 0 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix} + \gamma_3 \begin{pmatrix}
0 & & & & & & & \\
0 & 0 & & & & & & \\
0 & 0 & 0 & & & & & \\
1 & 0 & 0 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}$$

$$
+\gamma_9 \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
+\gamma_{10} \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

$$
+\gamma_{11} \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
+\gamma_{12} \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

$$
+\gamma_{13} \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
+\gamma_{14} \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

$$
+\gamma_{15} \begin{pmatrix} 0 \\ 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.
$$

For this specific rotation pattern the matrix Σ_u can be decomposed into 11 matrices and there will be 11 parameters to be estimated. It is worth mentioning that the way the matrices A_q are represented above does not guarantee positive definitiveness for Σ_u. There are ways to overcome this problem, either by modifying the matrices A_q, for example by considering the entries of every main diagonal as equal to 1, or by specifying start values for the parameters.

5.4.1 Computational Aspects

Multilevel models can be fitted through a range of widely available statistical computer packages. In Chapter 4 it was said that multilevel models can be fitted for example in `Stata` or `MLwiN` software. The same can be said about the longitudinal multilevel models. However, the fit of the multivariate multilevel model is not as straightforward in either of these two packages.

Three level growth curve models can be fitted using the `Stata` command `xtmixed`, which accommodates the various levels of the analysis. This command offers various options which can be changed in order to fit models using, for example, maximum likelihood estimation or restricted maximum likelihood estimation. Random coefficients can be declared at the various levels of the analysis and they can be set to be independent or set to covary. However, the user cannot change the default of this command to impose correlation among occasion level residuals. Hence, multivariate multilevel models cannot be fitted in `Stata` via `xtmixed` alone. Grilli and Rampichini (2006) showed different ways to set up a multivariate multilevel model using the package `GLLAMM` (Rabe-Hesketh et al., 2004) in `Stata`. This included setting up a latent factor model. One disadvantage is the computational time necessary for `GLLAMM` to fit these types of models.

Multilevel models, including the multivariate multilevel model, can be fitted using the software `MLwiN`. The multivariate model can be set up following the steps presented in the `MLwiN` user's guide (Rasbash et al., 2001). To follow these procedures the longitudinal data set needs to be in a wide format, where the rows represent the heads of household and their repeated measurements and other variables are the different columns. This allows for the inclusion of common or separated covariates and random intercept and slopes for the measurement occasion variable in any of the levels of the data set. One disadvantage, however, is that it is not clear in Rasbash et al. (2001) how to proceed to include occasion level covariates.

There is an alternative way to set up the multivariate multilevel model in `MLwiN` to produce the exact same results as the one described above (Singer and Willett, 2003, Chapter7). This can be achieved by setting up the multivariate model as a discrete growth curve model, where the occasion dummies are treated both as fixed and random effects in the model. For identification purposes, this model should not include a constant and does not contain the occasion level variance. For this set up, the longitudinal data set can be in a long format, where each occasion is a row of the data set, and the inclusion of occasion level variables is straightforward. In addition, it is also possible to include variables having either separate or common coefficients, by including the interaction terms of these variables with the occasion dummies. Moreover, this set up allows linear constraints to be imposed on the covariance matrix.

The statistical package `SAS`, on the other hand, can fit both the growth curve model and the multivariate multilevel model in a straightforward way (Singer, 1998). The former is achieved by adding the option `RANDOM` to the `PROC MIXED` command, and the latter by adding the option `REPEATED`. The `PROC MIXED` can also fit the two-level multivariate model in equation 2.24, where there is a random intercept at the PSU level. Another advantage of fitting multivariate multilevel models in `SAS PROC MIXED` command is that different covariance structures, including the AR(1) and the two alternative ones, can be imposed. Khattree and Naik (1999) and SAS Institute Inc,Version 8 (1999) presented a list of other structures that can be used.

The temporal power covariance structure can be imposed on a multivariate model when fitting it using `SAS PROC MIXED` by changing the `TYPE` option for the `REPEATED` statement to `TYPE = SP(POW)(`*`c-list`*`)`. The string *c-list* contains the numerical variable with the spatial coordinates (SAS Institute Inc,Version 8, 1999). In the PME case, this is the time variable indicating the lags. To fit a multivariate multilevel model imposing the general linear lag-dependent covariance structure the `TYPE` option needs to be changed to `TYPE = LIN(`*`q`*`)` where q is the number of parameters θ_q to be estimated (SAS Institute Inc,Version 8, 1999). In addition the user needs to provide an external file with all the A_q matrices. See Appendix A, which presents the `SAS` codes for fitting a multivariate model, such as model 5.2, imposing both the temporal power structure and the general linear lag-dependent structure.

5.5 Results

This section begins by illustrating the fit of the growth curve model and the multivariate model both accounting for the PME rotating design of the longitudinal data set. The different residual covariance structures presented in Section 5.4 are imposed on the multivariate multilevel models in this section. Furthermore, different ways of expressing the time variable are considered for comparison. This means that time is either considered to vary from 0 to 7 or is expressed as $(0, 1, 2, 3, 12, 13, 14, 15)$. The different models are compared and contrasted and their goodness-of-fit evaluated. It is worth noticing that from now on, when the time variable is expressed to account for the gap, as $(0, 1, 2, 3, 12, 13, 14, 15)$, this is referred as (0-15), otherwise (0-7).

With only the time variable as a covariate, the following models were fitted to the longitudinal data set:

Random Slope Model (0-7): This is the growth curve model described in equation 5.1. By considering time as varying from 0 to 7, this model does not account for the gap between the fourth and fifth occasions as it assumes an equally spaced data set.

Random Slope Model (0-15): This is the same random slope model as before but accounting for the irregularly spaced data of the longitudinal data set by considering time as a continuous variable varying from 0 to 15.

Unstructured (0-7): This is the multivariate multilevel model described in equation 5.2 imposing no constraints on the error covariance structure. Time was labelled as varying from 0 to 7 and is presented purely for comparison.

Unstructured (0-15): This is the same unstructured model as the previous model, but labelling the time variable varying from 0 to 15. The same results for both unstructured models are found, as the multivariate model treats time as a discrete variable and the error covariance structure for these two models are not lag-dependent.

First Order Autoregressive: This is the multivariate multilevel model 5.2 constraining the parameters of the error covariance matrix to represent an AR(1) structure. This model, as already mentioned, does not account for the rotation pattern as it assumes that the occasions are equally spaced in time.

Toeplitz: Like the AR(1) model, the multivariate multilevel model 5.2 imposing a Toeplitz structure does not fully account for the rotation pattern. Both of these models have lag-dependent error covariance structures but ignore the gap between the fourth and fifth interviews.

Temporal Power (0-7): This is the multivariate multilevel model 5.2 imposing a temporal power error-covariance structure. For comparison reasons, the time variable was here considered to vary from 0 to 7. This model provides the exact same results as the AR(1) model, which is omitted from the following tables.

Temporal Power (0-15): This is the multivariate multilevel model 5.2 imposing the temporal power error covariance structure that fully accounts for the PME rotating design. The time variable here is considered to vary from 0 to 15, and this variable is declared as the coordinate variable for the powering of ρ.

General Linear Lag-dependent (0-15): This is the multivariate multilevel model 5.2 imposing the general linear lag-dependent structure. Time is also considered to vary from 0 to 15 and the known matrices A_q were set to account for the temporal distance between the occasions.

Table 5.2: Model Selection: Goodness-of-fit Statistics

	Time varying as			
	from 0 to 7		from 0 to 15	
Model	AIC	BIC (SAS)	AIC	BIC (SAS)
Longitudinal Multilevel Model				
Random Slope	30,316	30,343	27,534	27,561
Multivariate Multilevel Model				
Unstructured	26,691	26,893	26,691	26,893
Toeplitz	29,392	29,441	-	-
Temporal Power	35,630	35,647	37,701	37,717
General Linear Lag-Dependent	-	-	26,950	27,016

Table 5.2 presents the goodness-of-fit statistics of the different models. All of these models were fitted using REML estimation as the main aim was to investigate the random parts of these models. Additionally, because these models are not necessarily nested, the likelihood ratio test cannot be applied here. For these reasons Table 5.2 presents the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) which are goodness-of-fit statistics that can be used instead. Under the column for time varying from 0 to 7 are the statistics for those models that do not account fully for the PME rotation pattern. For the random slope model, the model that accounts for the rotation pattern (0-15) has a better fit as AIC and BIC have smaller values than for the model that ignores the gap.

Comparing the multivariate models that ignore the gap, the temporal power structure is the least preferable, as it has the highest values for both AIC and BIC statistics. The Toeplitz structure is less affected than the temporal power structure, as it is less restrictive in the decay of the correlation over time. Table 5.3 presents the estimated error covariance matrices and error autocorrelation matrices for each of these models. Imposing the temporal power structure, assuming that the measurements in the longitudinal data set are equally spaced over time, underestimates the residual correlation. The unconstrained covariance structure, in turn, is unaffected by the incorrect specification of time. Hence, this model presents the best values for the goodness-of-fit statistics.

Observing the results for the models under the column for time varying from 0 to 15, once again, the least preferable structure is the temporal power. The model imposing an unstructured error covariance matrix has again the smallest values for AIC and BIC. However, this is the least parsimonious model. As already mentioned, the estimation of additional nuisance parameters may cause loss of efficiency in the inference for the fixed part of the model, and this model might not be a good choice. Comparing the general linear-dependent model with the unstructured model, the former does not do too badly, compared to the latter.

The general linear-dependent model fully accounts for the rotation pattern and has less estimated parameters than the unstructured model, therefore being a more parsimonious choice.

Observing Table 5.3 and the results for the random slope models, notice that they do provide different results as the continuous time variable is expressed differently in these models. The estimated autocorrelation matrix for the random slope model (0-7) presents an approximate banded structure that decays with increasing lag. This is a result of assuming time as varying from 0 to 7. In addition, the estimated covariance matrix shows some indication of heteroscedastic variance over time. The results for the unstructured model present similar trends as those for the random slope models. It is worth mentioning that because the models for unstructured (0-7) and unstructured (0-15) provided the exact same results, Table 5.3 presents these results only once. The temporal power (0-7) assumes that the variance is constant over time and the estimated autocorrelation matrix shows a fast decay towards zero as the temporal distance increases between the occasions. The results for the Toeplitz model shows a slower autocorrelation decay than the temporal power (0-7) structure, although it still assumes homoscedasticity and equally spaced data.

Table 5.3: Covariance Components and Autocorrelation Matrices

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
Random Slope (0-7)	$\begin{pmatrix} 0.8241 \\ 0.7595 & 0.7997 \\ 0.7457 & 0.7383 & 0.7817 \\ 0.7318 & 0.7277 & 0.7235 & 0.7701 \\ 0.7180 & 0.7171 & 0.7161 & 0.7152 & 0.7650 \\ 0.7042 & 0.7065 & 0.7088 & 0.7110 & 0.7133 & 0.7664 \\ 0.6904 & 0.6959 & 0.7014 & 0.7069 & 0.7124 & 0.7179 & 0.7742 \\ 0.6765 & 0.6853 & 0.6940 & 0.7027 & 0.7115 & 0.7202 & 0.7289 & 0.7884 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.9356 & 1.0000 \\ 0.9291 & 0.9338 & 1.0000 \\ 0.9186 & 0.9273 & 0.9325 & 1.0000 \\ 0.9043 & 0.9168 & 0.9261 & 0.9318 & 1.0000 \\ 0.8861 & 0.9024 & 0.9157 & 0.9255 & 0.9316 & 1.0000 \\ 0.8643 & 0.8844 & 0.9016 & 0.9155 & 0.9257 & 0.9320 & 1.0000 \\ 0.8393 & 0.8630 & 0.8840 & 0.9018 & 0.9161 & 0.9265 & 0.9330 & 1.0000 \end{pmatrix}$
Random Slope (0-15)	$\begin{pmatrix} 0.8084 \\ 0.7571 & 0.7987 \\ 0.7519 & 0.7479 & 0.7902 \\ 0.7468 & 0.7434 & 0.7399 & 0.7827 \\ 0.7007 & 0.7023 & 0.7039 & 0.7055 & 0.7661 \\ 0.6956 & 0.6977 & 0.6999 & 0.7020 & 0.7214 & 0.7698 \\ 0.6905 & 0.6932 & 0.6959 & 0.6986 & 0.7230 & 0.7258 & 0.7747 \\ 0.6853 & 0.6886 & 0.6919 & 0.6952 & 0.7246 & 0.7279 & 0.7312 & 0.7807 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.9421 & 1.0000 \\ 0.9408 & 0.9415 & 1.0000 \\ 0.9388 & 0.9402 & 0.9409 & 1.0000 \\ 0.8904 & 0.8978 & 0.9047 & 0.9111 & 1.0000 \\ 0.8817 & 0.8898 & 0.8974 & 0.9044 & 0.9394 & 1.0000 \\ 0.8725 & 0.8812 & 0.8894 & 0.8971 & 0.9386 & 0.9398 & 1.0000 \\ 0.8627 & 0.8720 & 0.8809 & 0.8893 & 0.9370 & 0.9389 & 0.9402 & 1.0000 \end{pmatrix}$
Unstructured	$\begin{pmatrix} 0.8085 \\ 0.7563 & 0.8047 \\ 0.7471 & 0.7537 & 0.7974 \\ 0.7378 & 0.7447 & 0.7470 & 0.7821 \\ 0.7039 & 0.7068 & 0.7056 & 0.7019 & 0.7801 \\ 0.6946 & 0.6982 & 0.6949 & 0.6953 & 0.7362 & 0.7689 \\ 0.6910 & 0.6943 & 0.6915 & 0.6890 & 0.7268 & 0.7281 & 0.7635 \\ 0.6913 & 0.6956 & 0.6929 & 0.6925 & 0.7264 & 0.7270 & 0.7285 & 0.7692 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.9376 & 1.0000 \\ 0.9305 & 0.9409 & 1.0000 \\ 0.9278 & 0.9387 & 0.9459 & 1.0000 \\ 0.8863 & 0.8921 & 0.8946 & 0.8986 & 1.0000 \\ 0.8810 & 0.8876 & 0.8875 & 0.8966 & 0.9506 & 1.0000 \\ 0.8795 & 0.8858 & 0.8862 & 0.8916 & 0.9417 & 0.9503 & 1.0000 \\ 0.8766 & 0.8841 & 0.8847 & 0.8928 & 0.9377 & 0.9453 & 0.9506 & 1.0000 \end{pmatrix}$

Table 5.3 – continued from previous page

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
Toeplitz	$\begin{pmatrix} 0.7906 & & & & & & & \\ 0.7432 & 0.7906 & & & & & & \\ 0.7322 & 0.7432 & 0.7906 & & & & & \\ 0.7205 & 0.7322 & 0.7432 & 0.7906 & & & & \\ 0.7106 & 0.7205 & 0.7322 & 0.7432 & 0.7906 & & & \\ 0.7016 & 0.7106 & 0.7205 & 0.7322 & 0.7432 & 0.7906 & & \\ 0.6939 & 0.7016 & 0.7106 & 0.7205 & 0.7322 & 0.7432 & 0.7906 & \\ 0.6852 & 0.6939 & 0.7016 & 0.7106 & 0.7205 & 0.7322 & 0.7432 & 0.7906 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.9400 & 1.0000 & & & & & & \\ 0.9261 & 0.9400 & 1.0000 & & & & & \\ 0.9113 & 0.9261 & 0.9400 & 1.0000 & & & & \\ 0.8988 & 0.9113 & 0.9261 & 0.9400 & 1.0000 & & & \\ 0.8874 & 0.8988 & 0.9113 & 0.9261 & 0.9400 & 1.0000 & & \\ 0.8777 & 0.8874 & 0.8988 & 0.9113 & 0.9261 & 0.9400 & 1.0000 & \\ 0.8667 & 0.8777 & 0.8874 & 0.8988 & 0.9113 & 0.9261 & 0.9400 & 1.0000 \end{pmatrix}$
Temporal Power (0-7)	$\begin{pmatrix} 0.7984 & & & & & & & \\ 0.7507 & 0.7984 & & & & & & \\ 0.7082 & 0.7507 & 0.7984 & & & & & \\ 0.6703 & 0.7082 & 0.7507 & 0.7984 & & & & \\ 0.6364 & 0.6703 & 0.7082 & 0.7507 & 0.7984 & & & \\ 0.6062 & 0.6364 & 0.6703 & 0.7082 & 0.7507 & 0.7984 & & \\ 0.5793 & 0.6062 & 0.6364 & 0.6703 & 0.7082 & 0.7507 & 0.7984 & \\ 0.5553 & 0.5793 & 0.6062 & 0.6364 & 0.6703 & 0.7082 & 0.7507 & 0.7984 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.9403 & 1.0000 & & & & & & \\ 0.8870 & 0.9403 & 1.0000 & & & & & \\ 0.8396 & 0.8870 & 0.9403 & 1.0000 & & & & \\ 0.7971 & 0.8396 & 0.8870 & 0.9403 & 1.0000 & & & \\ 0.7593 & 0.7971 & 0.8396 & 0.8870 & 0.9403 & 1.0000 & & \\ 0.7256 & 0.7593 & 0.7971 & 0.8396 & 0.8870 & 0.9403 & 1.0000 & \\ 0.6955 & 0.7256 & 0.7593 & 0.7971 & 0.8396 & 0.8870 & 0.9403 & 1.0000 \end{pmatrix}$

Table 5.3 – continued from previous page

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
Temporal Power (0-15)	$\begin{pmatrix} 0.8182 \\ 0.7791 & 0.8182 \\ 0.7434 & 0.7791 & 0.8182 \\ 0.7108 & 0.7434 & 0.7791 & 0.8182 \\ 0.5205 & 0.5347 & 0.5503 & 0.5674 & 0.8182 \\ 0.5075 & 0.5205 & 0.5347 & 0.5503 & 0.7791 & 0.8182 \\ 0.4957 & 0.5075 & 0.5205 & 0.5347 & 0.7434 & 0.7791 & 0.8182 \\ 0.4849 & 0.4957 & 0.5075 & 0.5205 & 0.7108 & 0.7434 & 0.7791 & 0.8182 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.9522 & 1.0000 \\ 0.9086 & 0.9522 & 1.0000 \\ 0.8687 & 0.9086 & 0.9522 & 1.0000 \\ 0.6362 & 0.6535 & 0.6726 & 0.6935 & 1.0000 \\ 0.6203 & 0.6362 & 0.6535 & 0.6726 & 0.9522 & 1.0000 \\ 0.6058 & 0.6203 & 0.6362 & 0.6535 & 0.9086 & 0.9522 & 1.0000 \\ 0.5926 & 0.6058 & 0.6203 & 0.6362 & 0.8687 & 0.9086 & 0.9522 & 1.0000 \end{pmatrix}$
General Linear Lag-dependent	$\begin{pmatrix} 0.7869 \\ 0.7446 & 0.7869 \\ 0.7390 & 0.7446 & 0.7869 \\ 0.7338 & 0.7390 & 0.7446 & 0.7869 \\ 0.6992 & 0.7007 & 0.7049 & 0.7051 & 0.7869 \\ 0.6965 & 0.6992 & 0.7007 & 0.7049 & 0.7446 & 0.7869 \\ 0.6942 & 0.6965 & 0.6992 & 0.7007 & 0.7390 & 0.7446 & 0.7869 \\ 0.6909 & 0.6942 & 0.6965 & 0.6992 & 0.7338 & 0.7390 & 0.7446 & 0.7869 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.9462 & 1.0000 \\ 0.9391 & 0.9462 & 1.0000 \\ 0.9325 & 0.9391 & 0.9462 & 1.0000 \\ 0.8886 & 0.8905 & 0.8958 & 0.8960 & 1.0000 \\ 0.8851 & 0.8886 & 0.8905 & 0.8958 & 0.9462 & 1.0000 \\ 0.8822 & 0.8851 & 0.8886 & 0.8905 & 0.9391 & 0.9462 & 1.0000 \\ 0.8780 & 0.8822 & 0.8851 & 0.8886 & 0.9325 & 0.9391 & 0.9462 & 1.0000 \end{pmatrix}$

When comparing the results for the models that fully account for the rotating design, the temporal power (0-15) model shows an even faster decay of the residual correlation. This is probably an indication why this is not a preferable structure when comparing the results of Table 5.2. In other words, this model assumes a relative faster decay for the autocorrelation for larger temporal distance than the actual data possibly presents. The general linear lag-dependent model, in turn, does not impose such a fast decay as the temporal power (0-15). The estimated autocorrelation matrix for the general linear lag-dependent model shows a similar decay in the correlation as those for the random slope model (0-15) and for the unstructured model. It also shows that the correlation for lags 9 and 10 seem to be very close, this could be an indication that these two parameters could be further constrained to be equal for future analysis of these data.

Table 5.4 presents the results for the fixed part of these models. For comparison this table also presents robust standard errors of the regression coefficients. This method is discussed in the next two chapters. As expected, the results for the two random slope models differ. These models are dependent on the values of the continuous variable for time. Observing the results for the multivariate models, note that the coefficient estimates for the occasion dummies are very similar regardless of the imposed covariance structure. The exception is for the results for the model imposing the temporal power (0-15) structure. The robust standard errors are also very similar for all the multivariate models. However, the same is not observed for the non-robust standard errors of these estimates. The non-robust standard errors are somewhat larger for both the temporal power structures when compared with the unstructured, the Toeplitz and the general linear lag-dependent structure. When comparing robust with non-robust standard errors, non-robust standard errors for temporal power structures are always larger than the robust ones, which is not the case for the other covariance structures.

Based on the results presented so far, the analysis continued with the inclusion of other covariates. The models were fitted only considering those structures that account for the gap in the PME rotating design. Therefore, only the random slope model (0-15), the unstructured and the general linear lag-dependent models are fitted. The temporal power model was initially considered but then discarded, bearing in mind that in the analysis presented above it was found not to be a good model for the longitudinal data set.

Due to methodological, rather than substantive, motivation the model selection strategy followed the same steps as those for the final cross-sectional analysis performed in Chapter 4, which served as the basis for the models considered in this

Table 5.4: Fixed Parameters Estimates

	RS (0-7)	RS (0-15)	Uns	Temp (0-7)	Toep	Temp (0-15)	GLLD
Intercept	6.626	6.629					
SE	(0.0167)	(0.0166)					
SE_{Rob}	(0.0171)	(0.0170)					
Wave	0.006	0.003					
SE	(0.0008)	(0.0003)					
SE_{Rob}	(0.0009)	(0.0004)					
Dummies							
d_0			6.629	6.628	6.629	6.627	6.629
SE			(0.0168)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0172)	(0.0172)	(0.0172)	(0.0172)	(0.0172)
d_1			6.632	6.631	6.632	6.630	6.632
SE			(0.0168)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0170)	(0.0171)	(0.0171)	(0.0171)	(0.0171)
d_2			6.631	6.630	6.631	6.630	6.631
SE			(0.0167)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0169)	(0.0169)	(0.0169)	(0.0169)	(0.0169)
d_3			6.637	6.636	6.636	6.635	6.636
SE			(0.0167)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0169)	(0.0169)	(0.0169)	(0.0169)	(0.0169)
d_{12}			6.667	6.666	6.667	6.666	6.667
SE			(0.0167)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0165)	(0.0165)	(0.0165)	(0.0165)	(0.0165)
d_{13}			6.662	6.661	6.661	6.660	6.661
SE			(0.0166)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0165)	(0.0165)	(0.0165)	(0.0165)	(0.0165)
d_{14}			6.661	6.660	6.661	6.659	6.660
SE			(0.0166)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0165)	(0.0165)	(0.0165)	(0.0165)	(0.0165)
d_{15}			6.667	6.666	6.667	6.665	6.667
SE			(0.0166)	(0.0171)	(0.0168)	(0.0173)	(0.0167)
SE_{Rob}			(0.0164)	(0.0164)	(0.0164)	(0.0164)	(0.0164)

chapter. The model selection strategy commenced with the inclusion of the main fixed effects. The variables considered as covariates were those described in Table 5.1. Their significance levels were tested sequentially, meaning, first considering the main effects, then the interaction effects, the contextual effects and finally the interaction of the time variable with other variables. Main effects were retained if significant at the 5% level. The next step was the inclusion of interaction effects between the retained covariates and the variables for males and whites. Only these interaction terms were tested, for the substantive reason, which is the belief that there is still the presence of gender and race discrimination in the Brazilian labour market. The significant interaction terms were retained and this step was followed by the inclusion of cluster level covariates. However, to facilitate the model fitting process, which at this point was already time-consuming, only significant contextual effects from the cross-sectional modelling in Chapter 4 were considered here. This model selection procedure still included tests for significant interaction effects between time and other covariates.

It is worth mentioning that the squared and the cubic terms for the time variable were tested in the random slope model but were not significant. Hence, only the linear term was retained. It is also worth mentioning that, apart from being a longitudinal sample, the sample considered in the longitudinal data set is relatively smaller than the one considered in Chapter 4. This might explain the loss of significance of some of the variables in the model. The following variables were excluded due to loss of significance: month at first interview (and its squared term), the interaction term between number of household members and the dummy for males, the interaction term between the squared term for age and dummy for males and the contextual effects of age, race and proxy respondent. If of substantive interest, extension of this analysis would be to test for the inclusion of other interaction terms between time and other variables and possibly other contextual effects. The selection of the fixed part of the model was performed by fitting these models using ML estimation. This allows for the performance of the likelihood ratio test. Once the fixed part of the model was selected, the random part of the model was finally estimated via REML.

Before presenting the results for the fixed part of the models, the random part of these models are examined. Table 5.5 presents the statistics for the goodness-of-fit for the four models fitted here. The temporal power model is the least preferable, reinforcing the previous findings about this structure. The choice is once again between the multivariate models with unstructured and general linear lag-dependent covariance structures. The same arguments made in favour of the general linear lag-dependent model are valid here. This model is more parsimonious when compared to the one imposing no constraints on the error covariance structure and fully accounts for the rotation pattern. Appendix B presents the plots for the residual diagnostics for both models. Both multivariate models produce similar plots.

Table 5.5: Model Selection: Goodness-of-fit Statistics (Model with more Covariates)

Model	AIC	BIC (SAS)
Random Slope (0-15)	21,246	21,273
Unstructured	20,446	20,649
Temporal Power (0-15)	30,178	30,195
General Linear Lag-Dependent	20,644	20,709

Table 5.6 presents the estimated residual covariance and autocorrelation matrices for the four tested models. When comparing these autocorrelation matrices with their respective matrices when no covariates were considered in the model other than time, it can be observed that now the autocorrelations are smaller but still quite strong. However, a faster decay for larger temporal distance between occasions is observed after controlling for covariates. Additionally, the autocorrelation matrix for the temporal power model has a much quicker decay towards zero for larger temporal distances than any other structures.

Figure 5.1: Autocorrelation Function

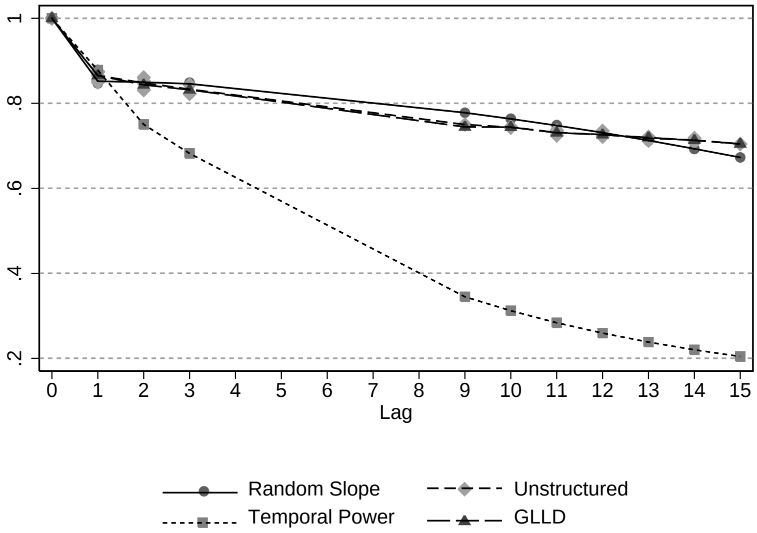


Figure 5.1 presents a plot for the autocorrelation function by lag. The series of this plot were constructed from the autocorrelation matrices for each of the structures. The line for the unstructured correlation represents the average over the points for the given lag. This figure shows how different the autocorrelation function for the temporal power is from the rest. It also shows that the autocorrelation functions for the general linear lag-dependent and the unstructured models are very similar as they lie almost on the top of each other. Additionally, for all the structures the autocorrelation shows a decreasing trend over time.

Table 5.7 presents the results for the fixed part of the model only for the unstructured, for comparison reasons, and the general linear lag-dependent models. A full interpretation of this table will be deferred to later analysis of this data set as for now the motivation was more methodological than substantive. However, by comparing the results of the two models in Table 5.7 a very small difference in magnitude of the fixed effects is observed while the standard errors of these estimates are nearly identical in both models.

Table 5.6: Covariance Components and Autocorrelation Matrices (Model with more Covariates)

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
Random Slope (0-15)	$\begin{pmatrix} 0.3237 & & & & & & & \\ 0.2742 & 0.3155 & & & & & & \\ 0.2698 & 0.2665 & 0.3083 & & & & & \\ 0.2655 & 0.2627 & 0.2598 & 0.3022 & & & & \\ 0.2261 & 0.2280 & 0.2299 & 0.2318 & 0.2938 & & & \\ 0.2218 & 0.2242 & 0.2266 & 0.2290 & 0.2506 & 0.2981 & & \\ 0.2174 & 0.2203 & 0.2232 & 0.2262 & 0.2525 & 0.2554 & 0.3034 & \\ 0.2130 & 0.2165 & 0.2199 & 0.2234 & 0.2544 & 0.2578 & 0.2613 & 0.3098 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8581 & 1.0000 & & & & & & \\ 0.8542 & 0.8546 & 1.0000 & & & & & \\ 0.8488 & 0.8507 & 0.8514 & 1.0000 & & & & \\ 0.7332 & 0.7489 & 0.7638 & 0.7779 & 1.0000 & & & \\ 0.7139 & 0.7310 & 0.7473 & 0.7629 & 0.8467 & 1.0000 & & \\ 0.6936 & 0.7121 & 0.7299 & 0.7469 & 0.8455 & 0.8492 & 1.0000 & \\ 0.6727 & 0.6924 & 0.7116 & 0.7300 & 0.8430 & 0.8483 & 0.8520 & 1.0000 \end{pmatrix}$
Unstructured	$\begin{pmatrix} 0.3189 & & & & & & & \\ 0.2708 & 0.3187 & & & & & & \\ 0.2646 & 0.2722 & 0.3178 & & & & & \\ 0.2565 & 0.2642 & 0.2696 & 0.3052 & & & & \\ 0.2254 & 0.2294 & 0.2309 & 0.2285 & 0.3042 & & & \\ 0.2198 & 0.2239 & 0.2232 & 0.2249 & 0.2633 & 0.2988 & & \\ 0.2173 & 0.2210 & 0.2208 & 0.2197 & 0.2553 & 0.2592 & 0.2948 & \\ 0.2165 & 0.2207 & 0.2209 & 0.2211 & 0.2535 & 0.2562 & 0.2588 & 0.2965 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8494 & 1.0000 & & & & & & \\ 0.8312 & 0.8553 & 1.0000 & & & & & \\ 0.8222 & 0.8471 & 0.8657 & 1.0000 & & & & \\ 0.7237 & 0.7368 & 0.7426 & 0.7499 & 1.0000 & & & \\ 0.7121 & 0.7256 & 0.7243 & 0.7448 & 0.8733 & 1.0000 & & \\ 0.7087 & 0.7210 & 0.7214 & 0.7325 & 0.8525 & 0.8733 & 1.0000 & \\ 0.7041 & 0.7180 & 0.7196 & 0.7350 & 0.8441 & 0.8608 & 0.8754 & 1.0000 \end{pmatrix}$

Table 5.6 – continued from previous page

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
Temporal Power	$\begin{pmatrix} 0.3199 & & & & & & & \\ 0.2809 & 0.3199 & & & & & & \\ 0.2473 & 0.2809 & 0.3199 & & & & & \\ 0.2182 & 0.2473 & 0.2809 & 0.3199 & & & & \\ 0.0830 & 0.0908 & 0.0998 & 0.1103 & 0.3199 & & & \\ 0.0762 & 0.0830 & 0.0908 & 0.0998 & 0.2809 & 0.3199 & & \\ 0.0704 & 0.0762 & 0.0830 & 0.0908 & 0.2473 & 0.2809 & 0.3199 & \\ 0.0653 & 0.0704 & 0.0762 & 0.0830 & 0.2182 & 0.2473 & 0.2809 & 0.3199 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8781 & 1.0000 & & & & & & \\ 0.7730 & 0.8781 & 1.0000 & & & & & \\ 0.6821 & 0.7730 & 0.8781 & 1.0000 & & & & \\ 0.2593 & 0.2837 & 0.3120 & 0.3447 & 1.0000 & & & \\ 0.2382 & 0.2593 & 0.2837 & 0.3120 & 0.8781 & 1.0000 & & \\ 0.2200 & 0.2382 & 0.2593 & 0.2837 & 0.7730 & 0.8781 & 1.0000 & \\ 0.2042 & 0.2200 & 0.2382 & 0.2593 & 0.6821 & 0.7730 & 0.8781 & 1.0000 \end{pmatrix}$
General Linear Lag-dependent	$\begin{pmatrix} 0.3068 & & & & & & & \\ 0.2655 & 0.3068 & & & & & & \\ 0.2600 & 0.2655 & 0.3068 & & & & & \\ 0.2552 & 0.2600 & 0.2655 & 0.3068 & & & & \\ 0.2228 & 0.2242 & 0.2282 & 0.2284 & 0.3068 & & & \\ 0.2207 & 0.2228 & 0.2242 & 0.2282 & 0.2655 & 0.3068 & & \\ 0.2187 & 0.2207 & 0.2228 & 0.2242 & 0.2600 & 0.2655 & 0.3068 & \\ 0.2162 & 0.2187 & 0.2207 & 0.2228 & 0.2552 & 0.2600 & 0.2655 & 0.3068 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8654 & 1.0000 & & & & & & \\ 0.8474 & 0.8654 & 1.0000 & & & & & \\ 0.8318 & 0.8474 & 0.8654 & 1.0000 & & & & \\ 0.7262 & 0.7307 & 0.7438 & 0.7444 & 1.0000 & & & \\ 0.7193 & 0.7262 & 0.7307 & 0.7438 & 0.8654 & 1.0000 & & \\ 0.7128 & 0.7193 & 0.7262 & 0.7307 & 0.8474 & 0.8654 & 1.0000 & \\ 0.7047 & 0.7128 & 0.7193 & 0.7262 & 0.8318 & 0.8474 & 0.8654 & 1.0000 \end{pmatrix}$

Table 5.7: Multivariate Multilevel Modelling: Fixed Parameters Estimates

	Unstructured		GLLD	
	Coeff	SE	Coeff	SE
Dummies for Occasion				
d_0	6.243	0.249	6.227	0.249
d_1	6.264	0.249	6.248	0.249
d_2	6.237	0.249	6.221	0.249
d_3	6.250	0.249	6.234	0.249
d_{12}	6.285	0.249	6.269	0.249
d_{13}	6.271	0.249	6.255	0.249
d_{14}	6.293	0.249	6.277	0.249
d_{15}	6.291	0.249	6.275	0.249
Males	0.649	0.075	0.650	0.075
White	0.224	0.023	0.226	0.023
Age (@ wave 1)	$4.209^\dagger$	$1.393^\dagger$	$4.310^\dagger$	$1.394^\dagger$
Squared term	$-0.440^\dagger$	$0.043^\dagger$	$-0.450^\dagger$	$0.043^\dagger$
Type of Worker				
(Employer as baseline)				
Informal	-0.139	0.016	-0.137	0.016
Formal	-0.047	0.016	-0.045	0.016
Military service	-0.008	0.021	-0.007	0.021
Self-Employed	-0.175	0.014	-0.174	0.014
Type of Activity				
(Manufacturing as baseline)				
Building	0.031	0.036	0.029	0.036
Commerce	-0.022	0.016	-0.023	0.016
Financial	0.039	0.020	0.039	0.020
Social Services	0.066	0.020	0.067	0.020
Domestic Services	0.000	0.019	-0.002	0.019
Other Services	0.001	0.018	0.000	0.018
Other Activities	0.028	0.061	0.038	0.061
Working Hours (in Log)	0.250	0.011	0.253	0.011
Proxy Respondent	0.003	0.006	0.003	0.006
Number of HH members	0.005	0.003	0.005	0.003
Metropolitan Region				
(Recife as baseline)				
Salvador	0.038	0.033	0.039	0.033
Belo Horizonte	0.245	0.032	0.244	0.032
Rio de Janeiro	0.211	0.029	0.212	0.029
São Paulo	0.359	0.032	0.359	0.032
Porto Alegre	0.226	0.034	0.226	0.034
Interaction Terms of Male and:				
Age (@ wave 1)	0.003	0.002	0.003	0.002
Education (@ wave 1)	0.054	0.013	0.054	0.013
Squared term	-0.003	0.001	-0.003	0.001
Type of Activity				
(Manufacturing as baseline)				
Building	-0.046	0.037	-0.044	0.037
Commerce	-0.021	0.018	-0.021	0.018
Financial	-0.056	0.022	-0.056	0.022
Social Services	-0.039	0.023	-0.042	0.023
Domestic Services	-0.122	0.034	-0.121	0.034
Other Services	-0.038	0.020	-0.036	0.020
Other Activities	-0.074	0.066	-0.084	0.065
Duration of Employment (×120)	0.018	0.009	0.019	0.009
Squared term	0.009	0.005	0.009	0.005
Working Hours (in Log)	-0.087	0.014	-0.087	0.014

Table 5.7 – continued from previous page

	Unstructured		GLLD	
	Coeff	SE	Coeff	SE
Proxy Respondent	-0.041	0.007	-0.041	0.007
Interaction Terms of White and:				
Type of Worker				
(Employer as baseline)				
Informal	-0.056	0.020	-0.060	0.020
Formal	-0.103	0.020	-0.104	0.020
Military service	-0.141	0.026	-0.144	0.026
Self-Employed	-0.013	0.017	-0.015	0.017
Occasion Dummies and:				
$d_0\times$ Education	-0.069	0.012	-0.069	0.012
$d_1\times$ "	-0.072	0.012	-0.072	0.012
$d_2\times$ "	-0.065	0.012	-0.065	0.012
$d_3\times$ "	-0.065	0.012	-0.065	0.012
$d_{12}\times$ "	-0.064	0.012	-0.064	0.012
$d_{13}\times$ "	-0.061	0.012	-0.061	0.012
$d_{14}\times$ "	-0.068	0.012	-0.068	0.012
$d_{15}\times$ "	-0.067	0.012	-0.067	0.012
$d_0\times$ Education^2	0.009	0.001	0.009	0.001
$d_1\times$ "	0.009	0.001	0.009	0.001
$d_2\times$ "	0.009	0.001	0.009	0.001
$d_3\times$ "	0.009	0.001	0.009	0.001
$d_{12}\times$ "	0.009	0.001	0.009	0.001
$d_{13}\times$ "	0.008	0.001	0.008	0.001
$d_{14}\times$ "	0.009	0.001	0.009	0.001
$d_{15}\times$ "	0.009	0.001	0.009	0.001
$d_0\times$ Duration of Employment	0.046	0.010	0.045	0.009
$d_1\times$ "	0.043	0.009	0.043	0.009
$d_2\times$ "	0.033	0.009	0.033	0.009
$d_3\times$ "	0.039	0.009	0.038	0.009
$d_{12}\times$ "	0.040	0.009	0.040	0.009
$d_{13}\times$ "	0.034	0.009	0.033	0.009
$d_{14}\times$ "	0.039	0.009	0.038	0.009
$d_{15}\times$ "	0.036	0.009	0.036	0.009
$d_0\times$ Duration of Employment^2	-0.022	0.005	-0.023	0.005
$d_1\times$ "	-0.022	0.005	-0.022	0.005
$d_2\times$ "	-0.021	0.005	-0.022	0.005
$d_3\times$ "	-0.024	0.005	-0.024	0.005
$d_{12}\times$ "	-0.020	0.005	-0.021	0.005
$d_{13}\times$ "	-0.018	0.005	-0.018	0.005
$d_{14}\times$ "	-0.019	0.005	-0.019	0.005
$d_{15}\times$ "	-0.017	0.005	-0.017	0.005
Contextual Effects:				
Proportion of Formal	-1.850	0.211	-1.847	0.211
Proportion of Informal	-2.187	0.247	-2.181	0.247
Proportion of Military	-2.080	0.233	-2.069	0.233
Proportion of Self-Employed	-2.274	0.245	-2.280	0.245
Average Education	0.072	0.006	0.072	0.006
Number of Clusters	1,761		1,761	
Number of Individuals	6,524		6,524	
-2× Log-Likelihood	19,760		20,007	

Note: † Values at 10^{-3}.

