Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2014)
63, *Part* 1, *pp.* 65–84

The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey

Alinne Veiga

Universidade do Estado do Rio de Janeiro, Brazil

and Peter W. F. Smith and James J. Brown

University of Southampton, UK

[Received April 2011. Final revision January 2013]

Summary. Longitudinal data from labour force surveys permit the investigation of income dynamics at the individual level. However, the data often originate from surveys with a complex multistage sampling scheme. In addition, the hierarchical structure of the data that is imposed by the different stages of the sampling scheme often represents the natural grouping in the population. Motivated by how income dynamics differ between the formal and informal sectors of the Brazilian economy and the data structure of the Brazilian Labour Force Survey, we extend the probability-weighted iterative generalized least squares estimation method. Our method is used to fit multivariate multilevel models to the Brazilian Labour Force Survey data where the covariance structure between occasions at the individual level is modelled. We conclude that there are significant income differentials and that incorporating the weights in the parameter estimation has some effect on the estimated coefficients and standard errors.

Keywords: Design weights; Labour force surveys; Longitudinal data; Multivariate multilevel models; Non-response weights; Probability-weighted iterative generalized least squares

1. Background

Labour force surveys of many countries are commonly collected by using rotating panel survey designs. With these designs, there is an element of repetition, creating longitudinal data, but units are rotated out with new units rotated in on the basis of a prescribed pattern so that over time the whole sample is being continually updated. The rotation is structured by stratifying the original sample into panels, each a representation of the overall design, so that individual panels can be rotated in and out of the sample. In the UK, the Labour Force Survey has had a rotating panel design based on addresses being selected for five successive quarters since 1992. As the data are actually collected uniformly across the three months of a quarter, there are effectively 15 monthly panels. The monthly rotation pattern can be denoted as 1–2–1–2–1–2–1–2–1, showing that a panel of addresses is in the sample for 1 month and dropped for 2 months, and that this is repeated until five interviews have taken place. The Current Population Survey in the USA has a 4–8–4 design. This maximizes the overlap between months 1 year apart and therefore supports analysis on income change at the individual and household level 1 year apart as well as gross flows for employment status year by year.

Address for correspondence: Alinne Veiga, Ciências Sociais, Universidade do Estado do Rio de Janeiro, Rua São Francisco Xavier, 524 Sala 9034-A, CEP 20550-013, Rio de Janeiro RJ, Brazil.
E-mail: alinneveiga@gmail.com

© 2013 Royal Statistical Society 0035–9254/14/63065

The Brazilian Labour Force Survey (BLFS) is another survey that can be characterized as a 4–8–4 rotating design, as shown in Table 1. Like the UK Labour Force Survey, the panel structure reflects addresses rather than individuals (households) so that out-movers from an address are replaced by in-movers. For example, the panel of addresses first entering the sample in May were interviewed on their first occasion in May 2004 and on their eighth occasion in August 2005. This creates problems with tracking the BLFS household and individual level data, with lower household matching rates partly caused by the out-movers and partly caused by traditional dropout of the stayers, leading to panel non-response (Antonaci and Silva, 2007). Although the survey is not designed to serve as a true longitudinal survey of individuals, it is possible to investigate change at the individual level. For that, we consider a longitudinal data set for heads of households, making use of further matching criteria to ensure more accurate matching at the individual level. This creates a hierarchical structure within the data with repeated observations for the head of household and an imposed gap in the data due to the 4–8–4 design as presented in Table 1. For example, heads of households who were first interviewed in March 2004 will have the gap between the months from July 2004 to February 2005 provided that they have not moved from their occasion 1 address or dropped out. Therefore, in the Brazilian context we can analyse short-term income dynamics at the individual level by using these data (Barros *et al*., 2000). Specifically, we investigate income dynamics at the individual level, focusing on the differences between the formal and informal sectors of the economy. Hence, the within- and between-individual variance and correlation structures are of substantive interest.

Table 1. Structure of the BLFS rotating design†

Interview time	*Structure for the following month of first entering the sample:*								
	January	*February*	*March*	*April*	*May*	*June*	*July*	*August*	*September*
January 2004	A1								
February 2004	A2	B1							
March 2004	A3	B2	C1						
April 2004	A4	B3	C2	D1					
May 2004		B4	C3	D2	E1				
June 2004			C4	D3	E2	F1			
July 2004				D4	E3	F2	G1		
August 2004					E4	F3	G2	H1	
September 2004						F4	G3	H2	I1
October 2004							G4	H3	I2
November 2004								H4	I3
December 2004									I4
January 2005	A5								
February 2005	A6	B5							
March 2005	A7	B6	C5						
April 2005	A8	B7	C6	D5					
May 2005		B8	C7	D6	E5				
June 2005			C8	D7	E6	F5			
July 2005				D8	E7	F6	G5		
August 2005					E8	F7	G6	H5	
September 2005						F8	G7	H6	I5
October 2005							G8	H7	I6
November 2005								H8	I7
December 2005									I8

†Letters A–I represent the cohorts (panels) followed over time and numbers 1–8 represent the measurement occasion.