5.6 Summary and Discussion

The main objective of this chapter was to incorporate the PME rotation pattern characterized as 4-8-4 into the analysis. For that a longitudinal data set which included all the eight interviews for the employed heads of household of the PME survey was considered. This chapter presented a discussion on how the models considered so far accommodate such rotation patterns and proposed the use of models that impose an error covariance structure that is dependent on the temporal distance between the occasions. For this, a lag-matrix for the PME design was presented and generalized to other rotation patterns. Two alternative covariance structures were presented. These were the temporal power and the general linear lag-dependent. Both of these structures can be imposed to data with rotation patterns different to that of the PME. Models were fitted with the usual structures and compared to the two alternative structures. The general linear lag-dependent model presented good results overall, although still assuming homoscedasticity over time. Extensions to the analysis presented here would be to consider heterogeneous covariance structures. This would extend the general linear lag-dependent model, by not constraining the variances to be constant over time. One other possibility would be, for example, to assume equal variance over the first four time points not constrained to be the same as the equal variance for the last four time points.

Chapter 6

Probability-weighted Iterative Generalized Least Squares

6.1 Introduction

Chapter 2 presented a review of the methods for the analysis of longitudinal data under the multilevel model framework. That represented the fundamental theory necessary for the analyses performed in Chapters 4 to 5. Chapter 7 deals with the estimation of longitudinal multilevel models, such as those in the Chapter 5, using methods which compensate for panel non-response. This chapter, therefore, presents a review of methods of analysis of complex survey multilevel, and in particular longitudinal, data. These are methods that account for the sampling design features, like clustering and stratification, and the sampling weights. This review acts as the basis for the method developed later in this chapter which involves the extension of the probability-weighted iterative generalized least squares estimation method presented in Pfeffermann et al. (1998) to fit multivariate multilevel models.

6.2 Complex Survey Methods for Longitudinal and Multilevel Data

Most surveys adopt a sampling design that is more complex than the simple random sampling (SRS) with replacement. This has motivated the adaptation of traditional inference methods, initially developed under the assumption of SRS (Kish and Frankel, 1974), to account for such design features (see for example

(Skinner, 1986, 1989b; Pfeffermann and La Vange, 1989; Pessoa and Silva, 1998; Pfeffermann, 1993; Lee and Forthofer, 2005). The complex sampling designs are usually of multiple stages including stratification and clustering of the target population and unequal selection probability of the sampling units in each of the sampling stages (LaVange et al., 2001).

Stratification is often used to guarantee a better representation of the target population. This procedure divides the target population into homogeneous groups and constitutes a more efficient design than the SRS. Clustering is usually justified as a means to reduce costs. The clusters are usually homogeneous groups and the units selected within the same cluster are expected to be correlated. Unequal probabilities of selection are also used as a means to ensure a better representation of the target population. The inverse of these probabilities determine the sampling weights. Some surveys, like the Brazilian labour force survey (the PME), are initially designed as self-weighting surveys. However, in the presence of non-response, the sampling weights can be adjusted and the sample loses its self-weighting characteristic. Consequently the sampling weights can correct for non-response and can also be post-stratification weights adjusting for extreme values or to a known population pattern.

A sampling design is said to be informative when the selection indicators depend on the survey response variables. In this case, ignoring the sampling design leads to biased inference about most population parameters. Pfeffermann et al. (1998) stated that informative sampling designs are usually the case when unequal probabilities of selection are employed.

For descriptive inference the usual procedure to account for informative sampling design is to utilize the Horvitz-Thompson estimator. This is an unbiased estimator that accounts for sampling weights; see Cochran (1977) for more details. Two other estimators commonly used are the Ratio and the Regression estimators (Pessoa and Silva, 1998) which are usually associated with the Taylor linearisation method for the estimation of the variance. Authors such as Skinner (1989a); Binder (1983) and Kish and Frankel (1974) also mentioned other variance estimation methods such as the Balanced Repeated Replications (BRR), the Jackknife and the Bootstrap (refer to these authors for details).

For analytical inference, or inference about model parameters, the usual approach to account for informative sampling designs is to employ pseudo likelihood estimation methods (described later in this section) which make use of the sampling weights (Binder, 1983). This procedure is usually combined with the Taylor linearisation method for the estimation of the variance of the parameter estimates producing design consistent estimates (Skinner, 1989a, page 18). Nonetheless, the use of the sampling weights for analytical inference is debatable (Pfeffermann, 1993). This practice is recommended when estimating the model parameters. However, the inclusion of the sampling weights may increase the variance of these estimates. Pfeffermann (1993) suggested testing whether the sampling designs are or are not informative, once the inclusion of the sampling weights in the analysis would result in loss of efficiency for estimates. Some authors (Rabe-Hesketh and Skrondal, 2006; Pfeffermann et al., 1998; Pfeffermann, 1993) also argued that the inclusion of the design variables and sampling weights as covariates in the model would suffice to account for the complex sampling design. However, this approach is only advised if no change would be caused in the interpretation of the parameters. Moreover, this approach might not be possible as design variables are not usually available in the official data sets released. Furthermore, Rabe-Hesketh and Skrondal (2006) mentioned that design consistency can only be guaranteed if both the design and the sampling weights are taken into account.

The approach based on pseudo likelihood estimation methods is applied to single level data under complex survey sampling schemes. Also based on this approach, Pfeffermann et al. (1998) described a method for the analysis of two-level models for a continuous outcome which accounts for the unequal selection probabilities of units in each of the two levels of the model. Their proposal is justified as in a multilevel model the finite population units are not independent and the overall selection probabilities do not contain enough information for bias correction. Their method can only be implemented if the selection probabilities in each of the levels are provided in the data set, so that the weights for the level two and level one units can be calculated. The proposed estimation procedure adapts an IGLS analogue of the pseudo-maximum likelihood (PML), therefore called probability-weighted IGLS (PWIGLS), which is fully described in the next subsection 6.2.1. In Pfeffermann et al. (1998), three scenarios were tested in a simulation study. The simulated samples were based respectively on a non-informative design, on an informative design at level two only and on an informative design at both levels. Results showed that the standard IGLS estimates for samples with informative designs produced biased estimates, whereas the PWIGLS for the same samples had better design-based asymptotic characteristics.

Drawbacks of the method developed in Pfeffermann et al. (1998), as identified by Rabe-Hesketh and Skrondal (2006), were that it only accounted for a two-level multilevel model and was developed for a continuous response variable.

Hence, Rabe-Hesketh and Skrondal (2006) proposed extensions to generalized linear mixed models with multiple levels, based on the PML estimation via adaptive quadrature (implemented in `Stata` via `gllamm`). They considered scaled weights, performed a simulation study and concluded by noting the importance of accounting for weights under an informative design. However, they also noted a decrease in efficiency of the weighted estimators and advised that post-stratification weights might not be useful in these methods.

Rabe-Hesketh and Skrondal (2006) added that their method could also be applied to longitudinal models, albeit without accounting for the complex level one covariance structure. Skinner and Holmes (2003), on the other hand, presented two approaches to incorporate the complex sampling design in longitudinal random effects models while accounting for correlated responses of the same individual. However, the proposed methods, once again, only accommodate two-level longitudinal data in which the individuals are the level two and occasions are the level one units. Their first approach was based on a multivariate model utilizing weighted mean and covariance matrices. However, this approach can only handle monotone non-response and for short series it produces unstable estimates for the covariance matrix. Their second approach, stated as the most efficient, is a multilevel approach for a random intercept model and corresponds to an extension of the method presented in Pfeffermann et al. (1998) to allow for serial correlation between repeated responses of the same individual. To begin with, this approach makes use of the econometric method of first-differences (Wooldridge, 2002) that eliminates the random effects and then estimates the correlation parameters using weighted least squares. In a second step the regression parameters are estimated via PWIGLS from a transformed data set based on the lagged responses and accounting for the correlation estimated in the first step. This approach generates consistent estimates.

Skinner and Holmes (2003) also discussed how to include individual weights as level two weights and longitudinal weights as level one weights. For their specific analysis, individual level weights were defined as the longitudinal weight in the first wave for every individual in order to prevent a large weight variability. Given the definition of the individual level weights, occasion level weights were defined by dividing the longitudinal weights for every occasion by that for the first occasion. This makes the conditional level one weight at the first occasion equal to one. This was an alternative to ensure that the sample selection and the response process at the first wave is assumed to be the selection process throughout the analysis. Their scaling method was similar to the scaling method applied by Pfeffermann et al. (1998) so that the average of the scaled level one weights equals one and their sum equals the number of occasions the individuals are in the panel.

Other authors, such as Feder et al. (2000) and Asparouhov and Muthen (2006), proposed alternative methods to account for the complex survey design in the analysis of longitudinal data. Both methods were based upon the pseudo-maximum likelihood estimation procedure. One other important issue to notice is that the likelihood ratio test can no longer be applied when examining the goodness-of-fit of models estimated via pseudo-maximum likelihood (Skinner, 1989b). The Wald test can still be used provided that both the estimates and their standard errors were estimated taking into account the sampling design and weights of the complex survey data.

6.2.1 A Review on the Probability-weighted Iterative Generalized Least Squares

This sub-section describes in detail the method proposed by Pfeffermann et al. (1998) for the analysis of two-level data under complex survey design: the probability-weighted iterative generalized least squares (PWIGLS). As was briefly mentioned before, this method is based on the IGLS estimation method adapting it to an analogue of the pseudo-maximum likelihood estimation (PMLE) (Skinner, 1989a). This subsection starts by describing the IGLS, which is followed by a description of the PMLE and finally by a description of the PWIGLS.

6.2.1.1 Iterative Generalized Least Squares (IGLS)

Consider the two-level model presented in equation 2.6 rewritten as:

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_j + e_{ij}, \tag{6.1}$$

where $\boldsymbol{x}_{ij}$ is the vector of the explanatory variables for both levels of the data hierarchy and $\boldsymbol{z}_{ij}$ is the sub-set of the vector of explanatory variables that are considered as random at the second level, the cluster level. The same assumptions about the errors terms as those specified in the model in equation 2.6 are adopted here:

$$\boldsymbol{u}_j \sim N(0, \Sigma_u) \text{ and } e_{ij} \sim N(0, \sigma_e^2).$$

This model can be expressed in matrix form as:

$$\boldsymbol{Y}_j = X_j\boldsymbol{\beta} + \boldsymbol{r}_j \tag{6.2}$$

where X_j is a matrix of all explanatory variables for units in cluster j, $\boldsymbol{Y}_j$ is a vector with the responses for units in cluster j and $\boldsymbol{r}_j$ is a vector of the composite errors given as:

$$\boldsymbol{r}_j = Z_j\boldsymbol{u}_j + \boldsymbol{e}_j,$$

where Z_j is a design matrix. For this model the total variance for the observations within cluster j is given as:

$$V_j = Z_j\Sigma_u Z_j^T + I_{n_j}\sigma_e^2,$$

where I_{nj} is an identity matrix size $n_j \times n_j$.

Defining $\boldsymbol{\theta}$ to be the row vector of s distinct covariance terms of V_j so that $\boldsymbol{\theta} = (\theta_1, ..., \theta_s) = (vech(\Sigma_u)^T, \sigma_e^2)$, V_j is expressed as a linear function of $\boldsymbol{\theta}$ such that:

$$V_j = \sum_{k=1}^{s} \theta_k G_{kj},$$

where $s = (q(q+1)/2) + 1$ as defined in Pfeffermann et al. (1998) and G_{kj} is given as:

$$G_{kj} = Z_j H_{kj} Z_j^T + I_{n_j}\delta_{ks} \qquad \text{for } k = (1, 2, \ldots, s),$$

where the H_{kj} are $q \times q$ matrices of zeroes and ones and δ_{ks} is the Kronecker delta defined as:

$$\delta_{ks} = \begin{cases} 1 & \text{if } k = s \\ 0 & \text{otherwise} \end{cases}.$$

Example 6.1 Consider a two-level random intercept model where $q = 1$ and $s = 2$. Hence, $\Sigma_u = [\sigma_u^2]$, $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\sigma_u^2, \sigma_e^2)$ and V_j can be expressed as:

$$V_j = \theta_1 G_{1j} + \theta_2 G_{2j}.$$

For this model formulation, H_{kj} and δ_{ks} are defined as:

$$H_{1j} = [1]\ ,\ H_{2j} = [0]\ ,\ \delta_{12} = [0]\ \text{ and } \delta_{22} = [1]\ .$$

■

Given a set of initial values (stage 0), the IGLS iterates between the estimation of $\hat{\boldsymbol{\beta}}$ (stage 1) and $\hat{\boldsymbol{\theta}}$ (stage 2) until convergence following the stages:

Stage 0: This stage calculates the initial values for both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$. The estimates for $\hat{\boldsymbol{\beta}}$ are usually taken to be:

$$\hat{\boldsymbol{\beta}}^{(0)}_{OLS} = \left(\sum_j X_j^T X_j\right)^{-1}\left(\sum_j X_j^T \boldsymbol{Y}_j\right).$$

For the initial values of $\hat{\boldsymbol{\theta}}$: $\hat{\sigma}_u^{2(0)}$ are usually taken to be 0 and $\hat{\sigma}_e^{2(0)}$ estimated as

$$\hat{\sigma}_e^{2(0)} = \frac{\sum_j \left(\sum_i \left(e_{ij}^{(0)} - \bar{e}_j\right)^2\right)}{\sum_j (n_j - 1)},$$

where $\sum_j$ denotes the sum over the $j = (1, 2, ..., n)$ clusters, $e_{ij}^{(0)}$ are the raw residuals calculated as $e_{ij}^{(0)} = y_{ij} - \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}^{(0)}_{OLS}$ and $\bar{e}_j$ is the average residual for cluster j. With $\hat{\boldsymbol{\theta}}^{(0)}$ estimated to compose $\hat{V}_j^{(r-1)}$ for the first iteration $(r = 1)$ the new estimates for the fixed effects are calculated in stage 1.

Stage 1: At this stage $\hat{\boldsymbol{\beta}}^{(r)}_{IGLS}$, where the superscript (r) denotes the iteration number, are calculated based on the general GLS formulation given as:

$$\hat{\boldsymbol{\beta}}^{(r)}_{IGLS} = \left(\sum_j X_j^T \hat{V}_j^{-1(r-1)} X_j\right)^{-1}\left(\sum_j X_j^T \hat{V}_j^{-1(r-1)} \boldsymbol{Y}_j\right). \quad (6.3)$$

The general form of $\hat{V}_j$ for the two-level random coefficients model is:

$$\hat{V}_j = Z_j \hat{\Sigma}_u Z_j^T + I_{n_j} \hat{\sigma}_e^2,$$

Pfeffermann and La Vange (1989) wrote its inverse as

$$\hat{V}_j^{-1} = \frac{1}{\hat{\sigma}_e^2} I_{n_j} - \frac{1}{\hat{\sigma}_e^2} Z_j (Z_j^T Z_j + \hat{\sigma}_e^2 \hat{\Sigma}_u^{-1})^{-1} Z_j^T.$$

Defining

$$A_j = (Z_j^T Z_j + \hat{\sigma}_e^2 \hat{\Sigma}_u^{-1})^{-1},$$

$\hat{V}_j^{-1}$ can then be written as:

$$\hat{V}_j^{-1} = \frac{1}{\hat{\sigma}_e^2} I_{n_j} - \frac{1}{\hat{\sigma}_e^2} Z_j A_j Z_j^T.$$

$\hat{V}_j^{-1(r-1)}$ can assume the same form as $\hat{V}_j^{-1}$ for each iteration (r). Substituting this expression for $\hat{V}_j^{-1(r-1)}$ in the equation 6.3, $\hat{\boldsymbol{\beta}}_{IGLS}^{(r)}$ can be re-expressed such that

$$\hat{\boldsymbol{\beta}}_{IGLS}^{(r)} = \hat{P}^{(r)-1}\hat{Q}^{(r)} \tag{6.4}$$

for $\hat{P}^{(r)}$ and $\hat{Q}^{(r)}$ written as

$$\begin{aligned}\hat{P}^{(r)} &= \sum_j X_j^T \hat{V}_j^{-1(r-1)} X_j \\ &= (\hat{\sigma}_e^2)^{-1} \sum_j (X_j^T X_j - X_j^T Z_j A_j Z_j^T X_j),\end{aligned}$$

and

$$\begin{aligned}\hat{Q}^{(r)} &= \sum_j X_j^T \hat{V}_j^{-1(r-1)} \boldsymbol{Y}_j \\ &= (\hat{\sigma}_e^2)^{-1} \sum_j (X_j^T \boldsymbol{Y}_j - X_j^T Z_j A_j Z_j^T \boldsymbol{Y}_j).\end{aligned}$$

The fixed coefficients are thereby estimated for iteration (r). Note in the above formulae that for simplicity of notation, the superscript $(r-1)$ was omitted from the components of $\hat{V}_j^{(r-1)}$: $\hat{\sigma}_e^2$ and $\hat{\Sigma}_u$. The subscript (r) was omitted from A_j.

Stage 2 After the estimation in stage 1, the vector for the raw residuals $\boldsymbol{e}_j^{(r)}$ is calculated to be used in the estimation of $\hat{\boldsymbol{\theta}}^{(r)}$ which can be estimated following the formulation presented in Pfeffermann et al. (1998)

$$\hat{\boldsymbol{\theta}}_{IGLS}^{(r)} = \hat{R}^{(r)-1}\hat{S}^{(r)}, \tag{6.5}$$

where $\hat{R}^{(r)}$ is a $s \times s$ matrix and $\hat{S}^{(r)}$ is a $s \times 1$ vector with elements respectively defined as:

$$\sum_j tr(\hat{V}_j^{(r-1)-1} G_{kj} \hat{V}_j^{(r-1)-1} G_{lj})$$

and

$$\sum_j tr(\boldsymbol{e}_j^{(r)^T} \hat{V}_j^{(r-1)-1} G_{kj} \hat{V}_j^{(r-1)-1} \boldsymbol{e}_j^{(r)}),$$

where subscripts k and l denote respectively the rows and columns of the matrices and

$$\boldsymbol{e}_j^{(r)} = (\boldsymbol{Y}_j - X_j \hat{\boldsymbol{\beta}}_{IGLS}^{(r)}). \tag{6.6}$$

These equations result from the partial derivatives of the log-likelihood of the model in equation 6.1 with respect to $\boldsymbol{\theta}$. The log-likelihood and the corresponding detailed algebra for obtaining the solution for the matrices $\hat{R}^{(r)}$ and $\hat{S}^{(r)}$ were presented in Zhu (2008). These solutions are the same as those introduced by Pfeffermann et al. (1998) in their Appendix A, which are reproduced below for the general form where $q > 1$. The element in the k^{th} row and l^{th} column of the $s \times s$ $\hat{R}^{(r)}$ matrix is given by:

$$\hat{R}^{(r)}[k,l] = \frac{1}{\hat{\sigma}_e{}^{4(r-1)}} \sum_j \{\delta_{ks}\delta_{ls}n_j + \delta_{ls}tr(Z_j^T D_j^{-1} Z_j C_{kj}) + \delta_{ks}tr(Z_j^T D_j^{-1} Z_j H_{lj})$$
$$+ tr(Z_j^T D_j^{-1} Z_j C_{kj} Z_j^T D_j^{-1} Z_j H_{lj})\},$$

where

$$C_{kj} = B_{kj} - \delta_{ks}A_j - B_{kj}Z_j^T D_j^{-1} Z_j A_j, \tag{6.7}$$
$$B_{kj} = \hat{\sigma}_e^2 A_j \hat{\Sigma}_u^{-1} H_{kj} - \delta_{ks} Aj, \tag{6.8}$$

and the k^{th} row element of the $s \times 1$ matrix $\hat{S}^{(r)}$, is given by:

$$\hat{S}^{(r)}[k] = \frac{1}{\hat{\sigma}_e{}^{4(r-1)}} \sum_j \{\delta_{ks}tr(\boldsymbol{e}_j^T D_j^{-1} \boldsymbol{e}_j) + tr(\boldsymbol{e}_j^T D_j^{-1} Z_j C_{kj} Z_j^T D_j^{-1} \boldsymbol{e}_j)\}.$$

It is worth noting that the superscript (r) for the current iteration was omitted from matrices A_j, B_{kj} and C_{kj} and from $\boldsymbol{e}_j$ in the above formulae. These matrices and this vector are calculated for each iteration based on the elements of $\hat{V}_j^{(r-1)}$. It is also important to notice that D_j^{-1} assumes the form of an identity matrix size $n_j \times n_j$, I_{nj}, for the case of the random intercept model under IGLS estimation.

The final iteration provides the estimates for both the fixed and random parts of the model in 6.1. These are the ML estimates if the normality assumption holds. The covariance matrix for the estimates of the fixed coefficients is given as

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_{IGLS}) = \hat{P}^{-1},$$

where $\hat{P}$ is taken from the final iteration. Furthermore, by analogy with the formulae presented in Goldstein (2003), it is straightforward to deduce that

$$\widehat{Var}(\hat{\boldsymbol{\theta}}_{IGLS}) = 2 \times \hat{R}^{-1},$$

where $\hat{R}$ is also taken from the final iteration.

6.2.1.2 Pseudo-maximum Likelihood (PML)

The pseudo-maximum likelihood (PML) estimation is one of the approaches that take into account the sampling weights and the sampling design characteristics when dealing with inference for regression parameters for survey data under complex survey design. Proposed by Skinner (1989a), based on the ideas proposed by Binder (1983), the PML modifies log-likelihood functions by incorporating the weights so that the estimated likelihood equations are equivalent to the census equations (Pfeffermann, 1993).

Consider the case of a parametric model for the population where $\boldsymbol{\varphi}$ represents all the parameters to be estimated. The census maximum likelihood estimators for $\boldsymbol{\varphi}$ are those that maximize the log-likelihood

$$l_U(\boldsymbol{\varphi}) = \sum_U \log f_i(y_i, \boldsymbol{\varphi}),$$

where $f_i(y_i, \boldsymbol{\varphi})$ is the density of y_i in the population. The census ML estimate is obtained by solving the census likelihood equations

$$\frac{\partial l_U(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} = \sum_U \boldsymbol{u}_i(\boldsymbol{\varphi}) = \sum_U \frac{\partial}{\partial \boldsymbol{\varphi}} \log f_i(y_i, \boldsymbol{\varphi}) = \mathbf{0},$$

where $\sum_U$ represents the sum across all the elements of the finite population U (Pessoa and Silva, 1998). These equations can only be solved if all the units in the population are observed. One approach to overcome this problem is to apply the sampling estimation principles and the Hovitz-Thompson (HT) estimator (Cochran, 1977) which uses the inverse of the selection probabilities (π_i) as weights (w_i). The HT estimator $\hat{\boldsymbol{T}}(\boldsymbol{\varphi})$ for the vector of the population totals, i.e. the sum of scores in the population,

$$\boldsymbol{T}(\boldsymbol{\varphi}) = \sum_U \boldsymbol{u}_i(\boldsymbol{\varphi})$$

is given by

$$\hat{\boldsymbol{T}}(\boldsymbol{\varphi}) = \sum_S w_i \boldsymbol{u}_i(\boldsymbol{\varphi}),$$

where $w_i = 1/\pi_i$. The pseudo-maximum likelihood estimator $\hat{\boldsymbol{\varphi}}_{PML}$ is the solution for

$$\hat{\boldsymbol{T}}(\boldsymbol{\varphi}) = 0.$$

Under general conditions, $\hat{\boldsymbol{\varphi}}_{PML}$ is a consistent estimator of $\boldsymbol{\varphi}$. In general the PML estimates do not share the same asymptotic characteristics as the ML estimates do (Skinner, 1989a). However, the PML estimates are robust.

For the estimation of the covariance matrix of the PML estimates, $V(\hat{\boldsymbol{\varphi}}_{PML})$, Skinner (1989a) adopted the delta method also known as the Taylor linearisation method, which in a general form is:

$$\hat{V}_L(\hat{\boldsymbol{\varphi}}_{PML}) = I(\hat{\boldsymbol{\varphi}}_{PML})^{-1} \hat{V}_L\left(\sum_S w_i \boldsymbol{u}_i(\hat{\boldsymbol{\varphi}}_{PML})\right) I(\hat{\boldsymbol{\varphi}}_{PML})^{-1} \tag{6.9}$$

where

$$I(\hat{\boldsymbol{\varphi}}_{PML})^{-1} = \frac{\partial \hat{\boldsymbol{T}}(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}^T} = \sum_S w_i \frac{\partial \boldsymbol{u}_i}{\partial \boldsymbol{\varphi}^T}\bigg|_{\boldsymbol{\varphi}=\hat{\boldsymbol{\varphi}}_{PML}},$$

and the middle term of 6.9 is defined as

$$\hat{V}_L\left(\sum_S w_i \boldsymbol{u}_i(\hat{\boldsymbol{\varphi}}_{PML})\right) = \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_j (\boldsymbol{d}_{hj} - \bar{\boldsymbol{d}}_h)(\boldsymbol{d}_{hj} - \bar{\boldsymbol{d}}_h)^T. \tag{6.10}$$

Here $\boldsymbol{d}_{hj}$ is defined in Skinner (1989a) as $\sum_{S_{hj}} w_i \boldsymbol{u}_i(\hat{\boldsymbol{\varphi}}_{PML})$ in each PSU (cluster) j in stratum h.

Pessoa and Silva (1998) added that PML estimation accounts for the weights through the estimating equations of the parameters and their covariance and accounts for the complex sampling design through the expression for the variance of the total scores and through the inclusion probabilities.

For the case of a multilevel model, Pfeffermann et al. (1998) stated that the PML estimation is not as straightforward as for the one-level model since the population values are no longer assumed to be independent. Therefore the census log-likelihood can no longer be expressed as simple sums of the units' contributions but as a sum across all levels instead. For a two-level model, for example, this could be written as (Grilli and Rampichini, 2006):

$$\ell(\boldsymbol{\varphi}) = \sum_{j=1}^{N} \log \int \left(\exp \left\{ \sum_{i=1}^{N_j} \log f_{ij}(y_{ij}, \boldsymbol{\varphi}|\boldsymbol{u}_j) \right\} \right) \phi(\boldsymbol{u}_j) d\boldsymbol{u}_j,$$

where $\phi(\boldsymbol{u}_j)$ is the multivariate normal distribution of the level two random effects and $\log f_{ij}(y_{ij}, \boldsymbol{\varphi}|\boldsymbol{u}_j)$ is the log-likelihood contributions of the level one units conditioned on the level two random effects (Rabe-Hesketh and Skrondal, 2006). For this reason, the log-likelihood for the sample units cannot be expressed as a simple weighted sum of the sample contributions either, but rather as:

$$\hat{\ell}(\boldsymbol{\varphi}) = \sum_{j} w_j \log \int \left(\exp \left\{ \sum_{S} w_{i|j} \log f_{ij}(y_{ij}, \boldsymbol{\varphi}|\boldsymbol{u}_j) \right\} \right) \phi(\boldsymbol{u}_j) d\boldsymbol{u}_j.$$

Therefore, the main difference is that the sum for each of the levels requires the respective conditional sampling weight. These weights are defined in Rabe-Hesketh and Skrondal (2006, page 811). For the above case, of a two-level model, the sum over the level one units are weighted by $w_{i|j}$, that is the inverse of the selection probability of the i^{th} unit at cluster j given that cluster j has been selected; and the sum over level two units are weighted by w_j, that is the inverse of the selection probability of cluster j.

6.2.1.3 Probability-weighted Iterative Generalized Least Squares (PWIGLS)

The probability-weighted iterative generalized least squares (PWIGLS) method adapts the IGLS estimation by incorporating the ideas of the PML estimation. Consider the same two-level model as the one described for IGLS in equation 6.2. Also consider that the selection probabilities for each of the levels are available for use in the multilevel data. Hence, π_j is the selection probability for cluster j and $\pi_{i|j}$ is the conditional selection probability of unit i in cluster j given that cluster j has been selected. These probabilities are such that

$$\pi_{ij} = \pi_j \times \pi_{i|j}$$

are the unconditional sample inclusion probabilities (Pfeffermann et al., 1998). The inverse of the conditional probabilities reflect the sampling weights for each of the levels of the data.

The underlying idea here is similar to that for the PML estimation. Firstly, Pfeffermann et al. (1998) wrote the IGLS estimation for the census parameters, that is for the case where all the units of the finite population U are observed. Secondly, the census estimators were substituted by weighted sample estimates by replacing each population sum over the level two units by the weighted sample sum using w_j and each population sum over the level one units by a weighted sample sum using $w_{i|j}$. Here w_j are the level two weights and $w_{i|j}$ are the level one weights.

The equivalent stages for the PWIGLS are then:

Stage 0: Like for the IGLS, in this stage the initial values for both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are calculated. First, let

$$D_j^{-1} = \text{diag}(w_{i|j}), \tag{6.11}$$

that is a $n_j \times n_j$ diagonal matrix with $w_{i|j}$ in the main diagonal. Then consider that w_j is a scalar, in this way the initial estimates for $\hat{\boldsymbol{\beta}}$ are given as:

$$\hat{\boldsymbol{\beta}}^{(0)}_{PWIGLS} = \left(\sum_j w_j (X_j^T D_j^{-1} X_j) \right)^{-1} \left(\sum_j w_j (X_j^T D_j^{-1} \boldsymbol{Y}_j) \right).$$

For the initial values of $\hat{\boldsymbol{\theta}}$ it is also assumed that $\hat{\sigma}_u^{2(0)}$ is equal 0. Given that the raw residuals can be calculated as

$$\boldsymbol{e}_j^{(0)} = \boldsymbol{Y}_j - X_j \hat{\boldsymbol{\beta}}^{(0)}_{PWIGLS}$$

and

$$\hat{\boldsymbol{u}}_j^{(0)} = \frac{\sum_i w_{i|j} e_{ij}^{(0)}}{\sum_i w_{i|j}},$$

following the suggestions in Pfeffermann et al. (1998), $\hat{\sigma}_e^{2(0)}$ is estimated as:

$$\hat{\sigma}_e^{2(0)} = \frac{\sum_j w_j \left(\sum_i (w_{i|j} (\boldsymbol{e}_j^{(0)} - \hat{\boldsymbol{u}}_j^{(0)})^2 \right)}{\sum_j w_j ((\sum_i w_{i|j}) - 1)},$$

where $\sum_j$ denotes the sum over the $j = (1, 2, ..., n)$ clusters in the sample and $\sum_i$ denotes the sum over the level one units sampled in cluster j. From this stage $\hat{\boldsymbol{\theta}}^{(0)}$ is estimated and that forms matrix $\hat{V}_j^{(r-1)}$ to be used in the next stage (iteration $r = 1$).

Stage 1: Using the estimated $\hat{V}_j^{(r-1)}$ from previous iteration, $\hat{\boldsymbol{\beta}}^{(r)}_{PWIGLS}$ is estimated adopting the same formulation as in 6.4, i.e.

$$\hat{\boldsymbol{\beta}}^{(r)}_{PWIGLS} = \hat{P}^{(r)-1}\hat{Q}^{(r)}, \tag{6.12}$$

where $\hat{P}^{(r)}$ and $\hat{Q}^{(r)}$ are now defined as:

$$\begin{aligned}\hat{P}^{(r)} &= \sum_j w_j(X_j^T \hat{V}_j^{(r-1)-1} X_j)\\ &= (\hat{\sigma}_e^2)^{-1} \sum_j w_j(X_j^T D_j^{-1} X_j - X_j^T D_j^{-1} Z_j A_j Z_j^T D_j^{-1} X_j),\end{aligned}$$

and

$$\begin{aligned}\hat{Q}^{(r)} &= \sum_j w_j(X_j^T \hat{V}_j^{(r-1)-1} \mathbf{Y}_j)\\ &= (\hat{\sigma}_e^2)^{-1} \sum_j w_j(X_j^T D_j^{-1} \mathbf{Y}_j - X_j^T D_j^{-1} Z_j A_j Z_j^T D_j^{-1} \mathbf{Y}_j).\end{aligned}$$

Here, D_j^{-1} is as defined in equation 6.11 and for each iteration (r), but for simplicity of notation omitting the superscripts (r) from A_j and $(r-1)$ from $\hat{V}_j^{-1}$,

$$A_j = (Z_j^T D_j^{-1} Z_j + \hat{\sigma}_e^2 \hat{\Sigma}_u^{-1})^{-1} \text{ and} \tag{6.13}$$

$$\hat{V}_j^{-1} = \hat{\sigma}_e^{-2} D_j^{-1} - \hat{\sigma}_e^{-2} D_j^{-1} Z_j A_j Z_j^T D_j^{-1}. \tag{6.14}$$

Note that $\hat{\sigma}_e^2$ and $\hat{\Sigma}_u$ are elements of matrix $\hat{V}_j^{(r-1)}$ and also have their $(r-1)$ superscripts omitted.

Stage 2 From the estimation in stage 1, the raw residuals are calculated to be used in the estimation of $\hat{\boldsymbol{\theta}}^{(r)}$, such that, as in equation 6.5, it is given as:

$$\hat{\boldsymbol{\theta}}^{(r)}_{PWIGLS} = \hat{R}^{(r)-1}\hat{S}^{(r)}, \tag{6.15}$$

where the elements of $\hat{R}^{(r)}$ and $\hat{S}^{(r)}$ are now written as:

$$\begin{aligned}\hat{R}^{(r)}[k,l] = \frac{1}{\hat{\sigma}_e{}^{4(r-1)}} \sum_j w_j\{&\delta_{ks}\delta_{ls}(\sum_i w_{i|j}) + \delta_{ls} tr(Z_j^T D_j^{-1} Z_j C_{kj})\\ &+ \delta_{ks} tr(Z_j^T D_j^{-1} Z_j H_{lj}) + tr(Z_j^T D_j^{-1} Z_j C_{kj} Z_j^T D_j^{-1} Z_j H_{lj})\}\end{aligned}$$

and

$$\hat{S}^{(r)}[k] = \frac{1}{\hat{\sigma}_e{}^{4(r-1)}} \sum_j w_j \{ \delta_{ks} tr(\hat{\boldsymbol{e}}_j^T D_j^{-1} \hat{\boldsymbol{e}}_j) + tr(\hat{\boldsymbol{e}}_j^T D_j^{-1} Z_j C_{kj} Z_j^T D_j^{-1} \hat{\boldsymbol{e}}_j) \},$$

where B_{kj} and C_{kj} are as in equations 6.8 and 6.7, A_j is defined as in equation 6.13, D_j^{-1} is defined as in 6.11 and $\hat{\boldsymbol{e}}_j$ are the raw residuals which can be calculated similarly to equation 6.6 but using $\hat{\boldsymbol{\beta}}_{PWIGLS}^{(r)}$ instead.

The final iteration provides the estimates for both the fixed and random parts of the model in 6.1. Like the PML estimates, the PWIGLS estimates are design consistent and model consistent under weak regularity conditions (Rabe-Hesketh and Skrondal, 2006, page 808). These conditions, however, require that the number of level two units n and the number of level one units n_j within cluster j increase, which might not usually be the case (Pfeffermann et al., 1998, page 29).

For the estimation of the covariance matrix for the PWIGLS estimates Pfeffermann et al. (1998) applied the Taylor linearisation methods as described in Skinner (1989a) for the PML estimation. This method is based on the randomization variance (Pfeffermann, 1993), assuming that the level two units were selected with replacement and that the contributions to the pseudo likelihood are independent. The Taylor linearisation method provides robust standard error estimates in the form of the sandwich estimator (Huber, 1967; White, 1982; Freedman, 2006). Pfeffermann et al. (1998) provided the formulation for these variances for the case of the random intercept model as:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_{PWIGLS}) = \hat{P}^{-1} \left(\frac{n}{n-1} \right) \left(\sum_j w_j^2 \boldsymbol{c}_j \boldsymbol{c}_j^T \right) \hat{P}^{-1}, \tag{6.16}$$

where $\boldsymbol{c}_j = (X_j \hat{V}_j^{(-1)} \hat{\boldsymbol{e}}_j)$, and by using the same kind of substitution as for $\hat{P}^{(r)}$ and $\hat{Q}^{(r)}$, $\boldsymbol{c}_j$ can be written as

$$\boldsymbol{c}_j = (\hat{\sigma}_e^2)^{-1} \sum_j (X_j^T D_j^{-1} \hat{\boldsymbol{e}}_j - X_j^T D_j^{-1} Z_j A_j Z_j^T D_j^{-1} \hat{\boldsymbol{e}}_j).$$

Given that $\hat{\boldsymbol{\theta}}_{PWIGLS}$ is estimated by equation 6.15 and $\hat{R}$ is a Jacobian matrix, an estimate for $Var(\hat{\boldsymbol{\theta}}_{PWIGLS})$ can be given as:

$$\begin{aligned}\widehat{Var}(\hat{\boldsymbol{\theta}}_{PWIGLS}) &= \widehat{Var}(\hat{R}^{-1}\hat{S}) \\ &= \hat{R}^{-1}\hat{V}_L(\hat{S})\hat{R}^{-1},\end{aligned}$$

where $\hat{V}_L(\hat{S})$ can be found using the same principle as in equation 6.10. Therefore $\widehat{Var}(\hat{\boldsymbol{\theta}}_{PWIGLS})$ is equal to

$$\hat{R}^{-1}\left(\frac{n}{n-1}\right)\left(\sum_j w_j^2(\hat{S}-\hat{R}\hat{\boldsymbol{\theta}}_{PWIGLS})(\hat{S}-\hat{R}\hat{\boldsymbol{\theta}}_{PWIGLS})^T\right)\hat{R}^{-1}. \tag{6.17}$$

It is worth mentioning that all the above formulation can also be applied to the random coefficient models where $q > 1$. Care must be taken however on the specification of matrices H_{kj} and on the determination of the initial values for $\hat{\boldsymbol{\theta}}^{(0)}$. When estimating random coefficient models through `gllamm` (Rabe-Hesketh et al., 2004) it was observed that the usual initial values for the level-two variance terms are 0.5 and for the covariance terms 0.

6.2.2 Scaled Weights

Most of the discussion presented in Pfeffermann et al. (1998) concerned scaling methods that can be applied to the multilevel weights. This issue was also raised by Rabe-Hesketh and Skrondal (2006); Grilli and Rampichini (2006) as an approach to reduce the bias generated due to small samples. However, Pfeffermann et al. (1998) showed that different scaling methods affect the estimates in different ways. Their so called “method two” was preferred as the best potential scaling method under informative weights. This particular method multiplies the level one weights $w_{i|j}$ by a constant λ_j that represents the inverse of the average weight in cluster j so that the sum of the scaled level one weights represents the actual cluster size n_j. Following this specification, the scaled weights can be writen as

$$w^*_{i|j} = \lambda_j \times w_{i|j} = \frac{w_{i|j} \times n_j}{\sum_i w_{i|j}}.$$

For a two-level multilevel model, scaling is only required for the level one weights, since the multiplication of level two weights by a constant would only re-scale the pseudo likelihood having no effects on the estimates (Rabe-Hesketh and Skrondal, 2006). However, the scaling of level one weights is expected not only to have an effect on the estimates of the fixed part of the model but also a bigger effect on the estimation of the random part.

The PWIGLS is not readily available in the traditional statistical computer packages. Chantala et al. (2006) presented a brief comparison of packages that can perform the PWIGLS. These include, for example, `MLwiN`, `gllamm`, MPLUS and LISREL. They estimated the same model in the different packages providing some highlights of the main differences and care that the user should take. It is worth mentioning that computer code developed for the PWIGLS estimation is available in Appendix D. These codes were written using the `Mata` language for `Stata`.

6.3 PWIGLS for Multivariate Multilevel Models

This section presents an extension to the estimation method of probability-weighted iterative generalized least squares (PWIGLS) to accommodate the estimation of two-level multivariate (longitudinal) models. The PWIGLS for the estimation of two-level random coefficients models was described in the previous section. The description presented here uses the same notation and the main differences are highlighted. This estimation method is used in the analysis presented in the next chapter.

First consider the two-level multivariate model with only the time variable as a covariate where random coefficients are allowed at the PSU level:

$$y_{tij} = \boldsymbol{d}_{tij}^T\boldsymbol{\beta} + \boldsymbol{z}_{tij}^T\boldsymbol{v}_j + \boldsymbol{d}_{tij}^T\boldsymbol{u}_{ij}. \tag{6.18}$$

This is a multivariate multilevel model equivalent to the model in equation 2.23. Here $\boldsymbol{z}_{tij}$ is a sub-set of the vector of explanatory variables which are considered as random at the PSU level. The vector $\boldsymbol{z}_{tij}$ is associated with the vector of random effects at the cluster level which may include the random intercept and random slopes. For the models in this section only the intercept is considered as random at the PSU level. Therefore $\boldsymbol{z}_{tij}$ is a vector of ones. Here it is also assumed that

$$\boldsymbol{v}_j \sim MN(0, \Sigma_v) \text{ and } \boldsymbol{u}_{ij} \sim MN(\boldsymbol{0}, \Sigma_u).$$

Therefore, when the vector of random effects $\boldsymbol{v}_j$ at the PSU level includes only the random intercepts, $\Sigma_v = \sigma_v^2$.

The multivariate multilevel model in equation 6.18 can be re-expressed in matrix form, as in equation 6.2, as

$$\boldsymbol{Y}_j = X_j\boldsymbol{\beta} + \boldsymbol{r}_j$$

where, as before, the subscript j is for the PSU, where X_j is the matrix of explanatory variables, $\boldsymbol{Y}_j$ is the vector with the response variable and $\boldsymbol{r}_j$ is the vector of composite residual now given as

$$r_{tij} = \boldsymbol{z}_{tij}^T\boldsymbol{v}_j + \boldsymbol{d}_{tij}^T\boldsymbol{u}_{ij}.$$

Here it is also assumed that

$$\boldsymbol{r}_j \sim N(0, V_j) \ .$$

For the multivariate model 6.18, V_j takes the form

$$V_j = Z_j\Sigma_v Z_j^T + D_j(I_{n_j} \otimes \Sigma_u),$$

where Z_j is the design matrix for PSU j and Σ_v is the covariance matrix for the PSU random effects. In addition consider that n_j is the number of heads of household in PSU j, each head of household is measured in T_{ij} occasions. Assume now that the number of occasions is fixed within heads of household and equal to T. Therefore, the number of observations within PSU j is $n_{tj} = T \times n_j$. Matrix D_j is the $n_{tj} \times n_{tj}$ matrix defined as

$$D_j = [I_{n_j} \otimes D_{ij}],$$

where D_{ij} is the matrix formed with the vectors of the occasion dummies $\boldsymbol{d}_{tij}$, I_{n_j} is the identity matrix with size $n_j \times n_j$ and the symbol $\otimes$ represents the Kronecker product. Furthermore, Σ_u is the $T \times T$ covariance matrix of the random effects at the heads of household level. Note that the cross-product between I_{n_j} and the covariance matrix Σ_u provides the block-diagonal structure at the cluster level, where each block represents one head of household.

Example 6.2 Consider a three-level balanced data set where $T_{ij} = 4$ for every head of household i in PSU j. Also consider that the model in equation 6.18 only includes the random intercept at the PSU level. Therefore $\Sigma_v = \sigma_v^2$ and

$Z_j = \mathbf{1}_{(n_{tj}\times 1)}$. Furthermore,

$$D_j = [I_{n_j} \otimes D_{ij}] = I_{n_j} \otimes \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence, for this special case D_j is equivalent to the $n_{tj} \times n_{tj}$ identity matrix $I_{n_{tj}}$. In this example V_j is given as:

$$V_j = J_{n_{tj}}\sigma_v^2 + D_j(I_{n_j} \otimes \Sigma_u),$$

where $J_{n_{tj}}$ is a $n_{tj} \times n_{tj}$ matrix of ones in every entry. ■

The matrix V_j can be expressed as a linear function of $\boldsymbol{\theta}$ such that:

$$V_j = \sum_{k}^{s} \theta_k G_{kj},$$

where $\boldsymbol{\theta}$ is the row vector formed with the s distinct elements of Σ_v and Σ_u. For the multivariate multilevel model, G_{kj} is given as:

$$G_{kj} = Z_j H_{kj} Z_j^T + D_j(I_{n_j} \otimes \Delta_{kj}).$$

The main differences between the PWIGLS for the multivariate model and the PWIGLS described in the previous section are in the form of V_j and the definition of matrices H_{kj} and Δ_{kj}. As before, H_{kj} are $q \times q$ matrices of zeroes and ones where $k = 1, ..., s$, where s is the total number of parameters in $\boldsymbol{\theta}$ and q, here, is the number of random effects at the PSU level. The Δ_{kj} matrices are s $T \times T$ matrices of zeroes and ones and are defined to determine the covariance structure being imposed in the multivariate model. Therefore, this procedure still allows for the different covariance structures such as those discussed in Chapter 5, to be imposed. For example, the general linear lag-dependent structure can be imposed by defining the matrices Δ_{2j} to Δ_{12j} assuming the same form as each one of the A_q matrices in Chapter 5 plus an additional Δ_{1j} matrix, with zeroes in every entry, necessary for the estimation of σ_v^2. It is worth remembering that, for the models considered here, $q = 1$.

Example 6.2 *continued* Consider for example that the Toeplitz structure is being imposed on the error covariance matrix for the balanced data set where

$T = 4$ and $q = 1$. Here, $\Sigma_v = \sigma_v^2$ and

$$\Sigma_u = \begin{pmatrix} \sigma^2 & & & \\ \sigma_1 & \sigma^2 & & \\ \sigma_2 & \sigma_1 & \sigma^2 & \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}.$$

Therefore, $\boldsymbol{\theta} = (\theta_1, ..., \theta_5)^T = (\sigma_v^2, \sigma^2, \sigma_1, \sigma_2, \sigma_3)^T$ and $s = 5$. The total variance is given as:

$$V_j = \sum_k^s \theta_k G_{kj} = \sum_k^s \theta_k (Z_j H_{kj} Z_j^T + D_j(I_{n_j} \otimes \Delta_{kj}))$$

where the matrices H_{kj} are

$$H_{1j} = [1]\ ,\ H_{2j} = [0]\ ,\ H_{3j} = [0]\ ,\ H_{4j} = [0]\ ,\ \text{and}\ H_{5j} = [0],$$

and the matrices Δ_{kj} are

$$\Delta_{1j} = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 \end{pmatrix},\ \Delta_{2j} = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix}\ \Delta_{3j} = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & 0 \end{pmatrix},$$
$$\Delta_{4j} = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ 1 & 0 & 0 & \\ 0 & 1 & 0 & 0 \end{pmatrix}\ \text{and}\ \Delta_{5j} = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ 0 & 0 & 0 & \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

■

The PWIGLS method iterates between the estimation of $\hat{\boldsymbol{\beta}}_{PWIGLS}^{(r)}$ and $\hat{\boldsymbol{\theta}}_{PWIGLS}^{(r)}$ as defined in equations 6.12 and 6.15 respectively. These equations are repeated here (note that the superscript (r) defines the current iteration and $(r-1)$ the previous iteration):

$$\hat{\boldsymbol{\beta}}_{PWIGLS}^{(r)} = \hat{P}^{(r)-1}\hat{Q}^{(r)},$$

$$\hat{\boldsymbol{\theta}}_{PWIGLS}^{(r)} = \hat{R}^{(r)-1}\hat{S}^{(r)},$$

where

$$\hat{P}^{(r)} = \sum_j w_j(X_j^T \hat{V}_j^{(r-1)-1} X_j) \ ,$$
$$\hat{Q}^{(r)} = \sum_j w_j(X_j^T \hat{V}_j^{(r-1)-1} \boldsymbol{Y}_j) \ ,$$

the k^{th} row and l^{th} column element of $\hat{R}^{(r)}$ and the k^{th} entry of $\hat{S}^{(r)}$ are given by

$$\hat{R}^{(r)}[k,l] = \sum_j w_j tr(\hat{V}_j^{(r-1)-1} G_{kj} \hat{V}_j^{(r-1)-1} G_{lj}) \text{ and}$$
$$\hat{S}^{(r)}[k] = \sum_j w_j tr(\hat{\boldsymbol{e}}_j^T \hat{V}_j^{(r-1)-1} G_{kj} \hat{V}_j^{(r-1)-1} \hat{\boldsymbol{e}}_j).$$

Here $\hat{\boldsymbol{e}}_j = \boldsymbol{Y}_j - X_j \hat{\boldsymbol{\beta}}_{PWIGLS}^{(r)}$ and $w_j = 1/\pi_j$ as defined in Pfeffermann et al. (1998). Note that the matrix $\hat{R}^{(r)}$ is a $s \times s$ matrix and $\hat{S}^{(r)}$ is $s \times 1$ matrix. For the multivariate multilevel model, following suggestions by Goldstein (1986) and Searle et al. (1992), $\hat{V}_j^{(r-1)-1}$ can be written as:

$$\hat{V}_j^{-1} = D_j^{-1}(I_{n_j} \otimes \hat{\Sigma}_u^{-1}) - D_j^{-1}(I_{n_j} \otimes \hat{\Sigma}_u^{-1}) Z_j A_j Z_j^T D_j^{-1}(I_{n_j} \otimes \hat{\Sigma}_u^{-1})$$

where

$$A_j = \left((\hat{\sigma}_v^2)^{-1} + Z_j^T D_j^{-1}(I_{n_j} \otimes \hat{\Sigma}_u^{-1}) Z_j\right)^{-1} \ ,$$

and D_j^{-1} is defined as

$$D_j^{-1} = \text{diag}(w_{i|j}) \otimes D_{ij} \ .$$

Here diag$(w_{i|j})$ is a diagonal matrix with the weights for the heads of household repeated in the main diagonal. Note that for simplicity of notation in the above formulae, the superscript (r) for the current iteration was omitted from matrix A_j and the superscript $(r-1)$ for the previous iteration was omitted from $\hat{\sigma}_v^2$ and $\hat{\Sigma}_u$.

The final iteration provides the estimates for $\hat{\boldsymbol{\beta}}_{PWIGLS}$ and $\hat{\boldsymbol{\theta}}_{PWIGLS}$. The variance estimators for the PWIGLS estimates are given in equations 6.16 and 6.17, and repeated here:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_{PWIGLS}) = \hat{P}^{-1} \left(\frac{n}{n-1}\right) \left(\sum_j w_j^2 \boldsymbol{c}_j \boldsymbol{c}_j^T\right) \hat{P}^{-1},$$

where $\boldsymbol{c}_j = (X_j \hat{V}_j^{-1} \hat{\boldsymbol{e}}_j)$ and

$$\widehat{Var}(\hat{\boldsymbol{\theta}}_{PWIGLS}) = \hat{R}^{-1} \left(\frac{n}{n-1} \right) \left(\sum_j w_j^2 (\hat{S} - \hat{R}\hat{\boldsymbol{\theta}})(\hat{S} - \hat{R}\hat{\boldsymbol{\theta}})^T \right) \hat{R}^{-1}.$$

Note that the subscript $PWIGLS$ was omitted from $\hat{\theta}_{PWIGLS}$ in the above equation for simplicity.

6.4 Summary

This chapter reviewed further topics to be considered in the analysis of a complex longitudinal data set. It presented a review of the methods for the analysis of longitudinal multilevel data under informative sampling designs. The probability-weighted iterative generalised least squares as presented in Pfeffermann et al. (1998) was described in detail. The topics reviewed in this chapter provided the fundamental theory necessary for the extension of the PWIGLS estimation method for multivariate multilevel models. This method was described in this chapter. The extended method can be applied to both longitudinal data sets and cross sectional data sets. The procedure accommodates different error covariance structures as long as they can be expressed as linear functions of the covariance parameters. Therefore, auto-regressive structures cannot be fitted here. Computer routines for the implementation of such a method were also developed.

Chapter 7

Longitudinal Multilevel Modelling Compensating for Panel Non-response

7.1 Introduction

In the analyses presented so far only complete-case data were considered. This involved the selection of employed heads of household who were employed at all times while in the Brazilian labour force survey (the PME) starting from January 2004. Furthermore, this subset was reduced to contain only heads of household classified as completers, meaning that they had valid data for all the occasions they were to be observed in the survey. Therefore, heads of household who failed to present data for any of the occasions or those who dropped out from the survey were not considered. The choice for the complete-case data was adopted for simplicity as the analysis of longitudinal data under the multilevel framework accommodates the analysis of unbalanced data sets.

The PME survey has quite a complex sampling design. Recall that this design involves a multistage rotating sample scheme with units selected through unequal probabilities. Previous chapters have dealt with some of the complexities of the PME sampling design. Chapter 4 dealt with the hierarchical structure of the PME data while Chapter 5 dealt with the rotating sampling scheme and with the complex correlation structure of the data. The main objective of this chapter is to implement the weighting of the observations in the PME survey mainly to account for panel non-response while still accounting for the aforementioned complexities.

As noted in Chapter 6, the use of the sampling weights for inference about model parameters is still under debate. Therefore, this chapter aims to provide a contribution to such discussion by presenting the analysis of the PME data through the fitting of multilevel longitudinal models utilising the PWIGLS estimation method described in Chapter 6. This chapter also aims to present comparisons between multilevel models fitted not taking into account the sampling weights but instead using methods for the robust estimation of the standard errors of the regression parameters. It is worth mentioning that, although containing a longitudinal component, the PME survey does not provide the set of longitudinal sampling weights. Hence, this chapter also presents some discussion and a suggested method to calculate such weights for the data analysed.

7.2 Weight Adjustments for Panel Non-response

As mentioned in Section 2.4, an alternative approach to compensate for the loss of data in a longitudinal survey is to make use of longitudinal sampling weights. This type of sampling weight is often supplied with longitudinal survey data. However, it is not available with the PME data. As described in Chapter 3, only the cross-sectional sampling weights are available to use. These weights are of two types: the sample base weight, which corrects for the cross-sectional non-response and the post-stratification, or calibrated, weights.

This section proposes an alternative method to calculate the longitudinal weights for the PME data under analysis. However, before implementing the proposed method, the next subsection presents a brief description of the sampling weights already provided in the PME survey. This description was also presented in Section 3.2.2. However, the following subsection discusses how to transform these weights into weights to be used in multilevel analysis.

7.2.1 Cross-sectional and Multilevel Weights for the PME

Chapter 3 described the PME design as being a stratified two stage cluster design in each of the six metropolitan regions of the survey. Within each of the metropolitan regions, municipalities compose the independent strata (subscript h) from which the census sectors, the PSUs, are selected. The PSUs are selected through PPS proportional to their total number of households as listed in the 2000 Census. Within each PSU the households are selected (the SSUs) via simple systematic sample. This is initially a self-weighting design. However, corrections for cross-sectional unit non-response are applied to the initial sampling weights resulting in the loss of the self-weighted characteristic.

Consider, here, the following sampling design variables:

H the number of strata;

n_h the number of PSUs selected in the h^{th} stratum;

m_{jh} the number of households selected in the j^{th} PSU;

M_{jh} the number of households counted in the sampling frame;

p_{jh} the relative size of PSU j in stratum h in the 2000 Census;

m^*_{jh} the number of households that were interviewed;

n^*_h the number of PSUs with at least 1 interviewed household.

The cross-sectional PME sampling weights corrected for unit non-response are of the form:

$$w^*_{ijh} = \frac{1}{n^*_h p_{jh} \frac{m^*_{jh}}{M_{jh}}} \ . \tag{7.1}$$

In the multilevel modelling framework, this set of weights makes up the unconditional level one weights. To be able to account for the sampling weights in a longitudinal multilevel analysis, it is necessary that a set of weights for each level of the data hierarchy is available. In the PME data set these sets of weights are not provided but they can be calculated from the unconditional weights using the sampling design variables provided. Based upon the PME design variables, the PSU level weights can be calculated as

$$w^*_{jh} = \frac{1}{n^*_h p_{jh}} \ , \tag{7.2}$$

and the household (individual) level weights can be calculated as

$$w^*_{i|jh} = \frac{M_{jh}}{m^*_{jh}} \ . \tag{7.3}$$

Due to the design, every household and individual within the same PSU has the same set of weights.

An investigation of the weights provided with the PME data files indicated a potential problem. The selection probabilities p_{jh} for some of the PSUs were found to change across the months of the survey. This change was observed to occur mainly in those panels included in the sample in 2005. It is understood from the design that this might be an indication of some field listing exercise or change in the geographic boundaries of the PSUs. However, no confirmation was found in the official documents of this survey. In order to address this problem, a new PSU identification variable was created[1].

Notice that the multilevel weights are calculated as recommended in Rabe-Hesketh and Skrondal (2006) using the sample base weights and not the post-stratification weights as these weights reflect more than just the sampling design. Hereafter the subscript h is omitted for simplicity.

7.2.2 Longitudinal Sampling Weights for the PME Survey

A brief review of some methods utilized for the construction of longitudinal weights was presented in Chapter 2. Based upon those methods, this section describes how the sets of longitudinal weights were calculated for the PME data under study.

One important decision to make when constructing longitudinal weights is which patterns of non-response to account for. As mentioned in Chapter 2, for a longitudinal survey with T occasions there would be up to 2^T patterns of non-response. For the case of the PME this sums to a total of 256 potential patterns. This could generate up to $2^T - 1$ sets of longitudinal weights, 255 for the PME, if all the patterns were to be accounted for. For this reason, it is quite common to construct the set of longitudinal weights which accounts only for the drop-out patterns and not for the intermittent non-response. However, this implies the exclusion of data since only individuals who respond up to occasion t, before dropping out of the panel, are retained. This set of individuals for each occasion t before the drop-out is called the attrition or drop-out sample (Skinner and Holmes, 2003).

Chapter 3 presented a brief analysis of the non-response patterns for the household units in the PME data set (see Section 3.3). From Table 3.7, which is also found summarized in Table 7.1 under the column of "All HoHH" (HoHH means heads of household), it was observed that, considering the total set of households which by design should have data in all eight occasions of the survey, 66% were completers, 7% dropped out at some point during the survey and 27%

[1] Models fitted later in this chapter use the new cluster identification variable for identifying the higher level units. Therefore, different results are found when comparing the models fitted in this chapter with those fitted in Chapter 5.

had partial non-response. It was also observed that the drop-out from the fifth interview was the most frequent. In addition, at this interview time a higher number of households which had partial non-response was observed.

Table 7.1: Non-response Patterns

Interview	Data from	
Time	All HoHH	Selected HoHH
Wave 8		
Completers	26,274	10,183
Wave NR	473	182
Wave 7		
Wave NR	233	85
Drop-out	228	99
Wave 6		
Wave NR	367	144
Drop-out	192	73
Wave 5		
Wave NR	1,598	675
Drop-out	1,615	820
Wave 4		
Wave NR	610	265
Drop-out	137	73
Wave 3		
Wave NR	722	281
Drop-out	127	71
Wave 2		
Wave NR	909	403
Drop-out	161	97
Wave 1		
Wave NR	5,905	-
Total	39,551	13,451

Table 7.1 also presents the non-response patterns for the data set under study in this chapter (explained in the following sub-section). It shows that 10,183 (76%) of the 13,451 selected heads of household provided data for all the interviews (eight in total). As before, there is a higher frequency of drop-out after the fourth interview, as well as a larger number of heads of household with partial non-response at the fifth interview. From the total of 13,451, 1,415 (11%) dropped out of the panel at some point and 1,853 (14%) had partial non-response. This shows that adjusting only for the drop-out patterns eliminates 14% of data. It is worth mentioning that, by construction in the selected data, there is no non-response at the first wave.

There are different methods to calculate the longitudinal weights. These methods mainly follow three steps which involve the definition and calculation of the base weights, the adjustment of the base weights by a non-response adjustment and later the calibration of the adjusted weights to population totals (Rizzo et al., 1996). It is also important to have a good definition of the respondents and non-respondents. This depends on the patterns of panel non-response that the weights are compensating for. After the definition of respondents and non-respondents the adjustment for the base weights can be calculated, for example, by using adjustment cells methods or logistic regression methods for the propensity of response. The latter is the method adopted in this chapter and the adjustment for the base weights is calculated as the inverse of the predicted probability of responding. This adjustment is applied only to the weights of the respondents in order to account for the non-respondents. The following sub-sections present each of the steps followed.

7.2.2.1 Base Weights and Definition of Respondents

The first step in the construction of any set of longitudinal weights is to define the set of base weights. In this analysis the base weights are defined as being the sample base weight at the first occasion for the units under analysis. The base weights are here called $w^*_{ij(1)}$. It is necessary at this point, however, to make a discussion on the data set selected to be analysed. Following the analysis presented in Chapter 5 this chapter shall also consider the longitudinal working data set, as described in Section 3.4. However, for the construction of the longitudinal weights, the data set used is the one before the filter for completers. It is important to mention that the validation of the matching of all eight occasions is not performed at this point either as this can only be done for the completers set.

The next step is to define the sets of respondents and non-respondents. This depends on which patterns of non-response the adjustment is accounting for. It can account for every non-response pattern or for only the attrition patterns. Kalton and Bryk (2000) and Lepkowski (1989) suggested transforming the non-attrition patterns into attrition patterns. This can be performed by considering the individual data until their first wave non-response. However, this was not undertaken here as it ignores the possibility that individuals who drop-out from the PME panel might be different from those who miss one interview but reappear later in the panel. Furthermore, given the complexity of accounting for each of the PME non-response patterns the non-respondents are defined here as those who drop-out from the survey. Therefore, no intermittent non-response is considered. As a result, only the drop-out samples, which are composed by those who respond up to occasion t, are retained for each occasion data. In this way, the response indicator is equal to one for those who do not drop-out from the panel and zero otherwise. The selection of those who respond up to the eighth occasion defines the set of completers, who have their weights adjusted to represent the sample observed at the first occasion.

7.2.2.2 Logistic Regression Model for the Predicted Probabilities of Response

The method chosen to adjust the base weights ($w^*_{ij(1)}$) to compensate for the drop-out is one of the methods mentioned in Lepkowski (1989). It involves performing logistic regressions of a response indicator in order to calculate the predicted probabilities and adjust the base weights by the inverse of these predicted probabilities. Because no intermittent non-response is considered, the probability of dropping out from the panel at occasion t is what is being predicted.

For each head of household on each of the occasions, there is an indicator of responding up to that specific wave (receiving value 1) against dropping out at that specific wave (receiving value 0). The set of completers are those who have response indicators equal to one in all the eight occasions. Logistic regression models are fitted for each response indicator as the outcome variable starting from the second occasion as all the units are observed at the first occasion. The covariates considered in each of the models are taken from the previous occasion. In this sense, using the data from the first occasion, a logistic model for the response at the second occasion is performed in order to adjust the base weights and calculate occasion 2 weights. The same procedure is performed using data of occasion 2 and response indicator for occasion 3 to adjust the already adjusted weights of occasion 2. This is repeated for each subsequent pair of occasions in order to conditionally adjust the weights of each occasion given the adjustment of the previous occasion.

The choice of the covariates for initial inclusion in each of the logistic models was made with the objective of predicting the probability of response given the characteristics of the heads of the households. The auxiliary variables initially considered were the same set of variables as those in Chapter 4. The exception is that all the variables were considered in the model as categorical variables. Therefore, continuous covariates were categorized. This followed the suggestions presented in Chapter 2.

Initial model selection showed that some of the variables were not statistically significant. Therefore they were not included subsequently. It was also observed that for some of the variables their categories could have been collapsed and this was pursued for both the education and the age variables. Further model selection was performed for each of the seven models, one for each consecutive pair of waves, separately. Statistically significant main effects were initially tested in one-level logistic regression models through forward selection. Using these models as a base, multivariate Wald tests for each categorical variable were performed and statistically significant terms at the 5% level were retained in the model. Table 7.2 presents the final main-effects models. This table shows that for each pair of occasions a different final main effects model was selected. No suggestion was found in the reviewed literature on the best approach to select these models. The belief that the probabilities of response might change over time motivated the selection of different models for each of the pairs of occasions.

Note from Table 7.2 that the variable for metropolitan region was kept in all the models even when the level of statistical significance was not met as this is an important variable in the design of the survey. Another design variable considered and tested was the variable representing the panel the units were selected to. This is the first set of variables shown in Table 7.2. The panel variable was only significant in the model (3,4) but not for the others. Although not shown in Table 7.2 the outcome variable income was also tested in the models. However, in none of the final main effects models did the variable income meet the significance criteria. This shows that the response probabilities are not related to the income for the set of heads of household considered here, which is an indication that the panel non-response is missing at random.

After selecting the one-level models in Table 7.2, two-level random intercept models were evaluated. There are different ways to perform the estimation of discrete random intercept models. The models presented in this section were all estimated using the `Stata` software and the `Gllamm` command, which estimates the random intercept models via adaptive quadrature methods as mentioned in Chapter 2. The two-level logistic models define the PSUs (clusters) as the second level units and include a random intercept for each of the clusters. However, after the selection of the data under analysis the number of observations per cluster is quite small and some of the clusters have no variation on the response indicator. Interaction effects between the significant main effects were then investigated at this stage for both one-level and two-level models. The two-level models which included the interaction terms did not converge and for some of the one-level models convergence problems were also met. That was an indication for the non-inclusion of the interaction effects. A similar problem with interactions terms was mentioned in Rizzo et al. (1996) where the final model for the predicted probabilities included only the significant main effects.

Table 7.2: Logistic Regression Model for the Response Propensity

	Occasion data, Response indicators						
	1,2	2,3	3,4	4,5	5,6	6,7	7,8
Panel 4.02			1.086				
			(2.05)*				
Panel 4.03			1.078				
			(2.03)*				
Panel 4.04			0.93				
			(1.86)				
Panel 4.05			0.589				
			(1.29)				
Panel 4.06			2.153				
			(2.82)**				
Panel 4.07			0.569				
			(1.33)				
Panel 4.08			0.763				
			(1.74)				
Panel 4.09			0.157				
			(0.41)				
Age 40 and over	0.565		0.649	0.431		0.468	0.433
	(2.69)**		(2.68)**	(5.82)**		(2.29)*	(2.85)**
+12 years of Education		-0.789	-0.862	-0.396		-0.539	-0.409
		(2.92)**	(3.27)**	(4.35)**		(2.28)*	(2.22)*
5 to 9 years of work		-0.063			0.003		
		(0.23)			(0.01)		
10 to 14 years of work		0.549			0.961		
		(1.33)			(2.02)*		
15 to 19 years of work		1.344			2.039		
		(1.85)			(2.01)*		
20 to 24 years of work		1.187			1.115		
		(1.63)			(1.53)		
+25 years of work		-			-0.183		
		-			(0.44)		
Proxy Respondent	0.643	0.64		0.162			
	(2.79)**	(2.57)*		(2.05)*			
2 members in the HH	0.602		0.921	0.336		0.545	0.509
	(1.99)*		(2.38)*	(2.68)**		(1.73)	(1.87)
3 members in the HH	0.995		1.101	0.543		0.767	0.602
	(3.24)**		(2.99)**	(4.39)**		(2.53)*	(2.35)*
4 members in the HH	1.124		1.201	0.895		1.496	0.641
	(3.47)**		(3.20)**	(6.84)**		(4.18)**	(2.51)*
5 members in the HH	1.524		0.425	0.705		0.941	0.389
	(3.24)**		(1.12)	(4.67)**		(2.43)*	(1.35)
+6 members in the HH	1.362		1.816	0.856		1.407	1.064
	(2.46)*		(2.39)*	(4.59)**		(2.52)*	(2.57)*

Table 7.2 – continued from previous page

	Occasion data, Response indicators						
	1,2	2,3	3,4	4,5	5,6	6,7	7,8
Salvador	1.528	-0.005	1.214	-0.206	-0.606	0.603	0.637
	(2.84)**	-0.01	(2.40)*	-1.37	-0.89	-1.37	-1.69
Belo Horizonte	-0.126	-0.538	0.269	0.136	-0.999	0.018	-0.424
	(0.36)	(1.07)	(0.72)	(0.91)	(1.59)	(0.05)	(1.46)
Rio de Janeiro	1.946	0.605	1.855	0.483	0.381	1.074	0.882
	(3.64)**	(1.06)	(3.46)**	(3.23)**	(0.54)	(2.55)*	(2.60)**
São Paulo	0.469	0.294	1.379	0.285	-0.881	0.625	0.497
	(1.33)	(0.56)	(3.25)**	(2.00)*	(1.43)	(1.67)	(1.62)
Porto Alegre	0.133	-0.476	-0.048	0.16	-1.211	-0.025	-0.61
	(0.35)	(0.91)	(0.13)	(1.01)	(1.92)	(0.07)	(2.07)*
Constant	2.876	4.744	2.669	1.567	5.435	3.273	3.219
	(8.00)**	(9.83)**	(5.66)**	(9.90)**	(9.21)**	(8.27)**	(9.64)**
Observations	11,598	10,808	11,430	11,357	10,537	10,464	10,365

Note: Absolute value of z statistics in parentheses.
*significant at 5%, **significant at 1%

With the decision to test main-effects models only, Table 7.3 presents the summary of predicted probabilities under the different model formulations. Table 7.4 presents the estimates for the between cluster variance in the two-level random intercept logistic models. It also presents the test for goodness-of-fit comparing the two model formulations. Note from Table 7.3 that the predicted probabilities for each of the models are not very different when comparing one-level with two-level models. A small difference is observed when, for the two-level models, the random effects are taken into account in calculating the probabilities. However, Table 7.4 shows that for three of the seven models the two-level model is not significantly different to the one-level model. This raises the issue of whether or not to account for the random intercepts in any of the models. The aim of this analysis is to calculate the marginal predicted probabilities of panel drop-out given the heads of household characteristics rather then providing inference on individual effects. Hence, the choice would be for the one-level model. Random effects models would be of interest if the probabilities of panel drop-out were thought to vary between clusters and also to control for the effects of the data hierarchy on these probabilities. By retaining the metropolitan region variable in the models, the regional effects are being controlled for. In addition, due to the similarities between the predicted probabilities, in order to maintain simplicity and consistency between the different models, those probabilities for the one-level model formulation are chosen at this stage to be the weight adjustments.

Table 7.3: Summary of the Predicted Probabilities

Model		Statistics			
		Min	Mean	Median	Max
1,2	Two-level				
	Fixed Only	0.9689	0.9963	0.9975	0.9998
	Fixed and Random	0.7052	0.9958	0.9977	0.9998
	One-level	0.9399	0.9916	0.9940	0.9995
2,3	Two-level				
	Fixed Only	0.9732	0.9949	0.9962	0.9995
	Fixed and Random	0.9625	0.9949	0.9961	0.9995
	One-level	0.9680	0.9938	0.9954	0.9993
3,4	Two-level				
	Fixed Only	0.8530	0.9936	0.9966	0.9999
	Fixed and Random	0.8530	0.9936	0.9966	0.9999
	One-level	0.8530	0.9936	0.9966	0.9999
4,5	Two-level				
	Fixed Only	0.7523	0.9393	0.9480	0.9770
	Fixed and Random	0.5272	0.9373	0.9483	0.9841
	One-level	0.7241	0.9278	0.9370	0.9717
5,6	Two-level				
	Fixed Only	0.9885	0.9954	0.9966	0.9998
	Fixed and Random	0.9641	0.9953	0.9965	0.9998
	One-level	0.9827	0.9931	0.9948	0.9996
6,7	Two-level				
	Fixed Only	0.9727	0.9969	0.9977	0.9995
	Fixed and Random	0.7007	0.9959	0.9979	0.9996
	One-level	0.9375	0.9905	0.9923	0.9982
7,8	Two-level				
	Fixed Only	0.9318	0.9891	0.9919	0.9980
	Fixed and Random	0.6812	0.9882	0.9924	0.9982
	One-level	0.9003	0.9824	0.9869	0.9963

Table 7.4: Between Cluster Variance and Goodness-of-fit Test

Model				Goodness-of-fit		
		$\hat{\sigma}_u^2$	SE($\hat{\sigma}_u^2$)	-2× *Log-Likelihood*	LRT	Half p-value
1,2	Two-level	1.879	0.753	1029.04	8.51	0.002
	One-level			1037.55		
2,3	Two-level	0.385	0.769	827.31	0.24	0.311
	One-level			827.55		
3,4	Two-level	0.000	0.000	796.69	0.00	0.500
	One-level			796.69		
4,5	Two-level	0.490	0.109	5675.18	32.26	0.000
	One-level			5707.44		
5,6	Two-level	0.875	0.844	836.57	1.06	0.151
	One-level			837.63		
6,7	Two-level	2.526	0.706	1045.26	25.29	0.000
	One-level			1070.56		
7,8	Two-level	1.077	0.341	1737.48	14.89	0.000
	One-level			1752.38		

7.2.2.3 Adjusted Longitudinal Weights

Once the predicted probabilities have been calculated, the next step is to adjust the weights by the inverse of the predicted probabilities. These adjustments are performed only on the set of completers. The first step is to adjust the base weights, $w^*_{ij(1)}$, multiplying them by the inverse of the predicted probabilities of the model "1,2", $p_{ij(1,2)}$, which uses data from the first occasion to predict the response at the second occasion:

$$w^*_{ij(2)} = w^*_{ij(1)} \times 1/p_{ij(1,2)} \ ,$$

$w^*_{ij(2)}$ are the adjusted weights for occasion 2 that compensates for the panel drop-out. The adjusted weights for occasion 3 are then calculated as:

$$w^*_{ij(3)} = w^*_{ij(2)} \times 1/p_{ij(2,3)} \ ,$$

for occasion 4

$$w^*_{ij(4)} = w^*_{ij(3)} \times 1/p_{ij(3,4)} \ ,$$

and so on, giving the adjusted weights for occasion 8:

$$w^*_{ij(8)} = w^*_{ij(7)} \times 1/p_{ij(7,8)} \ .$$

Adopting this procedure, Table 7.5 presents the summary statistics for the adjusted longitudinal weights for each of the occasions. The first row of Table 7.5 considers the base weights for completers only. These are the adjusted unconditional longitudinal weights for the PME data. Observe that these weights tend to increase over time. Table 7.6 presents the same arrangement of information as in Table 7.5 but for each of the adjustments, i.e. for each of the inverse probabilities in each of the occasions, for all the data used.

Table 7.5: Summary of the Adjusted (Unconditional) Weights - Completers

	Statistics					
	Min	Mean	Median	Max	SD	RSE
$w^*_{ij(1)}$	131.00	478.18	468.40	2500.00	270.54	56.58
$w^*_{ij(2)}$	131.42	481.54	470.76	2506.02	272.09	56.50
$w^*_{ij(3)}$	132.05	483.99	474.04	2515.98	272.94	56.39
$w^*_{ij(4)}$	133.56	486.24	476.65	2518.21	273.33	56.21
$w^*_{ij(5)}$	138.93	521.46	501.27	2691.98	291.71	55.94
$w^*_{ij(6)}$	140.96	524.89	502.04	2700.00	293.62	55.94
$w^*_{ij(7)}$	141.73	529.34	505.79	2721.89	295.64	55.85
$w^*_{ij(8)}$	145.29	537.20	511.51	2751.12	298.02	55.48

Table 7.6: Summary of the Inverse of the Predicted Probabilities

	Statistics					
	Min	Mean	Median	Max	SD	RSE
$1/p_{ij(1,2)}$	1.0005	1.0085	1.0060	1.0639	0.0090	0.8924
$1/p_{ij(2,3)}$	1.0007	1.0062	1.0046	1.0331	0.0050	0.4969
$1/p_{ij(3,4)}$	1.0001	1.0065	1.0034	1.1723	0.0090	0.8942
$1/p_{ij(4,5)}$	1.0291	1.0794	1.0672	1.3809	0.0430	3.9837
$1/p_{ij(5,6)}$	1.0004	1.0070	1.0052	1.0176	0.0050	0.4965
$1/p_{ij(6,7)}$	1.0018	1.0096	1.0078	1.0666	0.0080	0.7924
$1/p_{ij(7,8)}$	1.0037	1.0180	1.0133	1.1108	0.0130	1.2770

In Table 7.6, the predicted probabilities are not too variable. The exception is for $1/p_{ij(4,5)}$. This represents the probability of responding after the gap of eight months between the fourth and fifth occasions. The inverse of this probability shows a sudden increase and is the most variable one according to the relative standard error[2] (RSE). This is also reflected in the adjusted weights in Table 7.5. The variance of the longitudinal weights $w^*_{ij(8)}$ is larger than the initial base weights. However, this increase in the variance of the weights does not seem to be too serious.

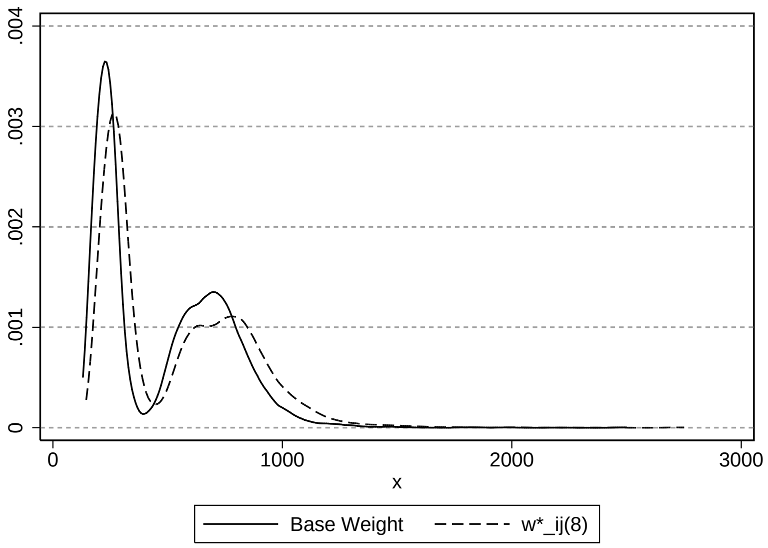
The plot in Figure 7.1 shows that the distribution of the adjusted weights for the 8^{th} occasion seems to be slightly shifted to the right in comparison to the distribution of the base weights. Note that the shape of the distribution is determined by the different sampling fractions for each of the metropolitan regions. The set of longitudinal weights $w^*_{ij(8)}$ is the one to be used in the analysis of the data set including data from the first to the eighth occasions for completers only. This set of weights compensates for the losses of data between each pair of previous occasions in order to keep the representativeness as at the first occasion.

The natural next step in the production of longitudinal weights would be to use some sort of calibration technique on the adjusted weights. However, the main objective for constructing such a set of weights here, is their use in a multilevel analysis. Furthermore, Rabe-Hesketh and Skrondal (2006) stated that calibration weights do not depend on selected clusters. For this reason, this step will not be performed here. In order to use the longitudinal weights in a multilevel model analysis, the different weights for the different levels need to be calculated in a similar way as in equations 7.2 and 7.3. Hence, the set of adjusted weights $w^*_{ij(8)}$ needs to be partitioned accordingly.

The cluster level weights w^*_j are still defined as in 7.2 as they represent the inverse of the PSU selection probability which should not change. The heads of

[2] RSE=(SD×100)/Mean.

Figure 7.1: Distribution of the Base Weights and Adjusted Weights at the 8th Occasion



household level weights $w^*_{i|j}$ are those that need to be adjusted, and this can be performed as in

$$w^*_{i|j(8)} = \frac{w^*_{ij(8)}}{w^*_j} \; . \qquad (7.4)$$

In this way, $w^*_{i|j(8)}$ represents the longitudinal weights at the heads of household level adjusted for the panel drop-out.

Once this adjustment is performed the occasion level weights $w^*_{t|ij}$ can be defined as equal to one for every occasion. One alternative way to proceed is not to perform this adjustment and instead of considering the occasion level weights equal to one, consider

$$w^*_{t|ij} = \frac{1}{p_{ij((t-1),t)}} \; ,$$

which is the actual adjustment, the inverse of the predicted probabilities. This alternative procedure will be considered in one application in Section 7.3, for comparison.

The method developed in this section to construct the longitudinal weights for the PME data considering attrition patterns only, is one of the possible methods and perhaps the simplest. The construction and evaluation of longitudinal weights using different methods goes beyond the scope of this thesis. However, this is an area for future research.

7.2.3 Computational Aspects

Appendix D presents the computer codes developed for the estimation of the longitudinal multilevel models using the PWIGLS method. The codes were written using the `Mata` language for `Stata` for the estimation of the two-level random intercept or the random slope multilevel models. In addition, codes were also written for the estimation of multivariate multilevel models imposing the Toeplitz and the General Lag-dependent structures to the error covariance matrix.

These routines were developed following the method described by Pfeffermann et al. (1998) and were based on the `SAS` codes using `IML` language presented in Corrêa (2001) created for the estimation of random intercept models only. Therefore, the computer codes developed in this thesis represent extensions of such work. It is worth mentioning that the codes were written to implement the PWIGLS estimation method using the so called *scaling method 2* as described by Pfeffermann et al. (1998). This scaling method was reviewed in Chapter 6.

Note that some of the models fitted in the following section utilized robust methods for the estimation of the standard error for the regression coefficients. These models were fitted using `SAS PROC MIXED` adding the “empirical” option.

7.3 Results

To fulfil the methodological motivation of this thesis, that of dealing with the data complexities of the PME data in one analysis, this chapter uses the same working longitudinal data set as the one used in the analyses of Chapter 5. Therefore, the same set of $6,524$ heads of household are hereafter analysed. The main aim of the Chapter 5 was to account for the rotating design by imposing alternative error covariance structures on multivariate multilevel models. The same goal is pursued here. For that, the general linear lag-dependent covariance structure as presented in 5.8 is considered.

The final model presented in Table 5.7 serves as the starting point for this section. However, for simplicity, the interaction terms between some of the covariates and the occasion variable are not considered at this point. They are re-included and tested in a final model selection procedure performed in the last subsection of this section. In the next two subsections, an application of the PWIGLS method to growth curve models and an application of the extended method to multivariate multilevel models are presented.

7.3.1 Application of PWIGLS to Growth Curve Models

The main aim of this subsection is to illustrate an application of PWIGLS to the analysis of the PME longitudinal data set. PWIGLS, as developed by Pfeffermann et al. (1998), allows the estimation of two-level random coefficients models for continuous outcomes. This method is described in detail in Section 6.2.1. It is recognized that the random coefficient model is not the most appropriate type of model to fit the PME data under study, as concluded in Chapter 5. However, this is performed in order provide some discussion on the application of the PWIGLS. This section also aims to provide a comparison of models estimated using the robust sandwich estimator for the standard errors of the regression parameters and models estimated via standard methods, neither accounting for the sampling weights. Note that the robust sandwich estimator protects against misspecification of the normality assumption of the higher level residuals, as discussed in Chapter 2. Rabe-Hesketh and Skrondal (2006) referred to these as being semi-robust standard errors.

As mentioned in the previous chapters, if no further clustering of higher level units are present in the data, growth curve models are two-level multilevel models where the variable representing time usually has a random coefficient. In this case level two units are the individuals i and the level one units are the occasions t. The PWIGLS, as described in Pfeffermann et al. (1998), can be readily applied. However, this is not the case for the PME data which includes the higher level for the PSU clusters j. Hence, this section starts by comparing two-level models, which do not account for the PSU level, and three-level models for the PME data. A similar analysis was presented in Skinner and Vieira (2007) while investigating the effects of clustering on the variance estimation of the regression coefficients in an analysis of longitudinal survey data. They compared robust estimation methods with standard estimation methods to investigate whether fitting multilevel models which do not account for any of the design features would be sufficient to account for the effects of clustering. They concluded that simply including the higher cluster level was not enough and that either the inclusion of random coefficients or the use of robust methods would be necessary. They also observed that the three-level model with robust methods provided results equivalent to those for the model estimated using methods which do account for the complex sampling design without accounting for the data hierarchy. However, none of their analyses included the sampling weights.

Table 7.7 presents the results for the two-level and three-level random coefficient models, still not considering the sampling weights, in order to compare the standard errors of the regression coefficient estimates (SEs). These models considered the time variable varying as $(0, 1, 2, 3, 12, 13, 14, 15)$. The last two columns of Table 7.7 present the ratios of the SEs for the three-level models to the two-level models. Values equal to one in these two last columns indicate that both SEs are equal and values greater than one indicate that those for the three-level model are larger.

The SEs for the two-level and three-level models (columns labelled SE) can be compared and they are virtually the same for most of the regression coefficients. Differences are most noticeable in the SEs for the contextual variables and for the metropolitan region variables, as they are cluster level covariates. Furthermore, difference is also observed for the SE of the intercept. The comparison between the robust standard errors and the non-robust shows that robust SEs (columns labelled Rob.SE) are slightly larger than non-robust. However, this is not observed for the variable of metropolitan regions where robust SEs are smaller than non-robust. Furthermore, some of the variables present a notable difference between robust and non-robust SEs. This will be investigated once again when the SEs under PWIGLS estimation are also examined.

Observing the values of the estimated regression coefficients in Table 7.7, notice that they are very similar but not identical when comparing the three-level with the two-level model. One other important point to observe, is the relatively small between cluster variability in the three-level model. The between individual variability is larger for this model, even larger than the within individual variability, reflecting that most of the variability is at level two. This very small cluster level variance, although statistically significant, is an indication that the fit of the two-level model for this data set would not be so bad. Table 7.8 therefore presents the results for the two-level models estimated via PWIGLS, hereafter referred to as PWIGLS models.

Table 7.7: Two and Three-level Random Slope Models

	2-level Model			3-level Model			*3-level/2-level*	
	Coeff	SE	Rob.SE	Coeff	SE	Rob.SE	SE	Rob.SE
Constant	6.198	0.230	0.270	6.227	0.249	0.283	1.084	1.051
Wave	0.003	0.000	0.000	0.003	0.000	0.000	1.000	1.056
Males	0.663	0.075	0.132	0.657	0.075	0.136	0.998	1.034
White	0.226	0.023	0.036	0.226	0.023	0.036	1.003	0.996
Age (@ wave 1)	4.265†	1.399†	1.477†	4.235†	1.395†	1.471†	0.997	0.996
Squared term	-0.443†	0.043†	0.054†	-0.442†	0.043†	0.056†	0.997	1.036
Education (@ wave 1)	-0.069	0.011	0.012	-0.066	0.011	0.012	0.998	1.008
Squared term	0.009	0.001	0.001	0.009	0.001	0.001	0.998	1.021
Type of Worker								
(Employer as baseline)								
Informal	-0.139	0.016	0.028	-0.139	0.016	0.029	1.000	1.010
Formal	-0.044	0.016	0.029	-0.044	0.016	0.029	0.999	1.005
Military service	-0.009	0.021	0.034	-0.009	0.021	0.034	1.000	1.011
Self-Employed	-0.175	0.014	0.024	-0.175	0.014	0.024	1.000	1.011
Type of Activity								
(Manufacturing as baseline)								
Building	0.023	0.036	0.040	0.022	0.036	0.040	1.000	1.013
Commerce	-0.023	0.016	0.027	-0.024	0.016	0.027	1.000	1.003
Financial	0.036	0.020	0.027	0.035	0.020	0.028	1.000	1.016
Social Services	0.065	0.020	0.027	0.065	0.020	0.027	0.999	0.999
Domestic Services	-0.002	0.019	0.032	-0.003	0.019	0.032	1.000	1.016
Other Services	-0.002	0.018	0.026	-0.003	0.018	0.027	0.999	1.043
Other Activities	0.038	0.061	0.063	0.037	0.061	0.063	1.000	1.005
Duration of Employment								
(× 120)	0.036	0.008	0.013	0.036	0.008	0.013	1.000	1.018
Squared term	-0.020	0.004	0.007	-0.020	0.004	0.007	1.000	1.000
Working Hours (in Log)	0.254	0.011	0.027	0.254	0.011	0.027	1.000	1.023
Proxy Respondent	0.003	0.006	0.008	0.003	0.006	0.008	1.000	1.023
Number of HH members	0.005	0.003	0.004	0.005	0.003	0.004	0.999	1.042
Metropolitan Region								
(Recife as baseline)								
Salvador	0.039	0.031	0.030	0.039	0.033	0.033	1.056	1.104
Belo Horizonte	0.248	0.030	0.028	0.244	0.032	0.029	1.048	1.052
Rio de Janeiro	0.215	0.028	0.025	0.212	0.029	0.028	1.054	1.091
São Paulo	0.360	0.030	0.028	0.359	0.032	0.030	1.055	1.078
Porto Alegre	0.231	0.033	0.031	0.225	0.034	0.033	1.048	1.072
Interaction Terms								
of Male and:								
Age (@ wave 1)	0.002	0.002	0.002	0.003	0.002	0.002	0.997	0.999
Education (@ wave 1)	0.053	0.013	0.014	0.054	0.013	0.014	0.996	1.004
Squared term	-0.003	0.001	0.001	-0.003	0.001	0.001	0.997	1.022
Type of Activity								
(Manufacturing as baseline)								
Building	-0.036	0.037	0.042	-0.035	0.037	0.044	1.000	1.037
Commerce	-0.021	0.018	0.029	-0.020	0.018	0.029	1.000	1.014
Financial	-0.052	0.022	0.030	-0.051	0.022	0.031	1.000	1.034
Social Services	-0.039	0.023	0.034	-0.039	0.023	0.033	0.999	0.990
Domestic Services	-0.125	0.034	0.055	-0.124	0.034	0.056	1.000	1.017
Other Services	-0.036	0.020	0.029	-0.035	0.020	0.031	1.000	1.055
Other Activities	-0.087	0.065	0.074	-0.085	0.065	0.074	1.000	0.997
Duration of Employment								
(× 120)	0.020	0.009	0.015	0.020	0.009	0.015	1.000	0.992
Squared term	0.009	0.005	0.008	0.009	0.005	0.008	1.000	1.000
Working Hours (in Log)	-0.090	0.014	0.032	-0.089	0.014	0.033	1.000	1.031

Table 7.7 – continued from previous page

	2-level Model			3-level Model			*3-level/2-level*	
	Coeff	SE	Rob.SE	Coeff	SE	Rob.SE	SE	Rob.SE
Proxy Respondent	-0.042	0.007	0.009	-0.042	0.007	0.009	1.000	1.023
Interaction Terms								
of White and:								
Type of Worker								
(Employer as baseline)								
Informal	-0.058	0.020	0.036	-0.059	0.020	0.035	1.000	0.982
Formal	-0.103	0.020	0.037	-0.105	0.020	0.037	1.000	0.995
Military service	-0.142	0.026	0.044	-0.142	0.026	0.045	1.000	1.015
Self-Employed	-0.015	0.017	0.031	-0.015	0.017	0.031	1.000	0.998
Contextual Effects:								
Prop of Formal	-1.817	0.194	0.220	-1.843	0.211	0.234	1.090	1.066
Prop of Informal	-2.159	0.227	0.252	-2.185	0.248	0.267	1.093	1.062
Prop of Military	-2.017	0.212	0.246	-2.065	0.233	0.274	1.098	1.111
Prop of Self-Employed	-2.209	0.225	0.250	-2.273	0.245	0.273	1.089	1.093
Average Education	0.071	0.006	0.006	0.072	0.006	0.007	1.079	1.144
$\hat{\sigma}^2_v$				0.011	$2.305^{\dagger}$	$2.305^{\dagger}$		
$\hat{\sigma}^2_{u0}$	0.278	$5.158^{\dagger}$	$5.158^{\dagger}$	0.267	$5.350^{\dagger}$	$5.350^{\dagger}$	1.037	1.037
$\hat{\sigma}^2_{u1}$	-0.004	$0.187^{\dagger}$	$0.187^{\dagger}$	-0.004	$0.187^{\dagger}$	$0.187^{\dagger}$	1.000	1.000
$\hat{\sigma}_{u01}$	0.001	$0.012^{\dagger}$	$0.012^{\dagger}$	0.001	$0.012^{\dagger}$	$0.012^{\dagger}$	1.000	1.000
$\hat{\sigma}^2_e$	0.045	$0.323^{\dagger}$	$0.323^{\dagger}$	0.045	$0.323^{\dagger}$	$0.323^{\dagger}$	1.000	1.000
$-2\times$ *Log-Likelihood*		20,747			20,719			
AIC		20,859			20,833			
BIC		21,239			21,145			

Note: † Values at 10^{-3}.

The first two columns in Table 7.8 are the coefficient estimates and the robust estimates of the standard errors for the two-level model in Table 7.7, which are presented for easy comparison. Because this is a two-level random coefficient model, only two sets of weights are needed: one for the individual level and another for the occasion level. Furthermore, two different weighting approaches were undertaken. The first took as level two weights the unconditional longitudinal weights $w^*_{ij(8)}$, calculated as mentioned earlier, and the occasion level weights were all equal to one. The second is the alternative weighting approach as mentioned in section 7.2. This means that the level two weights were the unadjusted w^*_{ij} weights and the level one weights were equal to $w^*_{ij(8)}/w^*_{ij}$, representing the inverse of the probability of responding up to occasion 8.

These two different weighting approaches generated different results for the models fitted. In the first weighting approach the level one weights are equal to one; hence, no effect of scaling can be identified here. In the second weighting approach, because the predicted probabilities are constant within heads of household, the scaling method makes the level one weights all equal to one. In consequence, the second weighting approach is equivalent to not using any of the panel non-response adjustments, since the level two weights are unadjusted. In this sense, the comparisons in Table 7.8 are for a model fitted using robust methods, a model fitted using weights which compensate for panel non-response and a model fitted using unadjusted weights. This last model does not account for the panel attrition, which is not correct for this application.

The SEs are virtually the same when comparing the two PWIGLS models in Table 7.8. The difference is mostly on the third decimal place, with few exceptions. The SEs after weighting are generally larger than the robust SEs not accounting for the weights[3]. Comparing the values of the estimated regression coefficients between the two PWIGLS models, leads to the conclusion that, for most of the variables, they are roughly the same. There are a few cases where these estimates differ. Examples are for the type of worker variable and for some of the categories of type of activity. This indicates that the different weights have different effects on these covariates. The parameter estimates for the model fitted without the weights are similar to the estimated coefficients after weighting for most variables. However, once again, for type of worker and type of activities the coefficient estimates differ. This could be an indication that the weights are compensating for the disproportionate sampling for the different types of workers or that they are compensating for the different drop-out patterns which exist between these different groups.

One other important point to raise for the comparison of the estimates in Table 7.8, and also with those in Table 7.7, is that some of the variables lose significance after either robust methods or weighting are adopted. This is observed for some of the interaction terms and some main effects indicating that further model selection would be necessary. This will be performed in the next section. One last comparison between the PWIGLS models is that both the parameter estimates and the SEs of the estimated random effects variances are virtually the same. However, differences in the SEs for the within individual variance are noticed when comparing these variance component estimates with those for the model fitted without the weights.

[3] The SEs for the PWIGLS method are also robust SEs. However, when the term "robust" is used in this chapter it refers to models estimated without accounting for the multilevel weights.

Table 7.8: Two-level Random Slope Models Fitted by Alternative Methods

	IGLS		PWIGLS (1)		PWIGLS (2)	
	Coeff	Rob.SE	Coeff	SE	Coeff	SE
Constant	6.198	0.270	5.900	0.298	5.935	0.298
Wave	0.003	0.000	0.002	0.000	0.002	0.000
Males	0.663	0.132	0.637	0.152	0.646	0.152
White	0.226	0.036	0.220	0.038	0.222	0.037
Age (@ wave 1)	$4.265^{\dagger}$	$1.477^{\dagger}$	$3.988^{\dagger}$	$1.778^{\dagger}$	$4.034^{\dagger}$	$1.786^{\dagger}$
Squared term	$-0.443^{\dagger}$	$0.054^{\dagger}$	$-0.431^{\dagger}$	$0.062^{\dagger}$	$-0.430^{\dagger}$	$0.062^{\dagger}$
Education (@ wave 1)	-0.069	0.012	-0.078	0.014	-0.077	0.014
Squared term	0.009	0.001	0.010	0.001	0.010	0.001
Type of Worker						
(Employer as baseline)						
Informal	-0.139	0.028	-0.098	0.029	-0.096	0.029
Formal	-0.044	0.029	-0.012	0.029	-0.010	0.029
Military service	-0.009	0.034	0.020	0.034	0.024	0.033
Self-Employed	-0.175	0.024	-0.156	0.024	-0.153	0.024
Type of Activity						
(Manufacturing as baseline)						
Building	0.023	0.040	0.004	0.048	0.001	0.048
Commerce	-0.023	0.027	-0.033	0.026	-0.031	0.026
Financial	0.036	0.027	0.029	0.026	0.032	0.026
Social Services	0.065	0.027	0.066	0.029	0.071	0.029
Domestic Services	-0.002	0.032	0.008	0.034	0.014	0.035
Other Services	-0.002	0.026	0.005	0.024	0.007	0.024
Other Activities	0.038	0.063	0.002	0.069	0.010	0.066
Duration of Employment (× 120)	0.036	0.013	0.029	0.016	0.029	0.017
Squared term	-0.020	0.007	-0.019	0.006	-0.019	0.006
Working Hours (in Log)	0.254	0.027	0.256	0.032	0.257	0.031
Proxy Respondent	0.003	0.008	0.010	0.008	0.012	0.008
Number of HH members	0.005	0.004	0.004	0.004	0.003	0.004
Metropolitan Region						
(Recife as baseline)						
Salvador	0.039	0.030	0.035	0.031	0.036	0.031
Belo Horizonte	0.248	0.028	0.246	0.029	0.251	0.029
Rio de Janeiro	0.215	0.025	0.220	0.026	0.222	0.026
São Paulo	0.360	0.028	0.367	0.029	0.372	0.029
Porto Alegre	0.231	0.031	0.233	0.032	0.240	0.032
Interaction Terms of Male and:						
Age (@ wave 1)	0.002	0.002	0.003	0.002	0.003	0.002
Education (@ wave 1)	0.053	0.014	0.063	0.016	0.062	0.016
Squared term	-0.003	0.001	-0.004	0.001	-0.004	0.001
Type of Activity						
(Manufacturing as baseline)						
Building	-0.036	0.042	-0.014	0.050	-0.013	0.050
Commerce	-0.021	0.029	-0.011	0.028	-0.014	0.028
Financial	-0.052	0.030	-0.047	0.030	-0.050	0.030
Social Services	-0.039	0.034	-0.045	0.036	-0.051	0.036
Domestic Services	-0.125	0.055	-0.107	0.044	-0.113	0.044
Other Services	-0.036	0.029	-0.047	0.028	-0.049	0.028
Other Activities	-0.087	0.074	-0.015	0.075	-0.024	0.072
Duration of Employment (× 120)	0.020	0.015	0.025	0.018	0.025	0.019
Squared term	0.009	0.008	0.008	0.007	0.008	0.008
Working Hours (in Log)	-0.090	0.032	-0.090	0.037	-0.091	0.037
Proxy Respondent	-0.042	0.009	-0.042	0.009	-0.044	0.009
Interaction Terms of White and:						
Type of Worker						

Table 7.8 – continued from previous page

	IGLS		PWIGLS (1)		PWIGLS (2)	
	Coeff	Rob.SE	Coeff	SE	Coeff	SE
(Employer as baseline)						
Informal	-0.058	0.036	-0.082	0.038	-0.086	0.037
Formal	-0.103	0.037	-0.117	0.039	-0.120	0.039
Military service	-0.142	0.044	-0.153	0.045	-0.154	0.044
Self-Employed	-0.015	0.031	-0.021	0.031	-0.025	0.031
Contextual Effects:						
Prop of Formal	-1.817	0.220	-1.619	0.243	-1.667	0.243
Prop of Informal	-2.159	0.252	-1.940	0.277	-1.974	0.278
Prop of Military	-2.017	0.246	-1.944	0.268	-1.963	0.269
Prop of Self-Employed	-2.209	0.250	-1.901	0.278	-1.935	0.278
Average Education	0.071	0.006	0.082	0.007	0.081	0.007
$\hat{\sigma}^2_{u0}$	0.278	$5.158^{\dagger}$	0.288	$7.306^{\dagger}$	0.287	$7.252^{\dagger}$
$\hat{\sigma}^2_{u1}$	-0.004	$0.187^{\dagger}$	-0.004	$0.264^{\dagger}$	-0.004	$0.263^{\dagger}$
$\hat{\sigma}_{u01}$	0.001	$0.012^{\dagger}$	0.001	$0.025^{\dagger}$	0.001	$0.025^{\dagger}$
$\hat{\sigma}^2_{e}$	0.045	$0.323^{\dagger}$	0.041	$1.280^{\dagger}$	0.041	$1.268^{\dagger}$
$-2\times$ *Log-Likelihood*	20,747					
AIC	20,859					
BIC	21,239					

Note: † Values at 10^{-3}.

7.3.2 Application of PWIGLS to Multivariate Multilevel Models

This section presents a similar analysis to the previous one but now considering the fit of multivariate multilevel models. The main difference here is that the extended PWIGLS procedure for the multivariate multilevel model accommodates the PSU level. Hence, this level does not need to be omitted from the analysis and the between PSU variance term can be estimated.

The model fitted in this section is an equivalent model to the one in equation 5.2 fitted in Chapter 5. As already defined, this is a two-level model where level two represents the PSUs, level one the heads of household and the different measurement occasions define the multivariate structure. The general linear lag-dependent covariance structure is imposed to the models fitted, which started with the same set of covariates as those in the models in Chapter 5. For comparison, the same model is fitted also imposing the Toeplitz covariance structure. However, it is not presented in Table 7.10. Table 7.10 presents the results for the models fitted through standard IGLS, robust methods and PWIGLS. When PWIGLS was adopted, the set of weights used were the w^*_j for the PSU level and the adjusted $w^*_{i|j(8)}$ for the heads of household level. These weights were defined in Section 7.2.2. Note that the level one weights (heads of household) are not necessarily constant within the level two units (PSUs). Therefore, the scaling method applied to the level one weights is expected to have an effect on the parameter estimates.

Table 7.9 presents the results for the estimated covariance and autocorrelation matrices for the models imposing first the general linear and then the Toeplitz correlation structure. Results are presented for models without accounting for the weights and models accounting for the weights. The comparison between the estimated covariance terms for the models with and without the weights indicates that the differences are usually of one point in the second decimal place. The small impact of weights on these estimates could be an indication that the sampling of higher level units is not informative (Pfeffermann et al., 1998). One other possibility is that, after the scaling, the level one weights become quite small and close to one. In addition, as also identified in Skinner and Holmes (2003), these level one weights are only weakly correlated (around 0.2) with the outcome variable.

Table 7.10 presents the estimated regression coefficients and their SEs for the multivariate model imposing the general linear lag-dependent covariance structure. A similar behaviour to that observed for the random coefficients model, in the previous section, when comparing robust and standard SEs is observed here. The SEs for the PWIGLS fit of the model are similar to those for the IGLS fit with robust SEs, but are larger for some of the variables. As before, the estimated regression coefficients differ when comparing the results for the two estimation procedures. Differences are also observed for the significance level of some of the interaction terms. This, as before, indicates the need for further model selection when applying the PWIGLS.

Extra model selection was then performed. It started by initially excluding the interaction terms between the variable for males and the variable for duration of employment and its squared term. It was also observed that the interaction terms between the categorical variable for type of activity and the variable for males are no longer significant. One other variable that was no longer significant was the one for the number of household members. The decision to exclude variables when applying the PWIGLS can only be through the examination of the Wald test statistic, for the single parameter, or through the multivariate Wald test statistic, for multiple parameters.

Table 7.9: Covariance Components and Autocorrelation Matrices - IGLS and PWIGLS

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
General Linear Lag-dependent without Weights	$\begin{pmatrix} 0.3058 \\ 0.2645 & 0.3058 \\ 0.2590 & 0.2645 & 0.3058 \\ 0.2543 & 0.2590 & 0.2645 & 0.3058 \\ 0.2219 & 0.2233 & 0.2273 & 0.2275 & 0.3058 \\ 0.2197 & 0.2219 & 0.2233 & 0.2273 & 0.2645 & 0.3058 \\ 0.2178 & 0.2197 & 0.2219 & 0.2233 & 0.2590 & 0.2645 & 0.3058 \\ 0.2152 & 0.2178 & 0.2197 & 0.2219 & 0.2543 & 0.2590 & 0.2645 & 0.3058 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.8649 & 1.0000 \\ 0.8471 & 0.8649 & 1.0000 \\ 0.8315 & 0.8471 & 0.8649 & 1.0000 \\ 0.7255 & 0.7301 & 0.7432 & 0.7438 & 1.0000 \\ 0.7184 & 0.7255 & 0.7301 & 0.7432 & 0.8649 & 1.0000 \\ 0.7122 & 0.7184 & 0.7255 & 0.7301 & 0.8471 & 0.8649 & 1.0000 \\ 0.7039 & 0.7122 & 0.7184 & 0.7255 & 0.8315 & 0.8471 & 0.8649 & 1.0000 \end{pmatrix}$
General Linear Lag-dependent with Weights	$\begin{pmatrix} 0.3145 \\ 0.2755 & 0.3145 \\ 0.2706 & 0.2755 & 0.3145 \\ 0.2666 & 0.2706 & 0.2755 & 0.3145 \\ 0.2336 & 0.2355 & 0.2381 & 0.2388 & 0.3145 \\ 0.2315 & 0.2336 & 0.2355 & 0.2381 & 0.2755 & 0.3145 \\ 0.2291 & 0.2315 & 0.2336 & 0.2355 & 0.2706 & 0.2755 & 0.3145 \\ 0.2269 & 0.2291 & 0.2315 & 0.2336 & 0.2666 & 0.2706 & 0.2755 & 0.3145 \end{pmatrix}$	$\begin{pmatrix} 1.0000 \\ 0.8760 & 1.0000 \\ 0.8606 & 0.8760 & 1.0000 \\ 0.8477 & 0.8606 & 0.8760 & 1.0000 \\ 0.7427 & 0.7490 & 0.7573 & 0.7594 & 1.0000 \\ 0.7360 & 0.7427 & 0.7490 & 0.7573 & 0.8760 & 1.0000 \\ 0.7286 & 0.7360 & 0.7427 & 0.7490 & 0.8606 & 0.8760 & 1.0000 \\ 0.7215 & 0.7286 & 0.7360 & 0.7427 & 0.8477 & 0.8606 & 0.8760 & 1.0000 \end{pmatrix}$

Table 7.9 – continued from previous page

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
Toeplitz without Weights	$\begin{pmatrix} 0.3066 & & & & & & & \\ 0.2605 & 0.3066 & & & & & & \\ 0.2500 & 0.2605 & 0.3066 & & & & & \\ 0.2390 & 0.2500 & 0.2605 & 0.3066 & & & & \\ 0.2297 & 0.2390 & 0.2500 & 0.2605 & 0.3066 & & & \\ 0.2220 & 0.2297 & 0.2390 & 0.2500 & 0.2605 & 0.3066 & & \\ 0.2156 & 0.2220 & 0.2297 & 0.2390 & 0.2500 & 0.2605 & 0.3066 & \\ 0.2090 & 0.2156 & 0.2220 & 0.2297 & 0.2390 & 0.2500 & 0.2605 & 0.3066 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8496 & 1.0000 & & & & & & \\ 0.8154 & 0.8496 & 1.0000 & & & & & \\ 0.7795 & 0.8154 & 0.8496 & 1.0000 & & & & \\ 0.7492 & 0.7795 & 0.8154 & 0.8496 & 1.0000 & & & \\ 0.7240 & 0.7492 & 0.7795 & 0.8154 & 0.8496 & 1.0000 & & \\ 0.7032 & 0.7240 & 0.7492 & 0.7795 & 0.8154 & 0.8496 & 1.0000 & \\ 0.6816 & 0.7032 & 0.7240 & 0.7492 & 0.7795 & 0.8154 & 0.8496 & 1.0000 \end{pmatrix}$
Toeplitz with Weights	$\begin{pmatrix} 0.3153 & & & & & & & \\ 0.2717 & 0.3153 & & & & & & \\ 0.2615 & 0.2717 & 0.3153 & & & & & \\ 0.2512 & 0.2615 & 0.2717 & 0.3153 & & & & \\ 0.2416 & 0.2512 & 0.2615 & 0.2717 & 0.3153 & & & \\ 0.2341 & 0.2416 & 0.2512 & 0.2615 & 0.2717 & 0.3153 & & \\ 0.2271 & 0.2341 & 0.2416 & 0.2512 & 0.2615 & 0.2717 & 0.3153 & \\ 0.2209 & 0.2271 & 0.2341 & 0.2416 & 0.2512 & 0.2615 & 0.2717 & 0.3153 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8615 & 1.0000 & & & & & & \\ 0.8292 & 0.8615 & 1.0000 & & & & & \\ 0.7966 & 0.8292 & 0.8615 & 1.0000 & & & & \\ 0.7661 & 0.7966 & 0.8292 & 0.8615 & 1.0000 & & & \\ 0.7423 & 0.7661 & 0.7966 & 0.8292 & 0.8615 & 1.0000 & & \\ 0.7201 & 0.7423 & 0.7661 & 0.7966 & 0.8292 & 0.8615 & 1.0000 & \\ 0.7004 & 0.7201 & 0.7423 & 0.7661 & 0.7966 & 0.8292 & 0.8615 & 1.0000 \end{pmatrix}$

Table 7.10: Multivariate Multilevel Models

	IGLS			PWIGLS	
	Coeff	SE	Rob.SE	Coeff	SE
Dummies for occasion					
d_0	6.232	0.249	0.285	5.979	0.292
d_1	6.237	0.249	0.285	5.982	0.292
d_2	6.235	0.249	0.285	5.974	0.293
d_3	6.241	0.249	0.285	5.986	0.292
d_{12}	6.270	0.249	0.285	6.011	0.292
d_{13}	6.265	0.249	0.285	6.007	0.292
d_{14}	6.264	0.249	0.285	6.004	0.293
d_{15}	6.269	0.249	0.285	6.011	0.293
Males	0.650	0.075	0.138	0.574	0.167
White	0.226	0.023	0.036	0.216	0.039
Age (@ wave 1)	$4.336^\dagger$	$1.395^\dagger$	$1.48^\dagger$	$2.827^\dagger$	$1.858^\dagger$
squared term	$-0.439^\dagger$	$0.043^\dagger$	$0.056^\dagger$	$-0.467^\dagger$	$0.069^\dagger$
Education (@ wave 1)	-0.067	0.011	0.012	-0.082	0.014
Squared term	0.009	0.001	0.001	0.010	0.001
Type of Worker					
(Employer as baseline)					
Informal	-0.137	0.016	0.029	-0.097	0.032
Formal	-0.045	0.016	0.029	-0.019	0.032
Military service	-0.007	0.021	0.034	0.042	0.037
Self-Employed	-0.174	0.014	0.024	-0.168	0.026
Type of Activity					
(Manufacturing as baseline)					
Building	0.030	0.036	0.037	0.024	0.053
Commerce	-0.023	0.016	0.026	-0.037	0.028
Financial	0.040	0.020	0.027	0.043	0.028
Social Services	0.067	0.020	0.026	0.074	0.031
Domestic Services	-0.001	0.019	0.031	0.008	0.034
Other Services	0.001	0.018	0.026	0.015	0.026
Other Activities	0.037	0.061	0.060	-0.018	0.069
Duration of Employment (× 120)	0.038	0.008	0.013	0.034	0.018
Squared term	-0.021	0.004	0.006	-0.023	0.007
Working Hours (in Log)	0.252	0.011	0.029	0.257	0.035
Proxy Respondent	0.002	0.006	0.008	0.008	0.008
Number of HH members	0.005	0.003	0.004	0.005	0.004
Metropolitan Region					
(Recife as baseline)					
Salvador	0.039	0.033	0.033	0.015	0.039
Belo Horizonte	0.244	0.032	0.029	0.254	0.032
Rio de Janeiro	0.212	0.029	0.028	0.217	0.029
São Paulo	0.359	0.032	0.030	0.368	0.032
Porto Alegre	0.226	0.034	0.033	0.239	0.035
Interaction Terms of Male and:					
Age (@ wave 1)	0.003	0.002	0.002	0.005	0.002
Education (@ wave 1)	0.054	0.013	0.014	0.074	0.016
Squared term	-0.003	0.001	0.001	-0.005	0.001
Type of Activity					
(Manufacturing as baseline)					
Building	-0.045	0.037	0.041	-0.046	0.057
Commerce	-0.021	0.018	0.028	-0.009	0.030
Financial	-0.057	0.022	0.030	-0.063	0.032
Social Services	-0.042	0.023	0.032	-0.056	0.038
Domestic Services	-0.122	0.034	0.054	-0.081	0.043
Other Services	-0.037	0.020	0.029	-0.061	0.032

Table 7.10 – continued from previous page

	IGLS			PWIGLS	
	Coeff	SE	Rob.SE	Coeff	SE
Other Activities	-0.083	0.065	0.072	0.004	0.075
Duration of Employment (× 120)	0.018	0.009	0.015	0.020	0.020
Squared term	0.010	0.005	0.007	0.011	0.008
Working Hours (in Log)	-0.087	0.014	0.034	-0.087	0.041
Proxy Respondent	-0.041	0.007	0.009	-0.039	0.010
Interaction Terms of White and:					
Type of Worker					
(Employer as baseline)					
Informal	-0.060	0.020	0.035	-0.086	0.040
Formal	-0.104	0.020	0.037	-0.114	0.043
Military service	-0.144	0.026	0.045	-0.179	0.050
Self-Employed	-0.015	0.017	0.031	-0.013	0.033
Contextual Effects:					
Prop of Formal	-1.847	0.211	0.234	-1.705	0.237
Prop of Informal	-2.181	0.248	0.267	-1.997	0.268
Prop of Military	-2.068	0.233	0.272	-2.094	0.300
Prop of Self-Employed	-2.281	0.245	0.273	-1.977	0.282
Average Education	0.072	0.006	0.007	0.086	0.008
$-2\times$ *Log-Likelihood*		20,050			
AIC		20,190			
BIC		20,573			

Note: † Values at 10^{-3}.

Table 7.11 presents the results for the "final" multivariate multilevel model selected using the PWIGLS fitting which accounts for the multilevel weights. This table also presents the results for the equivalent model estimated using IGLS with robust estimation methods for the SEs. The residual diagnostics for both models are presented in Appendix C. It is worth mentioning that this model does not include the interaction terms between the occasion dummies and the education and duration of employment variables. These terms were included and then tested through multivariate Wald test but were no longer significant. Table 7.13 presents the estimated residual covariance and autocorrelation matrices for the models in Table 7.11. The PSU level variance was estimated to be equal to 0.010 with SE equal to 0.003. These values are not shown in any of the tables.

7.3.3 Interpretation of the Fixed Parameters Estimates

The fixed part of the model estimated via PWIGLS in Table 7.11 can interpreted as follows:

Time Effect

As this is a multivariate model, the time effect is represented by the different effects of the occasion dummies. They represent different intercepts for each of the occasions. The results in Table 7.11 indicate that the average log-income of heads of household is different for the different occasions. There is an increasing trend with time (keeping all other variables constant and for a baseline head of household).

Gender and Skin Colour Effects

Like in the model presented in Chapter 4, the main effects of the variables for male heads of household and white heads of household cannot be interpreted on their own. These two variables interact with other variables in the model. Observe that the baseline income differential between males and females is now around 70% and between whites and others is around 24%.

Age effect - Experience in Labour Market

Like in the model presented in Chapter 4, age was also considered with a significant squared term and also interacts with the variable for males. This means that the effect of age on income is different for male and female heads of household. The linear term for age has a positive effect and the squared term has a negative effect. This indicates that for both males and females the impact of age on income has an inverted U-shape, as expected. Graphical investigation of these effects, in Figure 7.2(a), shows that by using age as a proxy for experience older male heads of household benefit more from extra years of experience in the labour market than female heads of household. However, at younger ages this behaviour is the opposite, benefiting female heads of household.

Education effect and Duration of Employment

The education variable interacts with the indicator for males. Hence, educational returns are different for male and female heads of household. Figure 7.2(b) shows that the more educated the heads of the household, the higher is the return for education on their labour income. In addition, these returns are even higher for males than for females. The impact of duration of employment on income has an inverted U-shape showing that income initially increases rapidly with experience. This variable is centred on the average of eight years of experience in the firm. This is a similar behaviour as in the model presented in Chapter 4.

Table 7.11: Final Multivariate Multilevel Model: Fixed Parameters Estimates

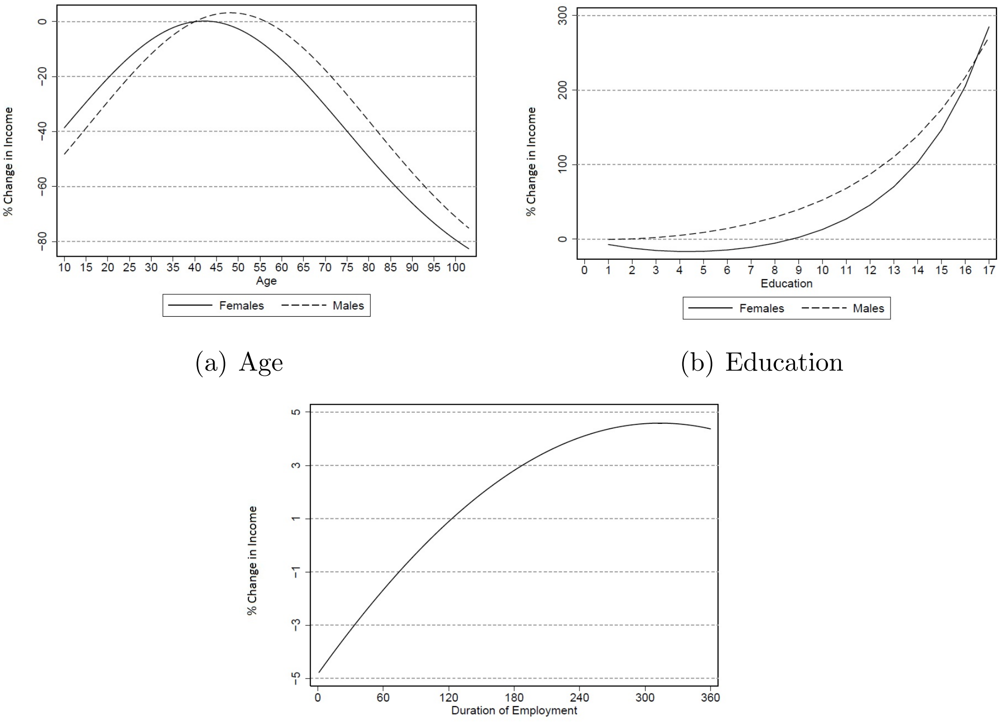
	IGLS		PWIGLS			
	Coeff	Rob.SE	Coeff	SE	z	Pr. > $\lvert z \rvert$
Dummies for occasion						
d_0	6.290	0.285	6.019	0.293	20.58	0.000
d_1	6.295	0.285	6.022	0.292	20.59	0.000
d_2	6.293	0.285	6.014	0.293	20.54	0.000
d_3	6.299	0.285	6.026	0.292	20.60	0.000
d_{12}	6.328	0.285	6.052	0.293	20.69	0.000
d_{13}	6.323	0.285	6.048	0.292	20.68	0.000
d_{14}	6.322	0.285	6.044	0.293	20.66	0.000
d_{15}	6.327	0.285	6.052	0.293	20.68	0.000
Males	0.599	0.137	0.530	0.164	3.22	0.001
White	0.226	0.036	0.216	0.039	5.51	0.000
Age (@ wave 1)	3.485†	1.429†	2.062†	1.767†	1.17	0.243
squared term	-0.45†	0.056†	-0.474†	0.07†	-6.77	0.000
Education (@ wave 1)	-0.069	0.012	-0.083	0.014	-5.91	0.000
squared term	0.009	0.001	0.010	0.001	11.69	0.000
Type of Worker (Employer as baseline)						
Informal	-0.135	0.029	-0.096	0.032	-2.98	0.003
Formal	-0.044	0.029	-0.018	0.032	-0.57	0.566
Military service	-0.006	0.034	0.042	0.037	1.12	0.261
Self-Employed	-0.175	0.024	-0.169	0.026	-6.47	0.000
Type of Activity (Manufacturing as baseline)						
Building	-0.008	0.014	-0.015	0.016	-0.94	0.345
Commerce	-0.041	0.009	-0.044	0.011	-4.02	0.000
Financial	-0.006	0.012	-0.009	0.012	-0.72	0.473
Social Services	0.033	0.015	0.028	0.016	1.74	0.082
Domestic Services	-0.045	0.023	-0.033	0.023	-1.47	0.142
Other Services	-0.029	0.011	-0.035	0.014	-2.43	0.015
Other Activities	-0.028	0.034	-0.011	0.024	-0.46	0.648
Duration of Employment (× 120)	0.053	0.006	0.050	0.009	5.85	0.000
squared term	-0.014	0.003	-0.014	0.004	-3.56	0.000
Working Hours (in Log)	0.251	0.029	0.257	0.036	7.20	0.000
Proxy Respondent	0.004	0.008	0.009	0.008	1.16	0.246
Metropolitan Region (Recife as baseline)						
Salvador	0.037	0.033	0.013	0.040	0.32	0.748
Belo Horizonte	0.244	0.029	0.254	0.032	8.03	0.000
Rio de Janeiro	0.211	0.028	0.217	0.029	7.46	0.000
São Paulo	0.360	0.030	0.368	0.032	11.55	0.000
Porto Alegre	0.226	0.033	0.239	0.035	6.76	0.000
Interaction Terms of Male and:						
Age (@ wave 1)	0.004	0.002	0.006	0.002	2.94	0.003
Education (@ wave 1)	0.057	0.014	0.076	0.016	4.69	0.000
squared term	-0.004	0.001	-0.005	0.001	-4.81	0.000
Working Hours (in Log)	-0.085	0.034	-0.088	0.041	-2.13	0.033
Proxy Respondent	-0.042	0.009	-0.040	0.010	-4.08	0.000
Interaction Terms of White and:						
Type of Worker (Employer as baseline)						
Informal	-0.062	0.035	-0.086	0.040	-2.14	0.032
Formal	-0.105	0.038	-0.114	0.043	-2.66	0.008
Military service	-0.146	0.045	-0.180	0.050	-3.59	0.000
Self-Employed	-0.015	0.031	-0.014	0.033	-0.41	0.684
Contextual Effects:						
Prop of Formal	-1.845	0.234	-1.690	0.237	-7.13	0.000
Prop of Informal	-2.175	0.268	-1.979	0.268	-7.39	0.000
Prop of Military	-2.069	0.273	-2.086	0.299	-6.97	0.000

Table 7.11 – continued from previous page

	IGLS		PWIGLS			
	Coeff	Rob.SE	Coeff	SE	z	Pr. > $\|z\|$
Prop of Self-Employed	-2.280	0.274	-1.963	0.282	-6.96	0.000
Average Education	0.072	0.007	0.086	0.008	10.64	0.000
$-2\times$ *Log-Likelihood*	20,098					
AIC	20,218					
BIC	20,547					
Number of Observations						
Clusters	1,762					
Individuals	6,524					
Time Points	8					

Note: † Values at 10^{-3}.

Figure 7.2: Impact of Education, Age and Duration of Employment on Income



(a) Age

(b) Education

(c) Duration of Employment

Type of worker and Type of Activity

Type of worker interacts with the indicator for white heads of household while type of activity no longer has a significant interaction with the indicator for male. Taking firstly the results for type of activity, compared with manufacturing activity, all other activities, except Social Services, have a negative effect on income. Compared to employers, almost any other type of worker earns less except for others in the military service. The income gap between informal or self-employed compared to employer is narrower for others than for whites. This is also observed for formal workers, where the gap between employers and formal others is quite small. However, this same gap is quite significant for whites. In the military service the race differential is about 2%.

Table 7.12: Percentage Impact on Average Real Labour Income

	Females	Males
***Interaction Terms of Male and*:**		
Working Hours (in Log)	29.31	18.46
Proxy Respondent	0.90	(3.02)
	Others	Whites
***Interaction Terms of White and*:**		
Type of Worker (Employer as baseline)		
Informal	(9.12)	(16.63)
Formal	(1.80)	(12.40)
Military service	4.26	(12.92)
Self-Employed	(15.55)	(16.69)

Note: Values in parentheses indicate % decrease.

Working Hours

Working hours is included in the model on the logarithm scale. The income elasticity is 0.26% for female heads of household and 0.16% for male heads of household (for an increase of 1% in the working hours).

Proxy Respondent

The variable for proxy respondent is a control for those heads of household which are hard to count. Table 7.12 shows that the use of proxy respondents have a negative impact on the income of male heads of household (it could be interpreted as the proxy tending to declare a lower income) but a small positive effect for female heads of household.

Metropolitan Regions

The metropolitan region of Salvador has an income differential of about 1% compared to the baseline. However, this is not significantly different from the baseline which is Recife. When compared to Recife, the Southern regions have income differentials in their favour varying from 24% to 28%. The richest region of São Paulo, however, has the highest differential of 45%.

Other Cluster level effects

Like in the previous models PSU or contextual variables were considered, mostly for empirical reasons, to improve the model fit. They represent how the make up of the PSU is associated with the income of the heads of household. In this final model, however, some of the previous PSU variables were not considered. In a general interpretation for those in Table 7.11, compared to the proportion of employers in the PSUs, the other proportions have a negative effect on the income of the heads of household. In addition, the higher the average education for the PSU the higher is the effect on income of the heads of household in the PSU.

Table 7.13 presents the estimated covariance and autocorrelation matrices for this final model. This table also contains, for comparison, the results for the equivalent model fitted using IGLS and robust methods for the estimation of the standard errors. There is a decreasing trend with time in the autocorrelation matrix. When comparing the results for the models estimated via IGLS with robust methods and PWIGLS the autocorrelations differ mostly in the second decimal place. This can also be observed in Figure 7.3, which presents the autocorrelation function by lag. The figure shows that the correlations for the model estimated via PWIGLS lie very slightly above those of the model estimated via IGLS with robust methods. However, this might still be an indication of a weak effect of weights. Furthermore, this figure shows the fall in the correlation between measurements taken at lag 9.

Figure 7.3: Autocorrelation Function

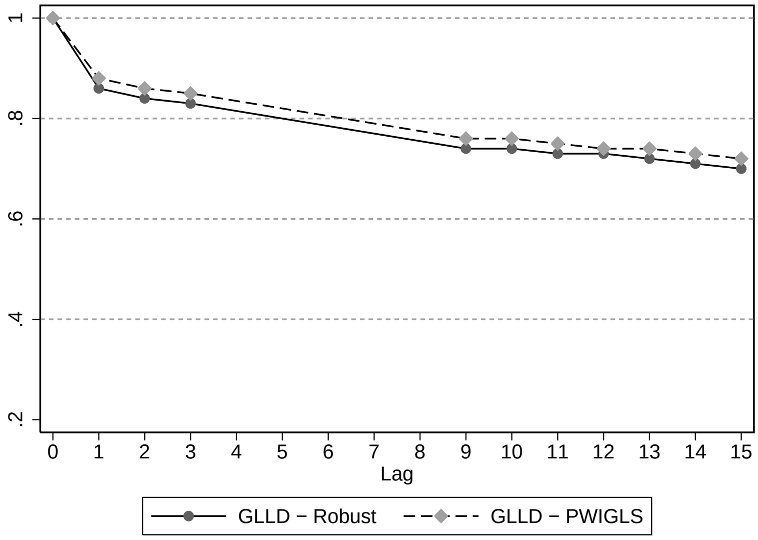


Table 7.13: Covariance Components and Autocorrelation Matrices - Final Multivariate Multilevel Model

	$\widehat{\Sigma}_r$	$\widehat{P}_r$
General Linear Lag-dependent IGLS Robust	$\begin{pmatrix} 0.3064 & & & & & & & \\ 0.2650 & 0.3064 & & & & & & \\ 0.2596 & 0.2650 & 0.3064 & & & & & \\ 0.2548 & 0.2596 & 0.2650 & 0.3064 & & & & \\ 0.2224 & 0.2238 & 0.2279 & 0.2281 & 0.3064 & & & \\ 0.2202 & 0.2224 & 0.2238 & 0.2279 & 0.2650 & 0.3064 & & \\ 0.2183 & 0.2202 & 0.2224 & 0.2238 & 0.2596 & 0.2650 & 0.3064 & \\ 0.2158 & 0.2183 & 0.2202 & 0.2224 & 0.2548 & 0.2596 & 0.2650 & 0.3064 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8649 & 1.0000 & & & & & & \\ 0.8472 & 0.8649 & 1.0000 & & & & & \\ 0.8316 & 0.8472 & 0.8649 & 1.0000 & & & & \\ 0.7258 & 0.7304 & 0.7438 & 0.7444 & 1.0000 & & & \\ 0.7187 & 0.7258 & 0.7304 & 0.7438 & 0.8649 & 1.0000 & & \\ 0.7124 & 0.7187 & 0.7258 & 0.7304 & 0.8472 & 0.8649 & 1.0000 & \\ 0.7043 & 0.7124 & 0.7187 & 0.7258 & 0.8316 & 0.8472 & 0.8649 & 1.0000 \end{pmatrix}$
General Linear Lag-dependent PWIGLS	$\begin{pmatrix} 0.3149 & & & & & & & \\ 0.2759 & 0.3149 & & & & & & \\ 0.2711 & 0.2759 & 0.3149 & & & & & \\ 0.2670 & 0.2711 & 0.2759 & 0.3149 & & & & \\ 0.2339 & 0.2359 & 0.2386 & 0.2393 & 0.3149 & & & \\ 0.2318 & 0.2339 & 0.2359 & 0.2386 & 0.2759 & 0.3149 & & \\ 0.2295 & 0.2318 & 0.2339 & 0.2359 & 0.2711 & 0.2759 & 0.3149 & \\ 0.2272 & 0.2295 & 0.2318 & 0.2339 & 0.2670 & 0.2711 & 0.2759 & 0.3149 \end{pmatrix}$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8760 & 1.0000 & & & & & & \\ 0.8608 & 0.8760 & 1.0000 & & & & & \\ 0.8477 & 0.8608 & 0.8760 & 1.0000 & & & & \\ 0.7428 & 0.7491 & 0.7575 & 0.7598 & 1.0000 & & & \\ 0.7360 & 0.7428 & 0.7491 & 0.7575 & 0.8760 & 1.0000 & & \\ 0.7285 & 0.7360 & 0.7428 & 0.7491 & 0.8608 & 0.8760 & 1.0000 & \\ 0.7215 & 0.7285 & 0.7360 & 0.7428 & 0.8477 & 0.8608 & 0.8760 & 1.0000 \end{pmatrix}$

7.4 One Further Comparison

This section presents the fit of a multivariate multilevel model to a less restrictive data set than the one analysed so far. The aim of this section is to provide one further model comparison, comparing the results presented here with those from the models in Tables 7.11 and 5.7.

The data set to be used in this section includes all employed heads of households who were employed when observed in the working data set (see Chapter 3). Furthermore, to maintain consistency with the longitudinal data set analysed in this chapter and Chapter 5, only the heads of households that by design should have all eight interviews completed were considered. However, here, a complete-case balanced data set is no longer required. Therefore, this unbalanced data set includes heads of household with intermittent wave non-response and also those who dropped out from the panel at some point of the survey. This data selection strategy generates a data set with a variable number of repeated observations for the heads of households in the sample. This variable number of observations is either due to wave non-response or panel attrition and is presented in Table 7.14. Notice that the total number of heads of households per interview time varies and these numbers can be compared to the total number of 6,524 heads of households analysed previously in this chapter and Chapter 5. It is also worth mentioning that only heads of household with valid income data were included. Furthermore, the same criteria to match heads of household used in Chapter 5 was used here.

Table 7.14: Number of Heads of Household per Interview Time

Interview Time	Number of HoHH
1	8,592
2	8,318
3	8,240
4	8,126
5	7,163
6	7,312
7	7,320
8	7,264

The multivariate multilevel model fitted to this unbalanced data set is equivalent to the one fitted in the previous section and presented in Table 7.11. The results for when the model is fitted to the unbalanced data set are presented in Table 7.15. Notice that the error correlation structure imposed in this model was the general linear lag dependent structure. This was to allow comparison. Furthermore, this is one of the structures that accommodates unbalanced data sets. For brevity, the model fitted in this section is referred to hereafter to the comparative model. One other important point to observe is that the comparative model was estimated via IGLS estimation. This is because no longitudinal weights were required to account for panel non-response. The less restrictive data set, used in this section, already accounts for panel non-response by incorporating the records for heads of households with incomplete responses.

Firstly, the results of the model presented in Table 7.11 are compared to the comparative model. Therefore, initially, the comparison is between a model fitted via PWIGLS to a balanced data set and using the longitudinal weights to account for panel non-response (as the one in Table 7.11) and a model fitted via IGLS to an unbalanced data set which includes the units with panel non-response (as the one in Table 7.15). The results for both models are very similar for most of the variables considered. The differences in the coefficient estimates are usually in the second decimal place which leads to no major differences in the interpretation already given in the previous section. Differences are observed mostly for the coefficients of the cluster level variables and for the occasion dummies. For the comparative model the average labour income for these heads of households was estimated to be slightly higher than for the PWIGLS model in Table 7.11. It is also observed that this average tends to increase with time on the first four occasions which is different from the previous model where the average income presented a small drop in the third occasion. The income dynamics behaviour in the second half for both models are generally quite similar.

Table 7.15 also provides the standard errors and the robust standard errors for the comparative model. For most of the variables in the models, the robust standard errors for the comparative model and the standard errors for the model estimated via PWIGLS are very similar, with those for the latter model always larger. Differences are found mostly in the third decimal place, except for the occasion dummies and some of the cluster level variables where the differences are at the second decimal place. Such differences were not enough to lead to different conclusions about the income dynamics for the heads of household.

In Chapter 5 multilevel models were fitted to the same balanced data set as the one used previously in this chapter. However, the robust methods for the estimation of the standard errors were not employed there. The main difference between the models in this chapter and the final model of Chapter 5 is the presence of interaction terms between the occasion dummies and some other variables. As already mentioned in the previous sections, these interaction terms lost their significance once robust methods for the estimation of the standard errors of the coefficient estimates were employed. This lack of consistent results indicates caution is necessary when including these terms in the models. However, comparison between the multivariate model in Table 5.7 and the comparative model, reveals that the effects for the variables in both models do not differ greatly.

Table 7.15: Multivariate Multilevel Model: Fixed Parameters Estimates

		IGLS	
	Coeff	SE	Rob.SE
Dummies for occasion			
d_0	6.306	0.216	0.263
d_1	6.311	0.216	0.263
d_2	6.312	0.216	0.263
d_3	6.317	0.216	0.263
d_{12}	6.346	0.216	0.263
d_{13}	6.339	0.216	0.263
d_{14}	6.339	0.216	0.263
d_{15}	6.343	0.216	0.263
Males	0.575	0.068	0.122
White	0.204	0.021	0.033
Age (@ wave 1)	$3.821^\dagger$	$1.168^\dagger$	$1.223^\dagger$
squared term	$-0.430^\dagger$	$0.037^\dagger$	$0.054^\dagger$
Education (@ wave 1)	-0.058	0.010	0.010
squared term	0.008	0.001	0.001
Type of Worker (Employer as baseline)			
Informal	-0.148	0.015	0.027
Formal	-0.057	0.015	0.027
Military service	-0.021	0.020	0.032
Self-Employed	-0.184	0.013	0.023
Type of Activity (Manufacturing as baseline)			
Building	-0.009	0.009	0.013
Commerce	-0.040	0.006	0.008
Financial	-0.011	0.008	0.011
Social Services	0.029	0.010	0.014
Domestic Services	-0.054	0.014	0.020
Other Services	-0.028	0.007	0.011
Other Activities	-0.042	0.021	0.029
Duration of Employment (× 120)	0.055	0.004	0.006
squared term	-0.012	0.002	0.003
Working Hours (in Log)	0.257	0.010	0.026
Proxy Respondent	-0.003	0.006	0.007
Metropolitan Region (Recife as baseline)			
Salvador	0.044	0.029	0.031
Belo Horizonte	0.225	0.027	0.027
Rio de Janeiro	0.199	0.025	0.026
São Paulo	0.332	0.028	0.028
Porto Alegre	0.214	0.029	0.030
Interaction Terms of Male and:			
Age (@ wave 1)	0.004	0.001	0.001
Education (@ wave 1)	0.053	0.012	0.012
squared term	-0.003	0.001	0.001
Working Hours (in Log)	-0.078	0.013	0.030
Proxy Respondent	-0.035	0.006	0.008
Interaction Terms of White and:			
Type of Worker (Employer as baseline)			
Informal	-0.052	0.018	0.033
Formal	-0.094	0.019	0.035
Military service	-0.120	0.024	0.042
Self-Employed	-0.016	0.016	0.030

Table 7.15 – continued from previous page

		IGLS	
	Coeff	SE	Rob.SE
Contextual Effects:			
Prop of Formal	-1.913	0.183	0.214
Prop of Informal	-2.252	0.216	0.247
Prop of Military	-2.168	0.205	0.241
Prop of Self-Employed	-2.421	0.215	0.250
Average Education	0.077	0.005	0.006
$-2\times$ *Log-Likelihood*		29,089	
AIC		29,209	
BIC		29,542	

Note: † Values at 10^{-3}.

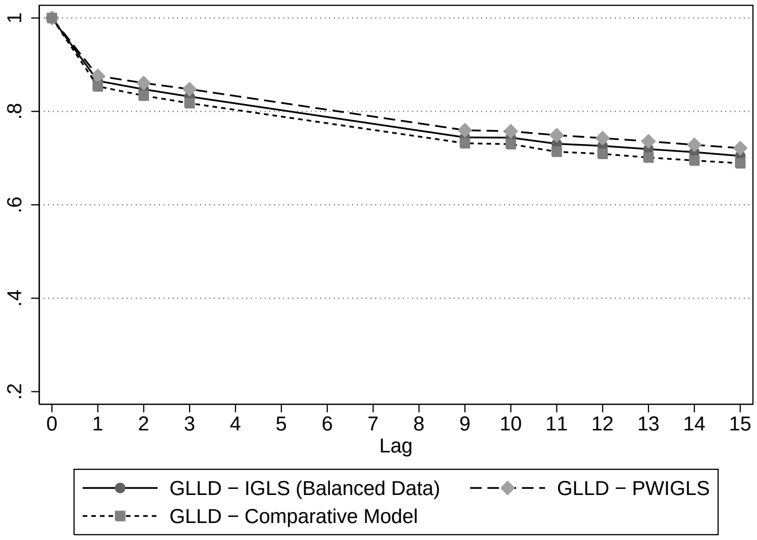
The main aim of Chapter 5 was to account for the panel design while also modelling the error covariance structure via multivariate multilevel models. Table 7.16 presents the estimated covariance and autocorrelation matrices for the comparative model. The results in this table can be compared to those in Tables 5.6, for the model imposing the general linear lag-dependent structure, and 7.13 for the model estimated via PWIGLS. Figure 7.4 presents the autocorrelation functions for these three models for better comparison. The model estimated via PWIGLS in the previous section has slightly larger correlations and the comparative model is the one with less strong correlations per lag. Still these differences are small and in the second decimal place.

The model fitted in this section explored the potential of the multilevel model approach in the analysis of an unbalanced complex panel data set. The results for the model fitted here to a less restrictive data set were compared to those for the models presented in Tables 7.11 and 5.7. The observed differences were not enough to lead to different insights about the income dynamics for the heads of households. However, it does reinforce the small effect of the sampling weights. The use of the sampling weights in the models presented in this chapter mainly increased the estimates of the standard errors of the coefficient estimates. However, the use of robust methods for the estimation of the standard errors, not accounting for the sampling weights, is still advisable as it protects against model misspecification. It is also important to notice that with this comparative model certain types of correlation structures could not be applied, such as the autoregressive ones. This is because the procedure used to fit the models with these structures require the data to be balanced and the observations to be measured at equally spaced time points, and this is not the case for the less restrictive data set used in this section.

Table 7.16: Covariance Components and Autocorrelation Matrices - Multivariate Multilevel Model

	General Linear Lag Dependent
$\widehat{\Sigma}_r$	$\begin{pmatrix} 0.3032 & & & & & & & \\ 0.2588 & 0.3032 & & & & & & \\ 0.2528 & 0.2588 & 0.3032 & & & & & \\ 0.2479 & 0.2528 & 0.2588 & 0.3032 & & & & \\ 0.2150 & 0.2164 & 0.2213 & 0.2219 & 0.3032 & & & \\ 0.2126 & 0.2150 & 0.2164 & 0.2213 & 0.2588 & 0.3032 & & \\ 0.2107 & 0.2126 & 0.2150 & 0.2164 & 0.2528 & 0.2588 & 0.3032 & \\ 0.2089 & 0.2107 & 0.2126 & 0.2150 & 0.2479 & 0.2528 & 0.2588 & 0.3032 \end{pmatrix}$
$\widehat{P}_r$	$\begin{pmatrix} 1.0000 & & & & & & & \\ 0.8536 & 1.0000 & & & & & & \\ 0.8338 & 0.8536 & 1.0000 & & & & & \\ 0.8176 & 0.8338 & 0.8536 & 1.0000 & & & & \\ 0.7091 & 0.7137 & 0.7299 & 0.7319 & 1.0000 & & & \\ 0.7012 & 0.7091 & 0.7137 & 0.7299 & 0.8536 & 1.0000 & & \\ 0.6949 & 0.7012 & 0.7091 & 0.7137 & 0.8338 & 0.8536 & 1.0000 & \\ 0.6890 & 0.6949 & 0.7012 & 0.7091 & 0.8176 & 0.8338 & 0.8536 & 1.0000 \end{pmatrix}$

Figure 7.4: Autocorrelation Function



7.5 Summary and Discussion

This chapter aimed to provide some contribution to the debate on the use of the sampling weights when fitting multilevel models. It started by developing a set of longitudinal weights for the PME data under study. Longitudinal weights were computed based on the logistic regression methodology for calculating response probabilities. Weight adjustments were computed as the inverse of the predicted response probabilities. The base weights were conditionally adjusted to account for the wave-by-wave panel drop-out. The final set of completers with data for eight occasions were weighted to compensate for panel attrition.

Longitudinal weights only accounted for the drop-out. Therefore, some data were eliminated because of wave non-response. An alternative way to account for wave non-response patterns is to use the data until the first non-response occurs and add these extra data to the attrition samples. This approach was not taken here. It is believed that, for the data under study, those who do not respond on one specific occasion but return to the survey afterwards, might be different to those who drop-out from the survey completely. This belief is based upon the problems identified when trying to perform data linkage for the PME data, as non-response for this specific survey might be due to mismatching of individual records or replacement of families in the selected sampled household, for example. However, it is recognized that this method is possibly the appropriate one for other longitudinal surveys. It is worth mentioning that logistic random intercept models were tested for the prediction of the response probabilities. If these were to be considered, one point for further discussion would be the inclusion or not of the estimated random effects in the calculation of the predicted response probabilities. However, for the data under study, these models were shown not to be better than one-level logistic models; so this issue did not arise. It is worth mentioning that there are other methods for calculating longitudinal weights in the literature. However, the exploration of such methods goes beyond the scope of this chapter and thesis.

The application of PWIGLS for both random slope models and multivariate longitudinal models was presented. For random slope models the procedure only accommodates two-level data structure. Further extensions are necessary to accommodate the extra cluster level. However, multivariate models allow for both the imposition of the general linear lag-dependent structure and the inclusion of the random intercept at the cluster level. Results for models fitted using weights were compared with those obtained using both robust and standard estimation methods. It was observed that the standard errors were larger for the models estimated via robust methods when compared to standard methods. However, these were smaller when comparing with those for the models estimated accounting for the sampling weights. Nonetheless, a strong effect of weights was not observed. Reasons were given as the possible non-informativeness of the sampling design. Other potential reasons were due to the very small and not too variable individual weights, which are weakly correlated with the outcome variable, log-income. The model fitted using the weights was also compared to an equivalent model fitted to an extended data set which included the non-respondents. However, the small differences found between these models were not enough to lead to different conclusions.

The final longitudinal analysis presented accounted for the PME data complexities: the data hierarchy was accommodated through the two-level multivariate model including the PSU random intercept; the 4-8-4 rotation scheme was accommodated by imposing the general linear lag-dependent structure for the error covariance; and finally, further features of the sampling design and panel non-response were accommodated via the estimation of the models through the PWIGLS method.

Chapter 8

Conclusions and Further Research

8.1 Introduction

Methods for the analysis of complex panel data were proposed in this thesis. These methods followed the multilevel modelling framework which is based on strong assumptions. This thesis demonstrated how some of these modelling assumptions can be relaxed in order to adapt the methods to better incorporate the complexities of the data: the hierarchical data structure; the complex structure of the residual correlation and the complex sampling design including the rotating panels, the panel attrition and the sampling weights. This chapter provides a summary of this thesis followed by a discussion of the main research contributions, limitations and areas for future research. The key conclusions drawn from this study are then summarised.

8.2 Summary of the Thesis

The data sets commonly used in the social sciences are often obtained from surveys with a complex multistage sampling design selected from populations with natural hierarchical structure. These surveys are sometimes conducted in a repeated manner, comprising separate waves of data collection from the same sampling units over time. When the resulting data set is formed of a large number of sampling units observed over a short period of time, this characterizes a typical panel data set. Panel data can also be generated from a rotating sampling scheme which involves the stratification of the selected sample into panels that are rotated in and out of the survey for a particular period of time. For example, the majority of the labour force surveys in the world adopt rotating sample designs, which are either of a consecutive or of a non-consecutive pattern. Compared to cross-sectional data analysis, panel data analysis has a number of advantages. This type of analysis allows the direct study of individual change and the identification of different effects on change. However, there are also disadvantages to the analysis of panel data, which can be affected by different sources of bias and require more complex methods of analysis.

There are different approaches available for the analysis of panel data and this thesis focussed on methods under the multilevel modelling framework, reviewed in Chapter 2. Panel data sets can be analysed under this framework by assuming that the repeated occasions represent the level one units t, which are nested within individuals i that can be nested within clusters j. Three main multilevel model formulations were reviewed. The random intercept model is the simplest model formulation and assumes that the residual terms are mutually independent, homoscedastic, normally distributed and uncorrelated with the covariates in the model. Furthermore, it assumes that the occasion level residuals are exchangeable, which is not always a plausible assumption particularly when the measurements are taken close in time. The growth curve model relaxes this assumption, by treating the effect of the continuous variable for time as random at the individual level. This causes the residual covariance matrix at the individual level to depend on the metric used for the time variable. The multivariate multilevel model extends the growth curve model by assuming conditional correlation between the repeated outcomes within the same individuals. In addition, this model usually treats time as a discrete variable and allows the modelling of the residual covariance matrix.

The Brazilian labour force survey (the PME) motivated many of the aims of this thesis. The data set from this survey was used throughout the thesis where the labour income dynamics of employed heads of households was investigated. Chapter 3 presented a description of the PME sampling design which includes a non-consecutive rotating panel scheme characterised as 4-8-4. Problems with the matching of the monthly PME data are well known and the low household matching rates are partly caused by panel non-response. In order to investigate change at the individual level an alternative was to consider data for only heads of household. These are the household reference unit and are easier to identify from the data set. Further matching criteria were used to ensure a more accurate matching of the set of heads of households.

Chapter 4 presented a brief review of the Brazilian economy, focussing on some studies of labour income determination mostly based upon the Mincer equation model. This chapter also presented a full cross-sectional multilevel analysis of the determinants of the log of the real labour income. This analysis accounted for the PSU and the heads of household levels. The fitted model was a more elaborate model than the one specified by the Mincer equation and it served as the basis for the applications in the later chapters. The data set considered in this analysis was composed of the first interview for employed heads of household that were first included in the PME survey from January 2004 till December 2005. The filter for employed heads of household was used because only those classified as employed have data for job earnings.

Following this cross-sectional modelling application, Chapter 5 analysed the longitudinal working data set containing complete sets of eight interviews. The main aim of this chapter was to fit multilevel longitudinal models incorporating the rotating sample design of the PME. This was achieved by expressing the time variable as $(0, 1, 2, 3, 12, 13, 14, 15)$ instead of $(0, 1, 2, 3, 4, 5, 6, 7)$ in the fitting of growth curve models. The incorporation of the rotating sample design into the fitting of multivariate multilevel models was not straightforward. However, this was achieved by constraining the residual covariance matrix to impose a lag-dependent structure. Different structures were considered, but most of them are only appropriate to equally spaced panel data, which is not the case for the PME rotating design. Two modified lag-dependent covariance structures were considered: the temporal power and the general linear lag-dependent. Multivariate multilevel models were fitted utilizing the alternative structures. The model using the general linear lag-dependent structure provided good results overall. However, the model using an unconstrained structure, again, provided a better fit. The unconstrained model imposed no constraint on the covariance structure in which the effect of the rotating design was completely determined by the data. The model imposing the general linear lag-dependent structure was a more parsimonious choice. Chapter 6 presented a detailed description of the probability-weighted iterative generalised least squares (PWIGLS) as presented by Pfeffermann et al. (1998). In order to allow for the estimation of multivariate multilevel models via PWIGLS, this method needed to be extended. This extension was presented in Chapter 6 and the necessary computer codes were presented in the appendix.

Chapter 7 dealt with the estimation of longitudinal multilevel models through PWIGLS. For that, longitudinal sampling weights that compensated for panel attrition in the PME data were constructed based upon the fitting of logistic regression models to estimate response probabilities. Two main applications of the

PWIGLS method were presented: for a two-level growth curve model and for multivariate multilevel models. In both applications only the set of completers from the PME longitudinal data set was considered. The results for the applications of the PWIGLS method were compared with results obtained by the fitting of models using robust estimation methods for the standard errors of the regression coefficients. In general, a strong effect of the sampling weights was not observed. This was identified as possibly being due to the non-informativeness of the sampling design or due to the individual level weights having values close to one and being not too variable and weakly correlated with the outcome variable. Those results were also compared to a less restricted data set that included the non-respondents but no different insights on income dynamics were raised.

The final longitudinal analysis presented in Section 7.3 accounted for the main data complexities in a single modelling exercise. The data hierarchy was accounted for by the fitting the two-level multivariate model which also accounted for the rotation scheme by constraining the error covariance to be lag-dependent. Furthermore, both the sampling weights and panel non-response were accounted for through the use of the extended PWIGLS method for fitting multivariate multilevel models.

8.3 Main Contributions

This thesis provides both methodological and substantive contributions. The main methodological contribution was the development of a modelling framework motivated by the data complexities generated by the design of complex panel surveys. This framework consists of the estimation of longitudinal multilevel models via the PWIGLS method, with the use of longitudinal sampling weights and with the complex residual covariance structure of the data being accounted for. The extended methods can be applied to both longitudinal data sets or cross sectional data sets. The developed framework accommodates different error covariance structures provided that they are expressed as linear functions of the covariance parameters. The specific contributions of this thesis are as follows.

- **The application of multivariate multilevel models in the analysis of complex panel data.** Multivariate multilevel models have been used in the analysis of longitudinal data sets but not vastly explored in the literature. They are extensions to growth curve models and the applications presented in this thesis provided a demonstration of the modelling of different covariance structures.

- **The proposal of alternative modified covariance structures.** This was pursued in order to adapt some of the conventional structures to accommodate the rotating design features. The alternative covariance structures were classified as lag-dependent structures, i.e. as being dependent on the time span between the different occasions. These structures can be further applied to any general rotating sampling schemes.

- **The construction of longitudinal weights.** The incorporation of the sampling weights in the analysis of multilevel models has been discussed. Furthermore, a method for constructing longitudinal weights, which accounted for the panel attrition in the PME data under study, was developed and implemented.

- **The extension of the PWIGLS method.** The PWIGLS method was extended to accommodate the estimation of multivariate multilevel models. Applications of the PWIGLS estimation method were pursued for both the growth curve model and the multivariate multilevel model using modified alternative covariance structures and the longitudinal weights constructed for the data under study. The extended PWIGLS method can also be applied to general multivariate regression models.

- **Computational aspects.** Computer codes were developed to fit the growth curve model and the multivariate multilevel model via PWIGLS. These codes were written in the `MATA` language for the `STATA` package.

Substantive contributions from this thesis are identified below.

- **Elaborate modelling and analysis of labour income dynamics.** Although the focus of this thesis was methodological, the applications presented a more elaborate model for the labour income determinants than the ones reviewed in this thesis. Interpretation of the effects of the coefficients estimated in the models was also provided.

- **Use of the PME data set.** Most of the analyses using the PME data set, as reviewed in the literature, consisted of repeated cross-sectional analyses. This thesis demonstrated how to take advantage of the PME panel component. Difficulties when dealing with the data set of this survey, with regard to the matching individual records, may represent a drawback. However, this data set can still be used in the investigation of important dynamic processes of the Brazilian economy.

8.4 Limitations and Recommendations for Further Research

The results of this thesis have led to the identification of limitations and areas requiring further research. These are summarised below.

The extended PWIGLS method only accommodates linear covariance structures. Therefore, structures such as the autoregressive or the temporal power cannot be modelled within the extended method. Further extensions would be necessary to allow for this. Extensions would also be necessary to the PWIGLS method to accommodate the estimation of three-level growth curve models. Computer codes also need to be developed for this situation.

The method adopted to construct the longitudinal weights only considered individuals who followed one of the attrition patterns and the completers. This resulted in the elimination of individuals with intermittent non-response patterns. An alternative method identified in the literature is to transform the wave non-response patterns into attrition patterns by considering the individual data until their first non-responses occur. As mentioned, this method was not undertaken in this thesis. The assumption that wave non-response and panel drop-out were fundamentally different in the data under study supported the decision to only deal with the attrition patterns. However, a possible extension of this analysis is to investigate the effect of considering the wave non-response patterns when calculating the longitudinal weights for the specific data under study. Another topic of further research is to explore the other methods for calculating longitudinal weights and other methods for modelling the drop-out mechanism.

In most of the applications presented in this thesis, the data set considered was a complete-case set. One of the advantages of analysing panel data under the multilevel model framework is that this framework allows the fitting of unbalanced and unequally spaced longitudinal data. The decision to keep a complete-case data set throughout the thesis was taken for simplicity and due to the methodological motivations. Although the comparison between models fitted to the complete-case data set and a data set including non-respondents was provided this could still be extended to also consider the inclusion of units that, by design, do not contain all the waves of the data. However, this would also represent not only panel non-response but also panel censoring.

It was also observed in the longitudinal analysis that the income for the heads of household was relatively stable over time. In other words, not much individual change was observed in the analyses presented in this thesis. This limitation might be due to the design of the PME, which collects monthly data on income, and its short time span. This income stability could also be a result of the selection of only heads of household as the units of analysis, due to the problems in the matching of the monthly PME data. If these problems could have been overcome, the selection of the units of analysis could be extended to include, for example, all household members. One other point to observe was the small effect of the weights on the estimates when compared with models fitted using robust or standard methods. This might also be a reflection of the sampling design of the PME, indicating that the sampling is not informative. However, the data set used in this thesis served its purpose which was the demonstration of the applicability of the methods developed in this thesis.

8.5 Concluding Remarks

This thesis aimed to develop methods for the analysis of complex panel data using the multilevel modelling framework. A modelling estimation procedure was developed to account for data complexities such as the hierarchical data structure, the complex residual correlation and the features of the sampling design which included the sampling weights, the rotating panels and the panel non-response. This procedure is based upon the probability-weighted iterative generalised least squares estimation (Pfeffermann et al., 1998) and can be applied to both complex longitudinal data and cross-sectional data sets.

The applicability of the different model formulations under the multilevel model framework for the analysis of longitudinal data was explored and compared. These methods are already quite robust. However, analyses accounting for the sampling design features can be compared to those using the iterative generalised least squares (IGLS) estimation method with and without robust estimation methods for the standard errors. Large differences in the results of these analyses are indications of either informativeness of the sampling weights or model misspecification.

Under informative panel sampling design and given the availability of the design variables and sampling weights, the PWIGLS framework developed in this thesis provides appropriate tools for the analysis of complex panel survey data. In the absence of such variables, robust methods that protect against model misspecification should be employed. Multivariate multilevel models which allow the modelling of the correlation between the repeated outcomes are a preferable approach. Lag-dependent residual covariance structures can be developed and used with the multivariate multilevel models. These can be compared to the conventional growth curve models, which when applied should carefully account for the appropriate metric of time. Random intercept models are not advisable if the main objective of the analysis is to account for the various data complexities of complex panel survey data.

Appendix A

SAS Computer Codes

A.1 Fitting the Temporal Power Structure

```
proc mixed data=long_w8_sas covtest noclprint noinfo method=reml;
class cluster HH_ID wave015;
model log_incomeR=wave015 / noint solution ddfm=bw;
random intercept / subject=cluster type=un;
repeated wave015 / subject=HH_ID(cluster) type=sp(pow)(wave015)
r;
run;
quit;
```

A.2 Fitting the General Linear Lag Dependent Structure

```
data glls;
input parm row col1-col8;

datalines;
1 1 1 0 0 0 0 0 0 0
1 2 0 1 0 0 0 0 0 0
1 3 0 0 1 0 0 0 0 0
1 4 0 0 0 1 0 0 0 0
1 5 0 0 0 0 1 0 0 0
1 6 0 0 0 0 0 1 0 0
1 7 0 0 0 0 0 0 1 0
1 8 0 0 0 0 0 0 0 1

2 1 1 1 0 0 0 0 0 0
2 2 1 1 1 0 0 0 0 0
2 3 0 1 1 1 0 0 0 0
2 4 0 0 1 1 0 0 0 0
2 5 0 0 0 0 1 1 0 0
2 6 0 0 0 0 1 1 1 0
2 7 0 0 0 0 0 1 1 1
2 8 0 0 0 0 0 0 1 1

3 1 1 0 1 0 0 0 0 0
3 2 0 1 0 1 0 0 0 0
3 3 1 0 1 0 0 0 0 0
3 4 0 1 0 1 0 0 0 0
3 5 0 0 0 0 1 0 1 0
3 6 0 0 0 0 0 1 0 1
3 7 0 0 0 0 1 0 1 0
3 8 0 0 0 0 0 1 0 1
```

```
4 1 1 0 0 1 0 0 0 0
4 2 0 1 0 0 0 0 0 0
4 3 0 0 1 0 0 0 0 0
4 4 1 0 0 1 0 0 0 0
4 5 0 0 0 0 1 0 0 1
4 6 0 0 0 0 0 1 0 0
4 7 0 0 0 0 0 0 1 0
4 8 0 0 0 0 1 0 0 1

5 1 1 0 0 0 0 0 0 0
5 2 0 1 0 0 0 0 0 0
5 3 0 0 1 0 0 0 0 0
5 4 0 0 0 1 1 0 0 0
5 5 0 0 0 1 1 0 0 0
5 6 0 0 0 0 0 1 0 0
5 7 0 0 0 0 0 0 1 0
5 8 0 0 0 0 0 0 0 1

6 1 1 0 0 0 0 0 0 0
6 2 0 1 0 0 0 0 0 0
6 3 0 0 1 0 1 0 0 0
6 4 0 0 0 1 0 1 0 0
6 5 0 0 1 0 1 0 0 0
6 6 0 0 0 1 0 1 0 0
6 7 0 0 0 0 0 0 1 0
6 8 0 0 0 0 0 0 0 1

7 1 1 0 0 0 0 0 0 0
7 2 0 1 0 0 1 0 0 0
7 3 0 0 1 0 0 1 0 0
7 4 0 0 0 1 0 0 1 0
7 5 0 1 0 0 1 0 0 0
7 6 0 0 1 0 0 1 0 0
7 7 0 0 0 1 0 0 1 0
7 8 0 0 0 0 0 0 0 1

8 1 1 0 0 0 1 0 0 0
8 2 0 1 0 0 0 1 0 0
8 3 0 0 1 0 0 0 1 0
8 4 0 0 0 1 0 0 0 1
8 5 1 0 0 0 1 0 0 0
8 6 0 1 0 0 0 1 0 0
8 7 0 0 1 0 0 0 1 0
8 8 0 0 0 1 0 0 0 1

9 1 1 0 0 0 0 1 0 0
9 2 0 1 0 0 0 0 1 0
9 3 0 0 1 0 0 0 0 1
9 4 0 0 0 1 0 0 0 0
9 5 0 0 0 0 1 0 0 0
9 6 1 0 0 0 0 1 0 0
9 7 0 1 0 0 0 0 1 0
9 8 0 0 1 0 0 0 0 1

10 1 1 0 0 0 0 0 1 0
10 2 0 1 0 0 0 0 0 1
10 3 0 0 1 0 0 0 0 0
10 4 0 0 0 1 0 0 0 0
10 5 0 0 0 0 1 0 0 0
10 6 0 0 0 0 0 1 0 0
10 7 1 0 0 0 0 0 1 0
10 8 0 1 0 0 0 0 0 1

11 1 1 0 0 0 0 0 0 1
11 2 0 1 0 0 0 0 0 0
11 3 0 0 1 0 0 0 0 0
11 4 0 0 0 1 0 0 0 0
11 5 0 0 0 0 1 0 0 0
11 6 0 0 0 0 0 1 0 0
11 7 0 0 0 0 0 0 1 0
```

```
11 8 1 0 0 0 0 0 0 1
;
run;

proc mixed data=long_w8_sas covtest noclprint noinfo method=reml;
class cluster HH_ID wave015;
model log_incomeR=wave015 / noint solution ddfm=bw;
random intercept / subject=cluster type=un;
repeated wave015 / subject=HH_ID(cluster) type=lin(11) ldata=glld r;
run;
quit;
```

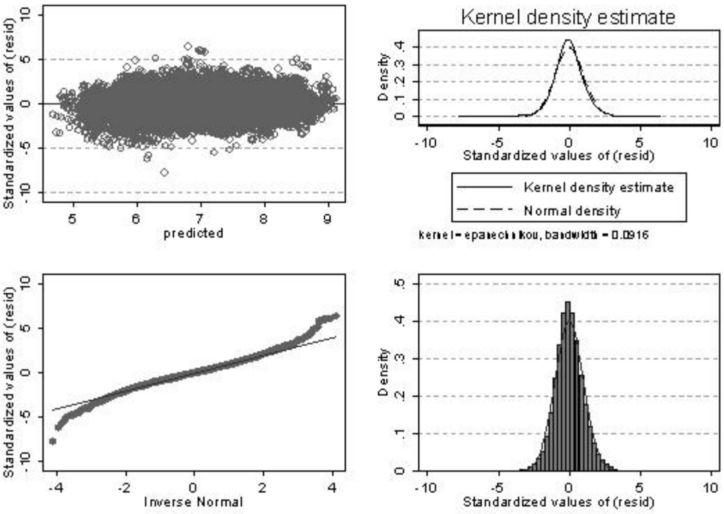
Appendix B

Additional Tables and Figures for Chapter 5

B.1 Level One Residuals

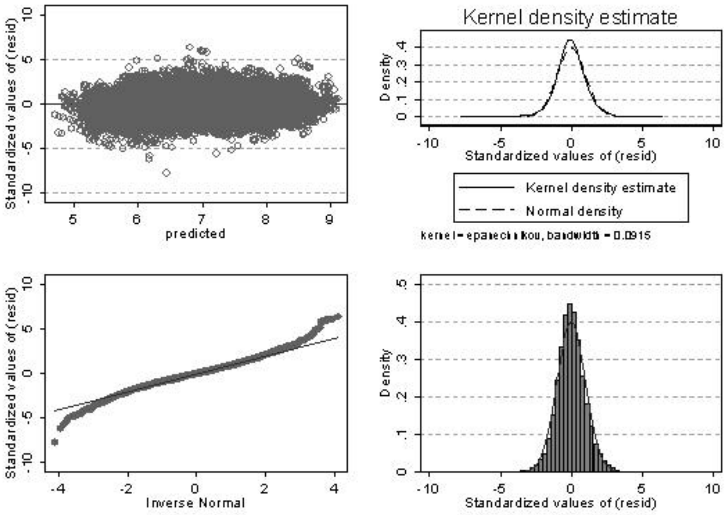
B.1.1 Multivariate Model with General Linear Lag Dependent Covariance Structure

Figure B.1: Residual Diagnostic - Level one residuals



B.1.2 Multivariate Model with Unconstrained Covariance Structure

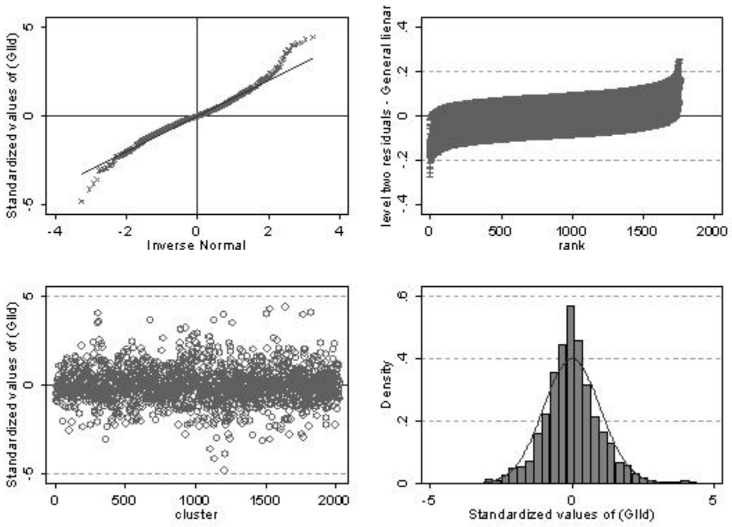
Figure B.2: Residual Diagnostic - Level one residuals



B.2 Level Two Residuals

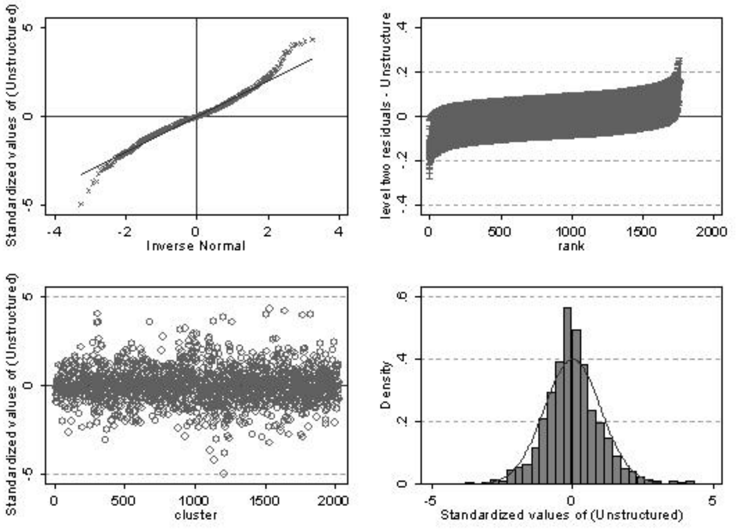
B.2.1 Multivariate Model with General Linear Lag Dependent Covariance Structure

Figure B.3: Residual Diagnostic - Level two residuals



B.2.2 Multivariate Model with Unconstrained Covariance Structure

Figure B.4: Residual Diagnostic - Level two residuals



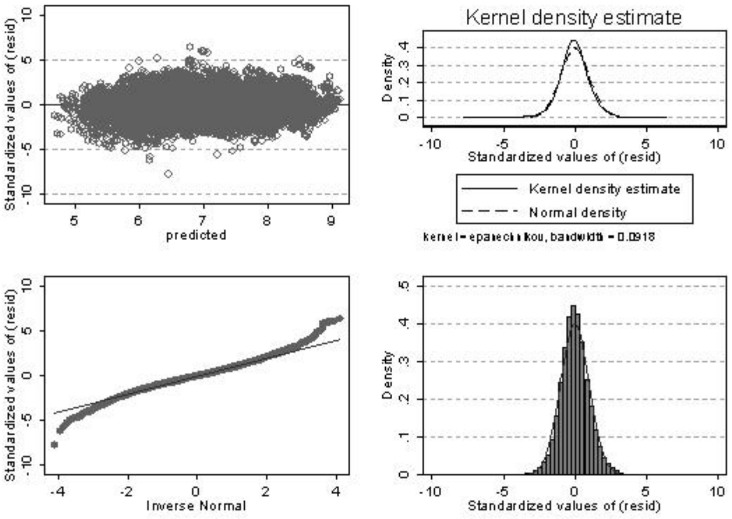
Appendix C

Additional Tables and Figures for Chapter 7

C.1 Level One Residuals

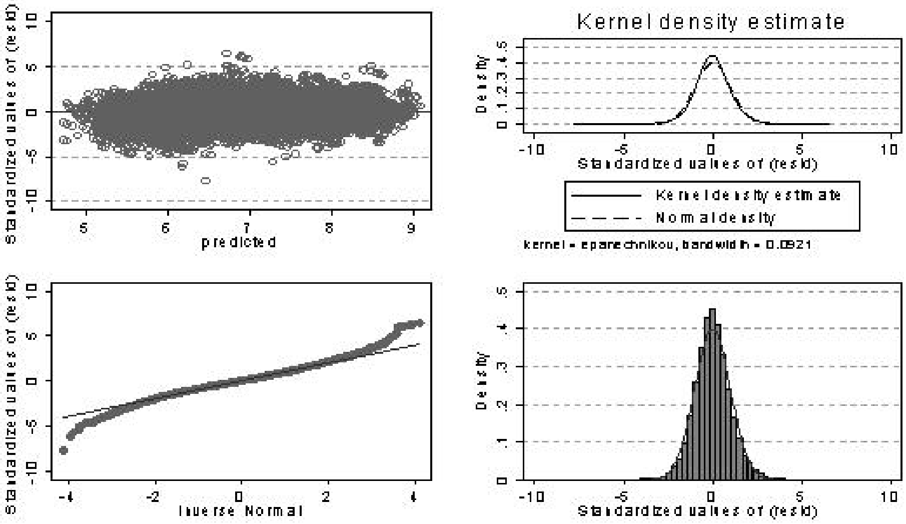
C.1.1 Multivariate Multilevel Model Estimated via IGLS with Robust Estimation Methods for the SEs

Figure C.1: Residual Diagnostic - Level one residuals



C.1.2 Multivariate Multilevel Model Estimated via PWIGLS

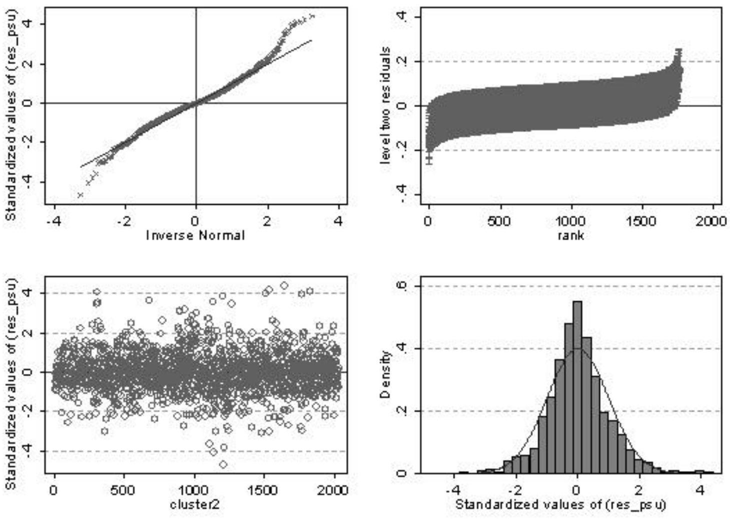
Figure C.2: Residual Diagnostic - Level one residuals



C.2 Level Two Residuals

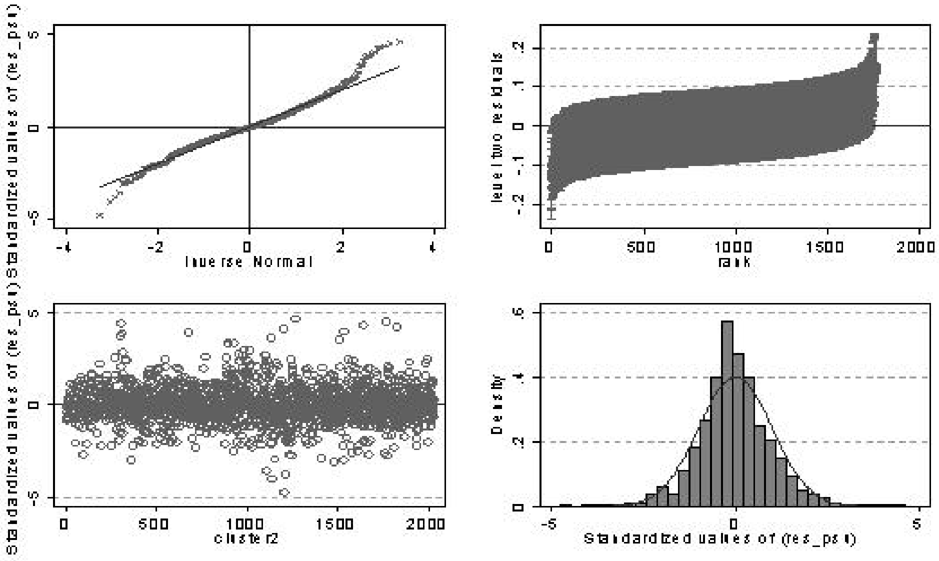
C.2.1 Multivariate Multilevel Model Estimated via IGLS with Robust Estimation Methods for the *SE*s

Figure C.3: Residual Diagnostic - Level two residuals



C.2.2 Multivariate Multilevel Model Estimated via PWIGLS

Figure C.4: Residual Diagnostic - Level two residuals



Appendix D

Computer Codes for the PWIGLS

D.1 Computer Code: Probability Weighted Iterative Generalized Least Squares for Random Slope Model

Based on the computer codes published in Corrêa (2001). These codes were developed in Mata for Stata.

```
/******************************************************************************/
/* Probability Weighted Iterative Generalized Least Squares                   */
/* for two level random coefficient models                                    */
/******************************************************************************/

version 9

mata:

void pwigls_2l_adcv( string varlist,  dep, string varlist1, cluster_var,  _wj_rep,  _wi_j )
{
start= st_global("c(current_time)")
today= st_global("c(current_date)")
x = st_data(., tokens(varlist))
y = st_data(.,tokens(dep))
z = st_data(.,tokens(varlist1))
cluster_var = st_data(.,tokens(cluster_var))
cluster= uniqrows(cluster_var)
wj_rep = st_data(.,tokens(_wj_rep))
wi_j = st_data(.,tokens(_wi_j))

nvar = cols(x)
ncluster = rows(cluster)
nsubjc_t=rows(y)
q = cols(z)
s = ((q*(q+1))/2)+1

h_matrix= ((I(s)[vec(makesymmetric(invvech(1::s-1))), ]))'

name1 = tokens(varlist)
name3 = ("sigma2_u0","sigma_u10","sigma2_u1" ,"sigma2_e")

lambidaj = J(nsubjc_t,1,0)
info_j=panelsetup(cluster_var,1)
cluster_wgt=J(ncluster,1,0)
```

```
/**************************************************************/
/*-------------- Calculating the Scaled Weights --------------*/
/*-------------- Scaling Method 2              --------------*/
/**************************************************************/

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 kwi_j=panelsubmatrix(wi_j, i, info_j)
 np = rows(nsubjc)
 k =mean(kwi_j)
 parte = J(np,1,k)

 a=info_j[i,1]
 b=info_j[i,2]
 lambidaj[a..b,]=parte

 wj_rep_j=panelsubmatrix(wj_rep, i, info_j)

cluster_wgt[i,]=uniqrows(wj_rep_j)
}

wi_j_star = wi_j :/ lambidaj
wj_star = cluster_wgt / mean(cluster_wgt)
wjesc_r = wj_rep / mean(cluster_wgt)

  /**************************************************************/
 /* --------- Calculating Beta_zero and Theta_zero--------------*/
/**************************************************************/

mat_t1j = J(ncluster,nvar^2,0)
mat_t3j = J(ncluster,nvar,0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 np = rows(nsubjc)

 wi_j_starj=panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(wi_j_starj)
 wj = wj_star[i,]

 mat_t1j[i,] = (vec( ( xj' * diag * xj ) * wj ))'
 mat_t3j[i,] = (vec( ( xj' * diag * yj ) * wj))'
}

somat1 = rowshape(colsum(mat_t1j),nvar)
somat3 = colsum(mat_t3j)

beta0 = cholinv(somat1) * (somat3')

vec_t6 = J(ncluster,1,0)
vec_aux = J(ncluster,1,0)
theta0 =  (vech(diag(0.5):*I(q))\0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 np = rows(nsubjc)

 wi_j_starj=panelsubmatrix(wi_j_star, i, info_j)
 wj = wj_star[i,]

 resid = yj - xj * beta0

 uj0 = (wi_j_starj' * resid ) / colsum(wi_j_starj)

 vec_t6[i,] = (wi_j_starj' * ( (resid :- uj0) :^ 2 ) ) * wj
 vec_aux[i,]= wj * (colsum(wi_j_starj)- 1 )
}
```

```
 theta0[s,]=colsum(vec_t6)/colsum(vec_aux)

  /*************************************************/
 /*----------- ITERATIVE MODULE-------------------*/
/*************************************************/

matp = J(ncluster,nvar^2,0)
matq = J(ncluster,nvar,0)

beta_ant = beta0
beta = beta_ant:*2

theta_ant=theta0
theta=theta_ant :* 2

itera = 1
 while (itera<=200 & (any(abs((theta-theta_ant)):> 0.000001) | any(abs((beta-beta_ant)):
 > 0.000001) ))
{

/*-------------looping for beta------------------*/

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)

 v=panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(v)

 wj = wj_star[i,]

 t1j = ( xj’ * diag * xj )
 t2j = ( xj’ * diag * zj )
 t3j = ( xj’ * diag * yj )
 t4j = ( zj’ * diag * yj )
 t5j = ( zj’ * diag * zj )

 if (itera ~= 1) {
 aj= cholinv( t5j + theta[s,]:*(cholinv(invvech(theta[1..(s-1),]))))
 }
 else {
 aj= cholinv( t5j + theta0[s,]:*(cholinv(invvech(theta0[1..(s-1),]))))
 }

 matp[i,] = (vec ( wj :* (t1j - t2j * aj * t2j’) ) )’
 matq[i,] = (vec ( wj :* (t3j - t2j * aj * t4j ) ) )’
}

/*----------- beta------------- */
s_matp = rowshape(colsum(matp),nvar)
s_matq = colsum(matq)

if (itera ~= 1){
beta_ant = beta
 }

beta = cholinv(s_matp) * (s_matq’)

/*-------- looping for theta=inv(R)x S -------------*/
 q = cols(z)
 s = ((q*(q+1))/2)+1

R = J(ncluster,s*s,0)
S = J(ncluster,s,0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)

 v=panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(v)

 wi_j_starj=panelsubmatrix(wi_j_star, i, info_j)
 wj = wj_star[i,]

 t1j = ( xj' * diag * xj )
 t2j = ( xj' * diag * zj )
 t3j = ( xj' * diag * yj )
 t4j = ( zj' * diag * yj )
 t5j = ( zj' * diag * zj )
 eij = yj - xj * beta

Rklj = J(s,s,0)
Skj = J(s,1,0)
B= J(s,q*q,0)
C = J(s,q*q,0)

H = h_matrix

delta = J(1,s,0)
delta[.,s]=1

 if (itera ~= 1)
{
aj= cholinv( t5j + theta[s,]:*(cholinv(invvech(theta[1..(s-1),]))))
for (k=1; k<=s ; k++){
B[k,]=(vec(theta[s,]*aj*cholinv(invvech(theta[1..(s-1),]))* rowshape(H[k,],q)- delta[,k]*aj ))'
C[k,]=(vec(-delta[,k]*aj + rowshape(B[k,],q)' - rowshape(B[k,],q)'* t5j * aj ))'
for (l=1; l<=s ; l++){
Rklj[k,l]= wj*(delta[.,k]*delta[.,l]*colsum(v) + delta[.,l]*trace(t5j*rowshape(C[k,],q)')+
 delta[.,k]*trace(t5j*rowshape(H[l,],q))+trace(t5j*rowshape(C[k,],q)'*t5j*rowshape(H[l,],q)))
}
Skj[k,]= wj*(delta[.,k]*trace(eij'*diag*eij) +trace(eij'*diag*zj * rowshape(C[k,],q)' *
 zj'*diag*eij))
 }
 }
 else
{
 aj= cholinv( t5j + theta0[s,]:*(cholinv(invvech(theta0[1..(s-1),]))))
for (k=1; k<=s ; k++){
B[k,]=(vec( theta0[s,]*aj*cholinv(invvech(theta0[1..(s-1),]))* rowshape(H[k,],q)- delta[,k]*aj ))'
C[k,]=(vec(-delta[,k]*aj + rowshape(B[k,],q)' - rowshape(B[k,],q)'*t5j*aj))'
for (l=1; l<=s ; l++){
Rklj[k,l]= wj*(delta[.,k]*delta[.,l]*colsum(v) + delta[.,l]*trace(t5j*rowshape(C[k,],q)')+
 delta[.,k]*trace(t5j*rowshape(H[l,],q))+trace(t5j*rowshape(C[k,],q)'*t5j*rowshape(H[l,],q)))
}
Skj[k,]= wj*(delta[.,k]*trace(eij'*diag*eij) +trace(eij'*diag*zj * rowshape(C[k,],q)' *
 zj'*diag*eij))
}
}
R[i,]=(vec(Rklj))'
S[i,]=(vec(Skj))'
}
 matr=colsum(R)
 mats=colsum(S)

 r_mat = rowshape(matr,s)
 s_mat = rowshape(mats,s)

if (itera ~= 1) {
 theta_ant = theta
}
theta = cholinv(r_mat) * s_mat
```

```
itera = itera + 1
}

/* Number of Iterations*/
n_it = itera - 1

  /************************************************/
 /*----------End of iterative process--------------*/
/************************************************/


  /***************************************************/
 /*------------------ Residuals-----------------------*/
/***************************************************/

u = J(ncluster,q, 0)
var_u = J(ncluster,q*q,0)
dp_u = J(ncluster,q,0)

yhat = J(nsubjc_t,1, 0)

v = J(nsubjc_t,1, 0)

var_v = J(nsubjc_t,1, 0)


for (i=1; i<=rows(info_j) ; i++){

 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)

 ej = yj - xj * beta

 a=info_j[i,1]
 b=info_j[i,2]

 Sigmau=(invvech(theta[1..(s-1),]))

 Rhj = Sigmau* zj'

 Vj = zj * Sigmau * zj' + theta[s,] :* I(np)

 aux =  Rhj * cholinv(Vj)


 u[i,] = (aux * ej)'

 var_u[i,] = (vec(Sigmau - aux * Rhj'))'
 dp_u[i,] = (vec(sqrt(diagonal(rowshape(var_u[i,],q) ))))'

 yhat1 = xj * beta + zj*u[i,]'
 yhat[a..b,] = yhat1

 vj = ej - zj*u[i,]'
 v[a..b,] = vj

 aux = theta[s,] :* ( 1 - (1/np) )

 var = J(np,1,aux)
 var_v[a..b,] = var
}

u_pad = u :/  dp_u
v_pad = v :/ sqrt(var_v)


/********************************************************/
```

```
/*-------------- Variances of Beta and Theta--------------*/
/********************************************************/

mat_c = J(ncluster,nvar^2,0)
mat_d = J(ncluster,s*s,0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)
 wj = wj_star[i,]
 v = panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(v)

/*----------beta----------*/

 ej = yj - xj * beta

 t2j = ( xj' * diag * zj )
 t4j = ( zj' * diag * yj )
 t5j = ( zj' * diag * zj )
 t7j = ( xj' * diag * ej )
 t8j = ( ej' * diag * zj )

 aj= cholinv( t5j + theta[s,]*(cholinv(invvech(theta[1..(s-1),]))))

 cj = t7j - ( t2j * aj *  t8j')

 mat_c[i,] = (vec( (wj:^2) * (cj * cj') ))'

/*----------theta----------*/

 Rklj=rowshape(R[i,],s)
 Skj=rowshape(S[i,],s)

 Dkj =  (Skj-Rklj*theta)*(Skj-Rklj*theta)'
 mat_d[i,]=(vec(Dkj))'
}

  /*********************************************************/
 /*---------------------Variances------------------------*/
/*********************************************************/
var_beta = cholinv(s_matp)*((ncluster /( ncluster-1))*rowshape(colsum(mat_c),nvar))
*cholinv(s_matp)
dp_beta = sqrt(diagonal(var_beta))


var_theta = cholinv(r_mat)*ncluster/(ncluster-1)* rowshape(colsum(mat_d),s)* cholinv(r_mat)
dp_theta = sqrt(diagonal(var_theta))

finish= st_global("c(current_time)")
today1= st_global("c(current_date)")

z_star = beta:/dp_beta
z_star2= theta:/dp_theta

z_star_l= beta-abs(invnormal(0.025)):*dp_beta
z_star_u= beta+abs(invnormal(0.025)):*dp_beta
pz_star= 2:*(1:-normal(abs(z_star)))

z_star2_l= theta-abs(invnormal(0.05)):*dp_theta
z_star2_u= theta+abs(invnormal(0.05)):*dp_theta
pz_star2= 2:*(1:-normal(abs(z_star2)))

st_matrix("beta",beta)
st_matrix("var_beta",var_beta)

printf("\n")
printf("{hline 87}\n")
printf("            Probability Weighted Iterative Generalized Least Squares            \n")
printf("{hline 87}\n")
printf("General Information\n")
printf("\n")
printf("Response Variable = {txt}%19s \n", dep)
printf("Weight at Level 2 = {txt}%19s \n", _wj_rep)
printf("Weight at Level 1 = {txt}%19s \n", _wi_j)
printf("\n")
printf("Start running on %s at  %s\n",  today , start)
printf("Number of Iterations    = %3.0f\n", n_it)
printf("Number of Level 1 units = %3.0f\n", nsubjc_t)
printf("Number of Level 2 units = %3.0f\n", ncluster)
printf("\n")
printf("{hline 87}\n")
printf("       Fixed Effects{c |}  Coef.    Std.Err.    z     P>|z|   [95%sConf.Interval]
 Init.Val.\n","%")
printf("{hline 20}{c +}{hline 66}\n")
for (mi=1; mi<= nvar; mi++) {
printf("{txt}%19s {c |} {res}%8.0g  %8.0g  %6.2f  %6.3f  %8.0g  %8.0g  %8.0g\n",name1[.,mi]',
 beta[mi,.], dp_beta[mi,.],z_star[mi,.], pz_star[mi,.] , z_star_l[mi,.],z_star_u[mi,.],
 beta0[mi,.] )
}
printf("{hline 87}\n")
printf("\n")

printf("{hline 87}\n")
printf(" Variance Components{c |}  Coef.    Std.Err.    z     P>|z|   [95%sConf.Interval]
  Init.Val.\n","%")
printf("{hline 20}{c +}{hline 66}\n")
for (mi=1; mi<= s; mi++) {
printf("{txt}%19s {c |} {res}%8.0g  %8.0g  %6.2f  %6.3f  %8.0g  %8.0g  %8.0g\n",name3[.,mi]',
 theta[mi,.], dp_theta[mi,.],z_star2[mi,.], pz_star2[mi,.] , z_star2_l[mi,.],z_star2_u[mi,.],
 theta0[mi,.] )
}
printf("{hline 87}\n")
printf("Note: Robust Standard Errors \n")
printf("      Finish running on %s at  %s\n",  today1 , finish)
printf("{hline 87}\n")
}
mata mosave  pwigls_2l_adcv(), replace
end
```

D.2 Computer Code: Probability-weighted Iterative Generalized Least Squares for General Linear Lag-dependent Covariance Structure

```
/*****************************************************************************/
/* Probability Weighted Iterative Generalized Least Squares for two level      */
/* multivariate models with General Lag Linear dependent covariance structure  */
/*****************************************************************************/
version 9
mata:

void pwigls_genlin_adcv( string varlist, dep, string varlist2, string varlist1, cluster_var,
 _wj_rep, _wi_j )

{
start= st_global("c(current_time)")
today= st_global("c(current_date)")
x = st_data(., tokens(varlist))
y = st_data(.,tokens(dep))
z = st_data(.,tokens(varlist1))
time_var = st_data(.,tokens(varlist2))
cluster_var = st_data(.,tokens(cluster_var))
```

```
cluster= uniqrows(cluster_var)
wj_rep = st_data(.,tokens(_wj_rep))
wi_j = st_data(.,tokens(_wi_j))

nvar = cols(x)
ncluster = rows(cluster)
nsubjc_t=rows(y)

q = cols(z)
t = cols(time_var)
s = 12

/*-------------- H matrices --------------*/
h_matrix= ((I(s)[1, ]))'
h_matrix

/*-------------- Delta matrices --------------*/
delta_matrix=J(s,64,0)
delta_matrix[2,]=(1,0,0,0,0,0,0,0,
0,1,0,0,0,0,0,0,
0,0,1,0,0,0,0,0,
0,0,0,1,0,0,0,0,
0,0,0,0,1,0,0,0,
0,0,0,0,0,1,0,0,
0,0,0,0,0,0,1,0,
0,0,0,0,0,0,0,1)


delta_matrix[3,]=(1,1,0,0,0,0,0,0,
1,1,1,0,0,0,0,0,
0,1,1,1,0,0,0,0,
0,0,1,1,0,0,0,0,
0,0,0,0,1,1,0,0,
0,0,0,0,1,1,1,0,
0,0,0,0,0,1,1,1,
0,0,0,0,0,0,1,1)

delta_matrix[4,]=(1,0,1,0,0,0,0,0,
0,1,0,1,0,0,0,0,
1,0,1,0,0,0,0,0,
0,1,0,1,0,0,0,0,
0,0,0,0,1,0,1,0,
0,0,0,0,0,1,0,1,
0,0,0,0,1,0,1,0,
0,0,0,0,0,1,0,1)

delta_matrix[5,]=(1,0,0,1,0,0,0,0,
0,1,0,0,0,0,0,0,
0,0,1,0,0,0,0,0,
1,0,0,1,0,0,0,0,
0,0,0,0,1,0,0,1,
0,0,0,0,0,1,0,0,
0,0,0,0,0,0,1,0,
0,0,0,0,1,0,0,1)

delta_matrix[6,]=(1,0,0,0,0,0,0,0,
0,1,0,0,0,0,0,0,
0,0,1,0,0,0,0,0,
0,0,0,1,1,0,0,0,
0,0,0,1,1,0,0,0,
0,0,0,0,0,1,0,0,
0,0,0,0,0,0,1,0,
0,0,0,0,0,0,0,1)

delta_matrix[7,]=(1,0,0,0,0,0,0,0,
0,1,0,0,0,0,0,0,
0,0,1,0,1,0,0,0,
0,0,0,1,0,1,0,0,
0,0,1,0,1,0,0,0,
0,0,0,1,0,1,0,0,
0,0,0,0,0,0,1,0,
0,0,0,0,0,0,0,1)

delta_matrix[8,]=(1,0,0,0,0,0,0,0,
0,1,0,0,1,0,0,0,
0,0,1,0,0,1,0,0,
0,0,0,1,0,0,1,0,
0,1,0,0,1,0,0,0,
0,0,1,0,0,1,0,0,
0,0,0,1,0,0,1,0,
0,0,0,0,0,0,0,1)

delta_matrix[9,]=(1,0,0,0,1,0,0,0,
0,1,0,0,0,1,0,0,
0,0,1,0,0,0,1,0,
0,0,0,1,0,0,0,1,
1,0,0,0,1,0,0,0,
0,1,0,0,0,1,0,0,
0,0,1,0,0,0,1,0,
0,0,0,1,0,0,0,1)

delta_matrix[10,]=(1,0,0,0,0,1,0,0,
0,1,0,0,0,0,1,0,
0,0,1,0,0,0,0,1,
0,0,0,1,0,0,0,0,
0,0,0,0,1,0,0,0,
1,0,0,0,0,1,0,0,
0,1,0,0,0,0,1,0,
0,0,1,0,0,0,0,1)

delta_matrix[11,]=(1,0,0,0,0,0,1,0,
0,1,0,0,0,0,0,1,
0,0,1,0,0,0,0,0,
0,0,0,1,0,0,0,0,
0,0,0,0,1,0,0,0,
0,0,0,0,0,1,0,0,
1,0,0,0,0,0,1,0,
0,1,0,0,0,0,0,1)

delta_matrix[12,]=(1,0,0,0,0,0,0,1,
0,1,0,0,0,0,0,0,
0,0,1,0,0,0,0,0,
0,0,0,1,0,0,0,0,
0,0,0,0,1,0,0,0,
0,0,0,0,0,1,0,0,
0,0,0,0,0,0,1,0,
1,0,0,0,0,0,0,1)

rowshape(delta_matrix[1,],8)
rowshape(delta_matrix[2,],8)
rowshape(delta_matrix[3,],8)
rowshape(delta_matrix[4,],8)
rowshape(delta_matrix[5,],8)
rowshape(delta_matrix[6,],8)
rowshape(delta_matrix[7,],8)
rowshape(delta_matrix[8,],8)
rowshape(delta_matrix[9,],8)
rowshape(delta_matrix[10,],8)
rowshape(delta_matrix[11,],8)
rowshape(delta_matrix[12,],8)

name1 = tokens(varlist)
name3 = ("Sigma_u_2","Genlin(1)","Genlin(2)","Genlin(3)","Genlin(4)","Genlin(5)","Genlin(6)",
"Genlin(7)","Genlin(8)","Genlin(9)","Genlin(10)","Genlin(11)","Genlin(12)")

lambidaj = J(nsubjc_t,1,0)

info_j=panelsetup(cluster_var,1)
cluster_wgt=J(ncluster,1,0)

/***********************************************************/
/*-------------- Calculating the Scaled Weights --------------*/
```

```
/*-------------- Scaling Method 2               --------------*/
/***************************************************************/
for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 kwi_j=panelsubmatrix(wi_j, i, info_j)
 np = rows(nsubjc)
 k =mean(kwi_j)
 parte = J(np,1,k)

 a=info_j[i,1]
 b=info_j[i,2]
 lambidaj[a..b,]=parte

wj_rep_j=panelsubmatrix(wj_rep, i, info_j)

cluster_wgt[i,]=uniqrows(wj_rep_j)
}

wi_j_star = wi_j :/ lambidaj
wj_star = cluster_wgt / mean(cluster_wgt)
wjesc_r = wj_rep / mean(cluster_wgt)

  /*********************************************************************/
 /*-------------- Calculating Beta_zero and Theta_zero --------------*/
/********************************************************************/
mat_t1j = J(ncluster,nvar^2,0) // it is necessary to declare a matrix
mat_t3j = J(ncluster,nvar,0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 np = rows(nsubjc)

 wi_j_starj=panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(wi_j_starj)
 wj = wj_star[i,]

 mat_t1j[i,] = (vec( ( xj' * diag * xj ) * wj ))'
 mat_t3j[i,] = (vec( ( xj' * diag * yj ) * wj))'
}
somat1 = rowshape(colsum(mat_t1j),nvar)
somat3 = colsum(mat_t3j)
beta0 = invsym(somat1) * (somat3')

vec_t6 = J(ncluster,1,0)
vec_aux = J(ncluster,1,0)
theta0 = J(s,1,.5)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 np = rows(nsubjc)

 wi_j_starj=panelsubmatrix(wi_j_star, i, info_j)
 wj = wj_star[i,]

 resid = yj - xj * beta0

 uj0 = (wi_j_starj' * resid ) / colsum(wi_j_starj)

 vec_t6[i,] = (wi_j_starj' * ( (resid :- uj0) :^ 2 ) ) * wj
 vec_aux[i,]= wj * (colsum(wi_j_starj)- 1 )
}
 theta0[2,]=colsum(vec_t6)/colsum(vec_aux)

  /*************************************************/
 /*----------- ITERATIVE MODULE-------------------*/
/*************************************************/
matp = J(ncluster,nvar^2,0)
```

```
matq = J(ncluster,nvar,0)

beta_ant = beta0
beta = beta_ant:*2

theta_ant=theta0
theta=theta_ant :* 2

itera = 1

while (itera<=200 & (any(abs((theta-theta_ant)):> 0.000001) |
any(abs((beta-beta_ant)):> 0.000001) ))
{

/*------------- looping for beta -------------*/
for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

  np = rows(nsubjc)

 v=panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(v)

 wj = wj_star[i,]

 if (itera ~= 1) {
 theta_genlin=theta[2,]*rowshape(delta_matrix[2,],8)+ theta[3,]*rowshape(delta_matrix[3,],8)+
  theta[4,]*rowshape(delta_matrix[4,],8)+ theta[5,]*rowshape(delta_matrix[5,],8)+
 theta[6,]*rowshape(delta_matrix[6,],8)+theta[7,]*rowshape(delta_matrix[7,],8)
 +theta[8,]*rowshape(delta_matrix[8,],8)+theta[9,]*rowshape(delta_matrix[9,],8)
 +theta[10,]*rowshape(delta_matrix[10,],8)+theta[11,]*rowshape(delta_matrix[11,],8)
 +theta[12,]*rowshape(delta_matrix[12,],8)
 sigma=I(np/t)#theta_genlin
 aj= invsym(invsym(theta[1,])+zj'*(diag*invsym(sigma))*zj)
 invvj=diag*invsym(sigma)-diag*invsym(sigma)*zj*aj*zj'*diag*invsym(sigma)
 }
 else {
 theta0_genlin=theta0[2,]*rowshape(delta_matrix[2,],8)+ theta0[3,]*rowshape(delta_matrix[3,],8)
 +theta0[4,]*rowshape(delta_matrix[4,],8)+ theta0[5,]*rowshape(delta_matrix[5,],8)+
 theta0[6,]*rowshape(delta_matrix[6,],8)+theta0[7,]*rowshape(delta_matrix[7,],8)
 +theta0[8,]*rowshape(delta_matrix[8,],8)+theta0[9,]*rowshape(delta_matrix[9,],8)
 +theta0[10,]*rowshape(delta_matrix[10,],8)+theta0[11,]*rowshape(delta_matrix[11,],8)
 +theta0[12,]*rowshape(delta_matrix[12,],8)
 sigma0=I(np/t)#theta0_genlin
 aj=invsym(invsym(theta0[1,])+zj'*(diag*invsym(sigma0))*zj)
 invvj=diag*invsym(sigma0)-diag*invsym(sigma0)*zj*aj*zj'*diag*invsym(sigma0)
 }
 matp[i,] = (vec ( wj :* xj'*invvj*xj ))'
 matq[i,] = (vec ( wj :* xj'*invvj*yj ))'
}

/*---------- beta------------- */
s_matp = rowshape(colsum(matp),nvar)
s_matq = colsum(matq)

if (itera ~= 1){
beta_ant = beta
 }
beta = invsym(s_matp) * (s_matq')

/*------------- looping for theta=inv(R)x S -------------*/
q = cols(z)
s = 12
R = J(ncluster,s*s,0)
S = J(ncluster,s,0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)

 v=panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(v)
 wj = wj_star[i,]
 ej = yj - xj * beta

 Rklj = J(s,s,0)
 Skj = J(s,1,0)

 H = h_matrix
 delta=delta_matrix

 if (itera ~= 1) {
 theta_genlin=theta[2,]*rowshape(delta_matrix[2,],8)+ theta[3,]*rowshape(delta_matrix[3,],8)
 +theta[4,]*rowshape(delta_matrix[4,],8)+ theta[5,]*rowshape(delta_matrix[5,],8)
 +theta[6,]*rowshape(delta_matrix[6,],8)+theta[7,]*rowshape(delta_matrix[7,],8)
 +theta[8,]*rowshape(delta_matrix[8,],8)+theta[9,]*rowshape(delta_matrix[9,],8)
 +theta[10,]*rowshape(delta_matrix[10,],8)+theta[11,]*rowshape(delta_matrix[11,],8)
 +theta[12,]*rowshape(delta_matrix[12,],8)
 sigma=I(np/t)#theta_genlin
 aj= invsym(invsym(theta[1,])+zj'*(diag*invsym(sigma))*zj)
 invvj=diag*invsym(sigma)-diag*invsym(sigma)*zj*aj*zj'*diag*invsym(sigma)
 }
 else {
 theta0_genlin=theta0[2,]*rowshape(delta_matrix[2,],8)+ theta0[3,]*rowshape(delta_matrix[3,],8)
 +theta0[4,]*rowshape(delta_matrix[4,],8)+ theta0[5,]*rowshape(delta_matrix[5,],8)
 +theta0[6,]*rowshape(delta_matrix[6,],8)+theta0[7,]*rowshape(delta_matrix[7,],8)
 +theta0[8,]*rowshape(delta_matrix[8,],8)+theta0[9,]*rowshape(delta_matrix[9,],8)
 +theta0[10,]*rowshape(delta_matrix[10,],8)+theta0[11,]*rowshape(delta_matrix[11,],8)
 +theta0[12,]*rowshape(delta_matrix[12,],8)
 sigma0=I(np/t)#theta0_genlin
 aj=invsym(invsym(theta0[1,])+zj'*(diag*invsym(sigma0))*zj)
 invvj=diag*invsym(sigma0)-diag*invsym(sigma0)*zj*aj*zj'*diag*invsym(sigma0)
 }

for (k=1; k<=s ; k++) {
for (l=1; l<=s ; l++) {
Rklj[k,l]= wj*(trace((invvj*(zj*h_matrix[k,]*zj' + I(np/t)#rowshape(delta_matrix[k,],t)' *
invsym(diag))) * (invvj*(zj*h_matrix[l,]*zj' + I(np/t)#rowshape(delta_matrix[l,],t)'
*invsym(diag)))))
}
Skj[k,]= wj*(trace(invvj*(zj*h_matrix[k,]*zj' +
I(np/t)#rowshape(delta_matrix[k,],t)'*invsym(diag))*invvj*(ej*ej')))
 }
R[i,]=(vec(Rklj))'
S[i,]=(vec(Skj))'
}

 matr=colsum(R)
 mats=colsum(S)

 r_mat = rowshape(matr,s)
 s_mat = rowshape(mats,s)

if (itera ~= 1) {
 theta_ant = theta
}

theta = invsym(r_mat) * s_mat

itera = itera + 1
}

n_it = itera - 1
  /**************************************************/
 /*----------End of iterative process--------------*/
```

```
/**************************************************/

  /************************************************/
 /*------------------ Residuals------------------*/
/************************************************/
u = J(ncluster,q, 0)
var_u = J(ncluster,q*q,0)
dp_u = J(ncluster,q,0)

yhat = J(nsubjc_t,1, 0)

v = J(nsubjc_t,1, 0)
var_v = J(nsubjc_t,1, 0)

for (i=1; i<=rows(info_j) ; i++){

 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)

 ej = yj - xj * beta

 a=info_j[i,1]
 b=info_j[i,2]

 Sigmau=(invvech(theta[1,]))

 Rhj = Sigmau* zj'

 theta_genlin=theta[2,]*rowshape(delta_matrix[2,],8)+ theta[3,]*rowshape(delta_matrix[3,],8)+
 theta[4,]*rowshape(delta_matrix[4,],8)+ theta[5,]*rowshape(delta_matrix[5,],8)+
 theta[6,]*rowshape(delta_matrix[6,],8)+theta[7,]*rowshape(delta_matrix[7,],8)
 +theta[8,]*rowshape(delta_matrix[8,],8)+theta[9,]*rowshape(delta_matrix[9,],8)
 +theta[10,]*rowshape(delta_matrix[10,],8)+theta[11,]*rowshape(delta_matrix[11,],8)
 +theta[12,]*rowshape(delta_matrix[12,],8)
 sigma=I(np/t)#theta_genlin

 Vj = zj * Sigmau * zj' + sigma

 aux = Rhj * invsym(Vj)

 u[i,] = (aux * ej)'

 var_u[i,] = (vec(Sigmau - aux * Rhj'))'
 dp_u[i,] = (vec(sqrt(diagonal(rowshape(var_u[i,],q) ))))'

 yhat1 = xj * beta + zj*u[i,]'
 yhat[a..b,] = yhat1

 vj = ej - zj*u[i,]'
 v[a..b,] = vj
}

resindex01 = st_addvar("float","u")
st_store((1,rows(u)),resindex01,u)

resindex0 = st_addvar("float","se")
st_store((1,rows(dp_u)),resindex0,dp_u)

resindex = st_addvar("float","resid")
st_store((1,rows(res)),resindex,res)

resindex1 = st_addvar("float","yhat1")
st_store((1,rows(yhat)),resindex1,yhat)

resindex2 = st_addvar("float","clusteru")
st_store((1,rows(cluster)),resindex2,cluster)
```

```
/*********************************************************/
/*-------------- Variances of Beta and Theta--------------*/
/*********************************************************/
mat_c = J(ncluster,nvar^2,0)
mat_d = J(ncluster,s*s,0)

for (i=1; i<=rows(info_j) ; i++){
 nsubjc=panelsubmatrix(cluster_var, i, info_j)
 yj=panelsubmatrix(y, i, info_j)
 xj=panelsubmatrix(x, i, info_j)
 zj=panelsubmatrix(z, i, info_j)

 np = rows(nsubjc)
 wj = wj_star[i,]
 v = panelsubmatrix(wi_j_star, i, info_j)
 diag = diag(v)

/*----------- beta-------------- */
 ej = yj - xj * beta

 theta_genlin=theta[2,]*rowshape(delta_matrix[2,],8)+ theta[3,]*rowshape(delta_matrix[3,],8)+
 theta[4,]*rowshape(delta_matrix[4,],8)+ theta[5,]*rowshape(delta_matrix[5,],8)+
 theta[6,]*rowshape(delta_matrix[6,],8)+theta[7,]*rowshape(delta_matrix[7,],8)
 +theta[8,]*rowshape(delta_matrix[8,],8)+theta[9,]*rowshape(delta_matrix[9,],8)
 +theta[10,]*rowshape(delta_matrix[10,],8)+theta[11,]*rowshape(delta_matrix[11,],8)
 +theta[12,]*rowshape(delta_matrix[12,],8)
 sigma=I(np/t)#theta_genlin
 aj= invsym(invsym(theta[1,])+zj'*(diag*invsym(sigma))*zj)
 invvj=diag*invsym(sigma)-diag*invsym(sigma)*zj*aj*zj'*diag*invsym(sigma)

 cj = xj'*invvj*ej

 mat_c[i,] = (vec( (wj:^2) * (cj * cj') ))'

/*----------- theta-------------- */
 Rklj=rowshape(R[i,],s)
 Skj=rowshape(S[i,],s)

 Dkj = (Skj-Rklj*theta)*(Skj-Rklj*theta)'
 mat_d[i,]=(vec(Dkj))'
}

  /*********************************************************/
 /*----------------------Variances-------------------------*/
/*********************************************************/
var_beta = invsym(s_matp)*((ncluster /( ncluster-1))*rowshape(colsum(mat_c),nvar))*
invsym(s_matp)
dp_beta = sqrt(diagonal(var_beta))

var_theta = invsym(r_mat)*ncluster/(ncluster-1)* rowshape(colsum(mat_d),s)* invsym(r_mat)
dp_theta = sqrt(diagonal(var_theta))

finish= st_global("c(current_time)")
today1= st_global("c(current_date)")

z_star = beta:/dp_beta
z_star2= theta:/dp_theta

z_star_l= beta-abs(invnormal(0.025)):*dp_beta
z_star_u= beta+abs(invnormal(0.025)):*dp_beta
pz_star= 2:*(1:-normal(abs(z_star)))

z_star2_l= theta-abs(invnormal(0.05)):*dp_theta
z_star2_u= theta+abs(invnormal(0.05)):*dp_theta
pz_star2= 2:*(1:-normal(abs(z_star2)))

theta_genlin=theta[2,]*rowshape(delta_matrix[2,],8)+ theta[3,]*rowshape(delta_matrix[3,],8)+
 theta[4,]*rowshape(delta_matrix[4,],8)+ theta[5,]*rowshape(delta_matrix[5,],8)+
 theta[6,]*rowshape(delta_matrix[6,],8)+theta[7,]*rowshape(delta_matrix[7,],8)
+theta[8,]*rowshape(delta_matrix[8,],8)+theta[9,]*rowshape(delta_matrix[9,],8)
+theta[10,]*rowshape(delta_matrix[10,],8)+theta[11,]*rowshape(delta_matrix[11,],8)
```

```
+theta[12,]*rowshape(delta_matrix[12,],8)

TOEP=theta_genlin
Sigma_r=theta[1]*J(t,t,1)+ TOEP
st_matrix("beta",beta)
st_matrix("var_beta",var_beta)

  /********************************************************/
 /*--------------- Printing results -----------------------*/
/*********************************************************/
printf("\n")
printf("{hline 87}\n")
printf("      Probability Weighted Iterative Generalized Least Squares      \n")
printf("{hline 87}\n")
printf("General Information\n")
printf("\n")
printf("Response Variable = {txt}%19s \n", dep)
printf("Weight at Level 2 = {txt}%19s \n", _wj_rep)
printf("Weight at Level 1 = {txt}%19s \n", _wi_j)
printf("\n")
printf("Start running on %s at %s\n", today , start)
printf("Number of Iterations  = %3.0f\n", n_it)
printf("Number of Time points  = %3.0f\n", t)
printf("Number of Level 1 units = %3.0f\n", nsubjc_t/t)
printf("Number of Level 2 units = %3.0f\n", ncluster)
printf("\n")
printf("{hline 87}\n")
printf("    Fixed Effects{c |}  Coef.  Std.Err.  z   P>|z|  [95%sConf.Interval]
 Init.Val.\n","%")
printf("{hline 20}{c +}{hline 66}\n")
for (mi=1; mi<= nvar; mi++) {
printf("{txt}%19s {c |} {res}%8.0g %8.0g %6.2f %6.3f %8.0g %8.0g %8.0g\n",name1[.,mi]',
 beta[mi,.], dp_beta[mi,.],z_star[mi,.], pz_star[mi,.] , z_star_l[mi,.],z_star_u[mi,.]
 ,beta0[mi,.] )
}
printf("{hline 87}\n")
printf("\n")
printf("{hline 87}\n")
printf(" Variance Components{c |}  Coef.  Std.Err.  z   P>|z|  [95%sConf.Interval]
 Init.Val.\n","%")
printf("{hline 20}{c +}{hline 66}\n")
for (mi=1; mi<= s; mi++) {
printf("{txt}%19s {c |} {res}%8.0g %8.0g %6.2f %6.3f %8.0g %8.0g %8.0g\n",name3[.,mi]',
 theta[mi,.], dp_theta[mi,.],z_star2[mi,.], pz_star2[mi,.] , z_star2_l[mi,.],z_star2_u[mi,.],
 theta0[mi,.] )
}
printf("{hline 87}\n")
printf("\n\n General Linear Matrix\n")
TOEP
printf("\n\n Total Variance\n")
Sigma_r
printf("{hline 87}\n")
printf("Note: Robust Standard Errors \n")
printf("   Finish running on %s at %s\n", today1 , finish)
printf("{hline 87}\n")
}
bb2=0
mata mosave pwigls_genlin_adcv(), replace
end
```

D.3 Output for the Final Multivariate Multilevel Model Estimated via PWIGLS

```
---------------------------------------------------------------------------------------
            Probability Weighted Iterative Generalized Least Squares
---------------------------------------------------------------------------------------
General Information

Response Variable =         log_incomeR
Weight at Level 2 =                  wj
Weight at Level 1 =                wi_j

Start running on  3 Jun 2009 at  23:03:47
Number of Iterations    =   7
Number of Time points   =   8
Number of Level 1 units = 6524
Number of Level 2 units = 1762
---------------------------------------------------------------------------------------
       Fixed Effects|   Coef.    Std.Err.     z     P>|z|   [95%Conf.Interval]  Init.Val.
--------------------+------------------------------------------------------------------
           wav015_1 |   6.01923   .292544   20.58   0.000   5.44585   6.59261   5.03989
           wav015_2 |   6.02227    .29244   20.59   0.000    5.4491   6.59544   5.04101
           wav015_3 |   6.01442   .292825   20.54   0.000   5.44049   6.58835   5.03244
           wav015_4 |   6.02598   .292496   20.60   0.000   5.45269   6.59926   5.04339
           wav015_5 |   6.05161   .292528   20.69   0.000   5.47826   6.62495   5.06866
           wav015_6 |    6.0475     .2924   20.68   0.000   5.47441   6.62059   5.06172
           wav015_7 |   6.04417   .292575   20.66   0.000   5.47073   6.61761   5.06012
           wav015_8 |   6.05173   .292677   20.68   0.000   5.47809   6.62536   5.06692
              RM_29 |   .012752   .039629    0.32   0.748   -.06492   .090424   .026273
              RM_31 |   .253822   .031615    8.03   0.000   .191859   .315786   .276608
              RM_33 |   .216725   .029061    7.46   0.000   .159765   .273684   .247782
              RM_35 |    .36807   .031863   11.55   0.000   .305621    .43052   .389173
              RM_43 |   .239461   .035441    6.76   0.000   .169997   .308925   .272588
             age_w1 |   .002062   .001767    1.17   0.243    -.0014   .005525  -.000731
          age_w1_sq |  -.000474    .00007   -6.77   0.000  -.000611  -.000337  -.000409
               male |   .529592   .164284    3.22   0.001   .207602   .851582    .61145
           white_w1 |   .216118   .039241    5.51   0.000   .139206   .293029   .170516
           informal |  -.095638    .03212   -2.98   0.003  -.158592  -.032685  -.386507
             formal |    -.0182   .031743   -0.57   0.566  -.080415   .044015  -.230513
          emp_other |   .041726   .037099    1.12   0.261  -.030986   .114438  -.002263
           self_emp |  -.169009   .026107   -6.47   0.000  -.220178  -.117841  -.522254
         int_vd20_2 |  -.015092   .015988   -0.94   0.345  -.046428   .016244  -.051027
         int_vd20_3 |   -.04361   .010839   -4.02   0.000  -.064853  -.022366  -.130819
         int_vd20_4 |  -.008754   .012194   -0.72   0.473  -.032653   .015145  -.033797
         int_vd20_5 |   .028476   .016354    1.74   0.082  -.003578    .06053   -.10124
         int_vd20_6 |  -.033176   .022618   -1.47   0.142  -.077506   .011154  -.161429
         int_vd20_7 |    -.0346    .01423   -2.43   0.015   -.06249   -.00671  -.081988
         int_vd20_8 |   -.01077   .023555   -0.46   0.648  -.056937   .035397  -.307713
           educa_w1 |  -.083486   .014133   -5.91   0.000  -.111186  -.055786  -.091485
        educa_w1_sq |   .009575   .000819   11.69   0.000    .00797    .01118   .009543
          dur_emp_c |   .000415   .000071    5.83   0.000   .000275   .000554   .001602
       dur_emp_c_sq |  -9.6e-07   2.7e-07   -3.63   0.000  -1.5e-06  -4.4e-07  -2.5e-06
          log_hours |   .257011   .035692    7.20   0.000   .187057   .326966   .471453
        entrevist_2 |   .008961    .00773    1.16   0.246   -.00619   .024112   .004061
    int_malXage_w_1 |   .005696    .00194    2.94   0.003   .001895   .009498   .004826
    int_malXeduca_1 |   .075799   .016174    4.69   0.000   .044098     .1075   .069797
    int_malXeducaa1 |  -.004585   .000953   -4.81   0.000  -.006453  -.002716  -.003996
    int_malXlog_h_1 |  -.087637   .041216   -2.13   0.033  -.168418  -.006856  -.137124
    int_malXent_1_2 |  -.039626   .009715   -4.08   0.000  -.058668  -.020585  -.002829
    int_whiXinfor_1 |  -.086287   .040238   -2.14   0.032  -.165152  -.007422  -.084352
    int_whiXforma_1 |  -.114203   .042926   -2.66   0.008  -.198336   -.03007  -.067152
    int_whiXemp_o_1 |  -.180046   .050087   -3.59   0.000  -.278214  -.081878  -.242086
    int_whiXself__1 |  -.013537   .033306   -0.41   0.684  -.078815   .051742   .044826
           CEformal |  -1.69013    .23689   -7.13   0.000  -2.15442  -1.22583  -1.20669
         CEinformal |  -1.97866   .267583   -7.39   0.000  -2.50311  -1.45421  -1.31177
        CEemp_other |  -2.08551   .299279   -6.97   0.000  -2.67209  -1.49893  -1.73421
         CEself_emp |  -1.96288   .281953   -6.96   0.000   -2.5155  -1.41026  -1.28446
            CEeduca |   .085814   .008067   10.64   0.000   .070002   .101625     .0933
---------------------------------------------------------------------------------------
```

```
---------------------------------------------------------------------------------------
 Variance Components|    Coef.    Std.Err.     z      P>|z|    [95%Conf.Interval]  Init.Val.
--------------------+------------------------------------------------------------------
          Sigma_u_2 |  .009565     .00318    3.01    0.003    .004334   .014795        .5
          Genlin(1) | -2.04916    .071627  -28.61    0.000   -2.16698  -1.93135   .238071
          Genlin(2) |  .266337    .008247   32.30    0.000    .252772   .279901        .5
          Genlin(3) |  .261529    .008083   32.36    0.000    .248234   .274824        .5
          Genlin(4) |  .257416    .008075   31.88    0.000    .244134   .270698        .5
          Genlin(5) |  .229741    .008488   27.07    0.000    .215779   .243702        .5
          Genlin(6) |     .229    .008283   27.65    0.000    .215376   .242624        .5
          Genlin(7) |  .226347    .008102   27.94    0.000     .21302   .239673        .5
          Genlin(8) |  .224365    .008007   28.02    0.000    .211194   .237535        .5
          Genlin(9) |  .222241    .007945   27.97    0.000    .209173   .235309        .5
         Genlin(10) |  .219888    .007872   27.93    0.000    .206939   .232836        .5
         Genlin(11) |  .217683    .007904   27.54    0.000    .204682   .230684        .5
---------------------------------------------------------------------------------------

 General Linear Matrix
[symmetric]
                1             2             3             4             5
    +-----------------------------------------------------------------------
  1 |  .3053804604
  2 |  .2663367367   .3053804604
  3 |  .2615288685   .2663367367   .3053804604
  4 |  .2574159714   .2615288685   .2663367367   .3053804604
  5 |  .2243645655   .2263465302    .228999806   .2297405771   .3053804604
  6 |  .2222409588   .2243645655   .2263465302    .228999806   .2663367367
  7 |  .2198878267   .2222409588   .2243645655   .2263465302   .2615288685
  8 |  .2176827283   .2198878267   .2222409588   .2243645655   .2574159714
    +-----------------------------------------------------------------------
                6             7             8
     -------------------------------------------+
  6    .3053804604                              |
  7    .2663367367   .3053804604                |
  8    .2615288685   .2663367367   .3053804604  |
     -------------------------------------------+

 Total Variance
[symmetric]
                1             2             3             4             5
    +-----------------------------------------------------------------------
  1 |    .31494527
  2 |  .2759015462     .31494527
  3 |  .2710936781   .2759015462     .31494527
  4 |   .266980781   .2710936781   .2759015462     .31494527
  5 |   .233929375   .2359113397   .2385646155   .2393053866     .31494527
  6 |  .2318057683    .233929375   .2359113397   .2385646155   .2759015462
  7 |  .2294526362   .2318057683    .233929375   .2359113397   .2710936781
  8 |  .2272475379   .2294526362   .2318057683    .233929375    .266980781
    +-----------------------------------------------------------------------
                6             7             8
     -------------------------------------------+
  6      .31494527                              |
  7    .2759015462     .31494527                |
  8    .2710936781   .2759015462     .31494527  |
     -------------------------------------------+
---------------------------------------------------------------------------------------
Note: Robust Standard Errors
      Finish running on  3 Jun 2009 at  23:47:54
---------------------------------------------------------------------------------------
```

Glossary

$\boldsymbol{\beta}$	the vector of all p fixed regression coefficients.
$\boldsymbol{r}_j$	the vector of composite residuals for cluster j.
$\boldsymbol{x}_{ij}$	the vector of all p explanatory variables.
$\boldsymbol{Y}_j$	the vector with the response variable.
$\boldsymbol{z}_{ij}$	the vector of explanatory variables considered as random.
$\hat{\Sigma}_r$	the estimated block diagonal matrix of the total covariance of the observations.
$\otimes$	the Kronecker product.
π_j	the selection probability for cluster j.
ρ	the intra-cluster correlation coefficient.
σ_e^2	the within level two units variance.
σ_u^2	the between level two units variance.
σ_v^2	the between level three units variance.
σ_{u01}	the covariance between the random intercepts and the random slopes.
e_{ij}	the raw residuals (level one).
i	the individual level subscript.
j	the cluster level subscript.
m	the total number of observations: $\sum_j n_j$.
n	the number of clusters.
n_j	the number of individuals within clusters.
p	the total number of covariates.
$p_{(1)}$	the level one covariates.
$p_{(2)}$	the level two covariates.
$p_{ij((t-1),t)}$	the predicted probabilities of the model using data of occasion $t-1$ and response indicator for occasion t.
p_{jh}	the PSU selection probability.
t	the occasion level subscript.
T_{ij}	the number of measurements occasions.
u_j	the cluster specific effects.

V	the variance matrix of the composite residuals.
$w^*_{ij(1)}$	the base weights.
$w^*_{ij(8)}$	the longitudinal weight to be used in the analysis of the data set including data from the first to the eighth occasion.
$w^*_{ij(t)}$	the adjusted weights for occasion t that compensates for the panel drop-out.
w_j	the inverse of π_j - the level two weights.
w^*_{ijh}	the PME sampling weights corrected for unit non-response.
w^*_{jh}	the PSU level weights.
log-income	Logarithmic of the Real Labour Income.
occasions	Measurement occasions representing the interview number.
wave	Interview time, also referred to as the time variable.
AR(1)	First Order Autoregressive.
CPS	Current Population Survey.
GEE	Generalized Estimation Equation.
GLS	Generalized Least Squares.
HH	Households
HoHH	Heads of Household.
IBGE	Instituto Brasileiro de Geografia e Estatítica.
IGLS	Iterative Generalized Least Squares.
ILO	International Labour Organization.
LFS	Labour Force Survey.
LRT	Likelihood Ratio Test.
ML	Maximum Likelihood.
OLS	Ordinary Least Squares.
PME	Pesquisa Mensal de Emprego.
PML	Pseudo-maximum likelihood.
PNAD	Pesquisa Nacional por Amostra de Domicílios.
PPS	Probability Proportional to Size Sampling.
PSU	Primary Sampling Units.
PWIGLS	Probability-weighted Iterative Generalized Least Squares.
R$	Brazilian currency *Real* (*Reais* in the plural).
REML	Restricted maximum likelihood.
RIGLS	Restricted Iterative Generalized Least Squares.
SE	Standard error of the coefficients estimated.
SSU	Secondary Sampling Units.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A*, 144(4):419–461.

Allen, T. M. and Petroni, R. J. (1994). Mover nonresponse adjustment research for the survey of income and program participation. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 662–667.

Antonaci, G. d. A. and Silva, D. B. d. N. (2007). Analysis of alternative rotation patterns for the Brazilian system of integrated household surveys. In *Proceedings of the 56th Session of the International Statistical Institute (ISI)*.

Arbache, J. S. (2001). Wage differentials in Brazil: theory and evidence. *The Journal of Development Studies*, 38(2):109–130.

Asparouhov, T. and Muthen, B. (2006). Multilevel modeling of complex survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 2718–2725.

Azzoni, C. R. and Servo, L. M. (2002). Education, cost of living and regional wage inequality in Brazil. *Papers in Regional Science*, 81(2):157–175.

Bailar, B. A. (1989). Information needs, surveys and measurement errors. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M., editors, *Panel Surveys*, chapter 1. J.W. Wiley and Sons, New York.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*. John Wiley and Sons Ltd, Chichester, third edition.

Barbosa, M. F. and Goldstein, H. (2000). Discrete response multilevel models for repeated measures: An application to voting intentions data. *Quality & Quantity*, 34(3):323–330.

Barros, R. P. d., Carvalho, M. d., Franco, S., and Mendonça, R. (2006). Uma análise das principais causas da queda recente na desigualdade de renda Brasileira. Texto Para Discussão 1203, Instituto de Pesquisa Econômica Aplicada. In Portuguese.

Barros, R. P. d., Corseuil, C. H., and Leite, P. G. (2000). Labor market and poverty in Brazil. Texto Para Discussão 723, Instituto de Pesquisa Econômica Aplicada.

Barros, R. P. d. and Mendonça, R. S. d. (1995). Os determinantes da desigualdade no Brasil. Texto para discussão, Instituto de Pesquisa Econômica Aplicada. In Portuguese.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods (Advanced Quantitative Techniques in the Social Sciences)*. Sage Publications, Inc.

Bureau of Labour Statistics (2002). *Technical Paper 63RV: Current Population Survey. Design and Methodology.*

Burkhauser, R. V. and Smeeding, T. M. (2000). Microdata panel data and public policy: National and cross-national perspectives. Center for Policy Research Working Papers 23, Center for Policy Research, Maxwell School, Syracuse University. `http://ideas.repec.org/p/max/cprwps/23.html`.

Chantala, K., Suchindran, C., and Blanchette, D. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 2815–2824.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley, New York, third edition.

Coelho, A. M. and Corseuil, C. H. (2002). Diferenciais salariais no Brasil: um breve panorama. In Corseuil, C. H., editor, *Estrutura Salarial: Aspectos Conceituais e Novos Resultados Para o Brasil*. Instituto de Pesquisa Econômica Aplicada, Rio de Janeiro. In Portuguese.

Corrêa, S. T. (2001). Modelos lineares hierárquicos em pesquisas por amostragem - relacionando o índice de massa corporal às variáveis da pesquisa sobre os padrões de vida. Master's thesis, Escola Nacional de Ciências Estatísticas. In Portuguese.

Corseuil, C. H. and Santos, D. D. d. (2002). Fatores que determinam o nível salarial no setor formal Brasileiro. In Corseuil, C. H., editor, *Estrutura Salarial: Aspectos Conceituais e Novos Resultados Para o Brasil*. Instituto de Pesquisa Econômica Aplicada, Rio de Janeiro. In Portuguese.

Diggle, P., Farewell, D., and Henderson, R. (2007). Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Journal of the Royal Statistical Society. Series C*, 56(5):499–550.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *The Analysis of Longitudinal Data*. Oxford Statistical Science. Oxford University Press, Oxford, second edition.

Dougherty, C. (2002). *Introduction to econometrics*. Oxford University Press, Oxford.

Duncan, G. (2000). Using panel studies to understand household behaviour and well-being. In Rose, D., editor, *Researching Social and Economic Change: the uses of household panel studies*, chapter 3. Routledge, London.

Efron, B. and Morris, C. (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.

Feder, M., Nathan, G., and Pfeffermann, D. (2000). Multilevel modelling of complex survey longitudinal data with time varying random effects. *Survey Methodology*, 26(1):53–65.

Fernandes, R. (2002). Desigualdade salarial: Aspectos teóricos. In Corseuil, C. H., editor, *Estrutura Salarial: Aspectos Conceituais e Novos Resultados Para o Brasil*. Instituto de Pesquisa Econômica Aplicada, Rio de Janeiro. In Portuguese.

Ferrao, M. E. (2002). Modelo multinível de resposta discreta para dados longitudinais: Uma aplicação aos dados da Pesquisa Mensal de Emprego. In *IX Congresso da Sociedade Portuguesa de Estatística*, Novos Rumos em Estatística - Actas do IX Congresso Anual da SPE, page 18.

Ferreira, F. H. (2000). Os determinantes da desigualdade de renda no Brasil: Luta de classes ou heterogeneidade educacional? Texto para discussão, Departamento de Economia, PUC-Rio.

Ferreira, F. H. G. and Barros, R. P. d. (1999). The slippery slope: Explaining the increase in extreme poverty in urban Brazil, 1976 - 96. *Brazilian Review of Econometrics*, 19(2).

Ferreira, F. H. G., Leite, P. G., and Litchfield, J. A. (2006). The rise and fall of Brazilian inequality, 1981-2004. World Bank Policy Research Working Paper 3867, World Bank.

Firebaugh, G. (1997). *Analyzing repeated surveys*. Thousand Oaks, Sage Publications, California.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York.

Folsom, R. E. (1989). A probability sampling perspective on panel data analysis. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M., editors, *Panel Surveys*, pages 108–138. J.W. Wiley and Sons, New York.

Fraine, B. D., Landeghem, G. V., Damme, J. V., and Onghena, P. (2005). An analysis of wellbeing in secondary school with multilevel growth curve models and multilevel multivariate models. *Quality and Quantity*, 39(3):297–316.

Freedman, D. A. (2006). On the so-called "huber-sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences.* Cambridge University Press, New York.

Fuller, W. and An, A. (1996). Adjustments for nonresponse in longitudinal surveys. In *Proceedings of the Survey Methods Section*, SSC Annual Meeting, pages 51–58.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1):43–56.

Goldstein, H. (2003). *Multilevel statistical models.* Kendall's Library of Statistics; 3. Arnold, London, 3rd edition.

Goldstein, H., Healy, M., and Rasbash, J. (1994). Multilevel time-series models with applications to repeated-measures data. *Statistics in Medicine*, 13(16):1643–1655.

Gonzaga, G., Filho, N. M., and Terra, M. C. (2005). Trade liberalization and the evolution of skill earnings differentials in Brazil. *Journal of International Economics*, 68:345–367.

Gouskova, E., Heeringa, S. G., McGonagle, K., and Schoeni, R. F. (2008). Panel study of income dynamics revised longitudinal weights 1993-2005. Technical report, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Griffiths, P. L., Brown, J. J., and Smith, P. W. F. (2004). A comparison of univariate and multivariate multilevel models for repeated measures of use of antenatal care in Uttar Pradesh. *Journal of the Royal Statistical Society. Series A*, 167(4):597–611.

Grilli, L. and Rampichini, C. (2006). A review of random effects modelling using gllamm in stata. Software reviews of multilevel analysis packages, Centre for Multilevel Modelling, University of Bristol. `http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/gllamm.shtml`.

Hanushek, E. A. (2006). Alternative school policies and the benefits of general cognitive skills. *Economics of Education Review*, 25(4):447–462.

Hawkes, D. and Plewis, I. (2006). Modelling non-response in the national child development study. *Journal of The Royal Statistical Society. Series A*, 127(3):479–491.

Hay, D. A. (2001). The post 1990 Brazilian trade liberalisation and the performance of large manufacturing firms: Productivity, market share and profits. *Economic Journal*, 111(473):20–41.

Heckman, J. J. and Singer, B. (1985). *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, Cambridge.

Hox, J. J. (2000). Multilevel analyses of grouped and longitudinal data. In Little, T. D., Schnabel, K. U., and Baumert, J., editors, *Modelling Longitudinal Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples*. Lawrence Erlbaum associates.

Hsiao, C. (2003). *Analysis of panel data*. Econometric Society Monographs (No. 34). Cambridge University Press, Cambridge, second edition.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In L. LeCam, J. N., editor, *Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability*, pages 221–233, Berkeley. University of California Press.

Hunter, L., Michaud, S., and Torrance, V. (1992). Modelling for non-response in a longitudinal survey. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 351–356.

IBGE (2001a). Notas metodológicas. Technical report, Instituto Brasileiro de Geofrafia e Estatística. http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova/pmemet1.pdf.

IBGE (2001b). Relatório sobre a transição metodológica. Technical report, Instituto Brasileiro de Geofrafia e Estatística. http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova/transmetod.pdf.

IBGE (2002). Pesquisa Mensal de Emprego. Série relatórios metodológicos. Technical report, Instituto Brasileiro de Geofrafia e Estatística. http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova/srmpme_2ed.pdf.

IBGE (2003). Estimação de intervalos de confiança para estimadores de diferenças temporais na Pesquisa Mensal de Emprego. Technical report, Instituto Brasileiro de Geofrafia e Estatística. http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova/Documentacao.pdf.

IBGE (2006a). O trabalho a partir dos 50 anos de idade. Pesquisa Mensal de Emprego (Recife, Salvador, Belo horizonte, Rio de Janeiro, São Paulo e Porto Alegre). Indicadores IBGE, Instituto Brasileiro de Geofrafia e Estatística. In Portuguese.

IBGE (2006b). O trabalho da mulher principal responsável no domicílio (Pesquisa Mensal de Emprego). Indicadores IBGE, Instituto Brasileiro de Geofrafia e Estatística. In Portuguese.

IBGE (2006c). Perfil dos trabalhadores domésticos nas seis regiões metropolitanas investigadas pela Pesquisa Mensal de Emprego (Recife, Salvador, Belo horizonte, Rio de Janeiro, São Paulo e Porto Alegre). Indicadores IBGE, Instituto Brasileiro de Geofrafia e Estatística. In Portuguese.

IBGE (2009). Principais destaques da evolução do mercado de trabalho nas regiões metropolitanas abrangidas pela pesquisa. (Recife, Salvador, Belo horizonte, Rio de Janeiro, São Paulo e Porto Alegre, 2003-2008). Indicadores IBGE, Instituto Brasileiro de Geofrafia e Estatística. In Portuguese.

IPEA (2005). Análise do mercado de trabalho. Technical report, Instituto de Pesquisa Econômica Aplicada. `http://www.ipea.gov.br/pub/bcmt/mt_28e.pdf`.

Jenkins, S. P. (2000). Modelling household income dynamics. *Journal of Population Economics*, 13(4):529 – 567.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820.

Kalton, G. and Bryk, M. (2000). Weighting in household panel surveys. In Rose, D., editor, *Researching Social and Economic Change: the uses of household panel studies*, chapter 5. Routledge, London.

Khattree, R. and Naik, D. N. (1999). *Applied Multivariate Statistics with SAS Software*. SAS Institute, Cary, NC, USA, second edition.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society. Series B*, 36(1):1–22.

Kobilarcik, E. L. and Singh, R. P. (1996). SIPP: Longitudinal estimation for persons' characteristics. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 214–219.

LaRoche, S. (2003). Longitudinal and cross-sectional weighting of the survey of labour and income dynamics. *Income Research Paper Series*. Households Survey Methods Division.

LaVange, L. M., Koch, G. G., and Schwartz, T. A. (2001). Applying sample survey methods to clinical trials data. *Statistics In Medicine*, 20(17-18):2609–2623.

Lee, E. S. and Forthofer, R. N. (2005). *Analyzing Complex Survey Data*. Number 71 in Quantitative Applications in the Social Sciences. Sage Publications Pvt. Ltd.

Lemieux, T. (2006). The mincer equation thirty years after schooling experience, and earnings. In *Jacob Mincer A Pioneer of Modern Labor Economics*, chapter 11. Springer US.

Lemos, S. (2002). The effects of the minimum wage on wages and employment in Brazil a menu of minimum wage variables. Discussion Papers in Economics 02-02, University College London.

Lemos, S. (2006). Anticipated effects of the minimum wage on prices. *Applied Economics*, 38(3):325–337.

Lepkowski, J. M. (1989). The treatment of wave nonresponse in panel surveys. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M., editors, *Panel Surveys*, chapter 5. J.W. Wiley and Sons, New York.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data-analysis using generalized linear-models. *Biometrika*, 73(1):13–22.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of The American Statistical Association*, 90(431):1112–1121.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data.* Wiley, Hoboken, N.J.

Longford, N. T. (1993). *Random Coefficient Models.* Clarendon Press, Oxford.

Luke, D. A. (2004). *Multilevel Modeling.* Thousand Oaks, Sage Publications, California.

Maas, C. J. M. and Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, 46(3):427–440.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Madrian, B. C. and Lefgren, L. J. (1999). A note on longitudinally matching current population survey (CPS) respondents. Technical working paper series, National Bureau of Economic Research.

McLaren, C. and Steel, D. (2000). The impact of different rotation patterns on the sampling variance of trend estimates. *Survey Methodology*, 26(2):163–172.

Mehran, F. (2007). Longitudinal analysis of employment and unemployment based on matched rotation samples. *Labour - Review of Labour Economics and Industrial Relations*, 3(1):3–20.

Menard, S. W. (2002). *Longitudinal research.* Sage university papers series. Quantitative applications in the social sciences; no. 07-76. Sage Publications, Thousand Oaks, California, second edition.

Menezes-Filho, N. (2002). Equações de rendimentos: Questões. In Corseuil, C. H., editor, *Estrutura Salarial: Aspectos Conceituais e Novos Resultados Para o Brasil.* Instituto de Pesquisa Econômica Aplicada, Rio de Janeiro. In Portuguese.

Menezes-Filho, N., Fernandes, R., and Picchetti, P. (2000). A evolução da distribuiição de salários no Brasil: Fatos estilizados para as décadas de 80 e 90. In Henrriques, R., editor, *Desigualdade e Pobreza no Brasil.* Instituto de Pesquisa Econômica Aplicada, Rio de Janeiro. In Portuguese.

Mincer, J. and Polachek, S. (1974). Family investment in human capital: Earnings of women. *Journal of Political Economy*, 82(2):S76–S108.

Nascimento, M. A. d. and Souza, A. P. F. d. (2005). Medidas e determinantes da mobilidade dos rendimentos do trabalho no Brasil. *Anais do XXXIII Encontro Nacional de Economia.*

O.E.C.D. (2006). Economic survey of Brazil, 2006. Policy brief, Organization For Economic Co-Operation and Development. `http://www.oecd.org/dataoecd/47/47/37667205.pdf`.

ONS, Office for National Statistics (2009). Guidance on the use of labour force survey microdata pending full reweighting following the 2001 census.

Passos, A. F. d., Ansiliero, G., and Paiva, L. H. (2005). Mercado de trabalho evolução recente e perspectivas. Boletim do mercado de trabalho - notas técnicas, Instituto de Pesquisa Econômica Aplicada. In Portuguese.

Pessoa, D. and Silva, P. (1998). *Análise de Dados Amostrais Complexos.* Associação Brasileira de Estatística (ABE), São Paulo, Brasil. In Portuguese.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337.

Pfeffermann, D. and La Vange, L. (1989). Regression models for stratified multi-stage cluster samples. In Skinner, C., Holt, D., and Smith, T., editors, *Analysis of Complex Surveys*, chapter 12. Wiley, Chichester.

Pfeffermann, D. and Sikov, A. (2008). Estimation and imputation under non-ignorable nonresponse with missing covariate information. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 90–101.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B*, 60(1):23–40.

Plewis, I. (1985). *Analysing change: Measurement and explanation using longitudinal data.* Wiley, Chichester.

Plewis, I. (2005). Modelling behaviour with multivariate multilevel growth curves. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(2):71–80.

Plewis, I. and Fielding, A. (2003). What is multilevel modelling for? a critical response to Gorard. *British Journal of Educational Studies*, 51(4):408–418.

Pourahmadi, M. (2007). Graphical diagnostics for modeling unstructured covariance matrices. *International Statistical Review*, 70(3):395–417.

Rabe-Hesketh, S. and Skrondal, A. (2005). *Multilevel and Longitudinal Modeling using Stata*. Stata Press, College Station, Tex.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society. Series A*, 127(4):805–827.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). *GLLAMM Manual*. U.C. Berkeley, division of biostatistics working paper series 160 edition.

Ramos, L. (2007). O desempenho recente do mercado de trabalho Brasileiro: tendências, fatos estilizados e padrões espaciais. Discussion Papers 1255, Instituto de Pesquisa Econômica Aplicada. `http://ideas.repec.org/p/ipe/ipetds/1255.html`.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., and Lewis, T. (2001). *A User's Guide to MlwiN*. London, England.

Rizzo, L., Kalton, G., and Brick, J. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22(2):43–53.

Rochon, J. and Helms, R. W. (1989). Maximum likelihood estimation for incomplete repeated-measures experiments under an arma covariance structure. *Biometrics*, 45(1):207–218.

Rowe, G. and Nguyen, H. (2004). Longitudinal analysis of labour force survey data. *Survey Methodology*, 330(1):4–13.

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: Changes in methadone treatment practices. *Journal of the American Statistical Association*, 97(457):40–52.

Royall, R. M. (1986). Model Robust Confidence Intervals Using Maximum Likelihood Estimators. *International Statistical Review*, 54(2):221–226.

SAS Institute Inc,Version 8 (1999). Sas onlinedoc®. Online manual, SAS Institute Inc., Cary, NC.

Schabenberger, O. and Pierce, F. J. (2001). *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press, Boca Raton, FL.

Schwartzman, S. (1999). Fora de foco: Diversidade e identidades Étnicas no Brasil. *Novos Estudos Cebrap*, 55:83–96. In Portuguese.

Scully, G. W. (1981). Interstate wage differentials: A cross section analysis. *The Journal of Human Resources*, 16(2):238–259.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley and Sons Ltd, Ithaca, New York.

Sedlacek, G., Barros, R. P. d., and Varandas, S. (1989). Segmentação e mobilidade no mercado de trabalho: a carteira de trabalho em São Paulo. *Pesquisa e Planejamento Econômico*, 20(1):87 – 103. In Portuguese.

Silva, P. L. N. and Moura, F. A. S. (1988). Redução da amostra da Pesquisa Mensal de Emprego: Estratégia para reduzir o custo da pesquisa. *Revista Brasileira de Estatística*, 49(192):65–95.

Silveira-Neto, R. and Azzoni, C. R. (2006). Location and regional income disparity dynamics: The Brazilian case. *Papers in Regional Science*, 85(4):599–613.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23(4):323–355.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Occurrence*. Oxford University Press, New York.

Skinner, C. (1986). Design effects of two-stage sampling. *Journal of the Royal Statistical Society. Series B*, 48(1):89–99.

Skinner, C. (1989a). Domain means, regression and multivariate analysis. In Skinner, C., Holt, D., and Smith, T., editors, *Analysis of Complex Surveys*, chapter 3. Wiley, Chichester.

Skinner, C. (1989b). Introduction to part A. In Skinner, C., Holt, D., and Smith, T., editors, *Analysis of Complex Surveys*, chapter 2. Wiley, Chichester.

Skinner, C. (2003). Introduction part D. In Chambers, R. and Skinner, C., editors, *Analysis of Survey Data*, chapter 13. Wiley, Chichester.

Skinner, C. and Holmes, D. (2003). Random effects models for longitudinal survey data. In Chambers, R. and Skinner, C., editors, *Analysis of Survey Data*. Wiley, Chichester.

Skinner, C. and Vieira, M. (2007). Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology*, 33(1):3–12.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Interdisciplinary statistics. Chapman and Hall/Crc, Boca Raton, Fla.

Smith, J. and Vinhosa, F. (2002). *History of Brazil, 1500 - 2000. Politics, economy, society, diplomacy.* Pearson Education Limited 2002, London.

Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications Ltd, Thousand Oaks, CA.

Solon, G. (1989). The value of panel data in economic research. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M., editors, *Panel Surveys*, pages 486–496. J.W. Wiley and Sons, New York.

STATA Press (2005). *STATA Base Reference Manual, Release 9*. STATA Press, College Station, TX.

Steel, D. (1997). Producing monthly estimates of unemployment and employment accounting to the international labour definition. *Journal of the Royal Statistical Society. Series A*, 160(1):5–46.

Steele, F. (2008). Multilevel models for longitudinal data. *Journal of the Royal Statistical Society. Series A*, 127(1):5–19.

Taylor, M. F., Brice, J., Buck, N., and Prentice-Lane, E. (2009). *British household panel survey user manual: Volume A: Introduction, Technical Report and Appendices - Weighting, Imputation and Sampling Errors*. University of Essex, Colchester.

Tufte, E. R. (1974). *Data analysis for politics and policy*. Prentice-Hall, Englewood Cliffs, N.J.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Winkels, J. and Davies, S. (2000). Panel attrition. In Rose, D., editor, *Researching Social and Economic Change: the uses of household panel studies*, chapter 4. Routledge, London.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, Mass.

World Bank (2002). Brazil jobs report (in two volumes). 1: Policy Briefing No. 24408 - Br, The World Bank and Instituto de Pesquisa Econômica Aplicada.

World Bank (2003). Inequality and economic development in Brazil. Volume 1: Policy report, The World Bank and Instituto de Pesquisa Econômica Aplicada.

World Bank (2008). World Bank home page. `http://devdata.worldbank.org/AAG/bra_aag.pdf`.

Yang, M., Goldstein, H., Browne, W., and Woodhouse, G. (2002). Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society. Series A*, 165(1):137–153.

Yang, M., Goldstein, H., and Heath, A. (2000). Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society. Series A*, 163(1):49–62.

Zhu, L. (2008). Application des modèles de régression sur des données d'enquête. Master's thesis, Université Laval, Département de mathématiques et de statistique, Québec, Canada. In French.