Over the last 30 years, the Brazilian economy has gone through important macroeconomic changes which included an increase in the informal sector (Passos *et al.*, 2005). In general terms, informality can be defined as the lack of a formal contract or registration and social security and, more precisely, when the employment is

> 'not subject to national labour legislation, income taxation, social protection or entitlement to certain employment benefits (advance notice of dismissal, severance pay, paid annual or sick leave, etc.)'

(International Labour Organisation, 2012). This definition may vary from country to country. For example: in Brazil, informal employees are those 'without formal contract (*carteira assinada*)'; in Mexico, they are those 'without access to public or private health services by virtue of their job' and, in South Africa, they are those

> 'without written employment contract, or for whom the employer does not contribute to the pension/retirement fund or to medical aid benefits'.

Multivariate multilevel models provide the appropriate tools for the analysis of a longitudinal data set where the structure of the error covariance matrix is also of interest. These extend general multivariate regression models (Longford, 1993) by treating the successive measurements within the same individual as part of a joint multivariate normal vector. Therefore, their use is a special case of multiple-outcomes data analysis. In addition, multivariate multilevel models deal with rotating panel data, such as those of the BLFS, where missing data are determined by design. Standard multilevel models do not accommodate this feature of the sampling design in the same way, and we demonstrate that this can be handled by estimating a constrained error covariance matrix (Yang *et al.*, 2002).

In parallel to the development of multilevel approaches for longitudinal data, there has also been a growing recognition within the social sciences of the importance of the complex nature of sampling designs that generate the data which are often analysed. Most surveys adopt a sampling design that is more complex than simple random sampling with replacement. The complex sampling designs are usually stratified with multiple stages of selection, reflecting clustering in the target population, and unequal selection probabilities of the sampling units at each stage (LaVange *et al.*, 2001). For example, the initial selection of the BLFS has a stratified two-stage cluster design. Samples of addresses are selected separately from six metropolitan regions, where municipalities compose the independent strata from which the census sectors, the primary sampling units (PSUs), are selected. The PSUs are selected with probability proportional to their total number of households as listed in the 2000 census. Within each PSU the addresses are selected via systematic sampling. Although designed to be self-weighting, the BLFS still requires corrections for cross-sectional unit non-response, resulting in the loss of the self-weighting characteristic. In addition, panel non-response occurs over the subsequent occasions due to traditional dropout as well as movers. The desire to account for such design features has motivated the adaptation of traditional inference methods (see, for example, Kish and Frankel (1974), Skinner (1986, 1989a), Pfeffermann and La Vange (1989), Pessoa and Silva (1998), Pfeffermann (1993) and Lee and Forthofer (2005)). For inference about model parameters, the usual approach to account for the aspects of the sampling design is to employ pseudolikelihood estimation methods which make use of the sampling weights (Binder, 1983). This procedure is usually combined with the Taylor series linearization method for estimating the covariance matrix of the parameter estimates, which produces design consistent estimates (Skinner (1989b), page 18).

Aspects of the sampling design, beyond the natural clustering, are often ignored in the analysis of multilevel models. Motivated by this argument, Pfeffermann *et al.* (1998) developed a method

called probability-weighted iterative generalized least squares (PWIGLS) for the analysis of multilevel models of continuous outcomes. This method adapts an iterative generalized least squares (IGLS) analogue of pseudo-maximum-likelihood estimation which accounts for the unequal selection probabilities of units in each of the levels of the data structure. The use of unequal probabilities of selection in any of the stages of the sample might cause bias in the IGLS estimates. The approach of Pfeffermann *et al.* (1998) reduces or removes this bias.

The method developed in Pfeffermann *et al.* (1998) is presented for two-level multilevel models, but for longitudinal analysis with complex survey designs we typically need three levels with complex covariance structures for the errors. Other researchers have proposed alternative methods to account for the complex survey design in the analysis of longitudinal data. Skinner and Vieira (2007) investigated the effects of clustering on the variance estimation of the regression coefficients in an analysis of longitudinal survey data. Although not accounting for sampling weights, they concluded that the simple inclusion of the higher cluster level was not enough and that either the inclusion of random coefficients or the use of robust methods would be necessary to account for the effect of clustering. Vieira and Skinner (2008) took account of the sampling weights, extending the approach presented by Skinner (1989a) for panel survey data. However, their work considered the standard longitudinal modelling strategy as described by Diggle *et al.* (2002), where the model parameters are estimated under the presence of a working covariance structure. Therefore, they did not follow the multilevel modelling approach. Rabe-Hesketh and Skrondal (2006) proposed extensions to generalized linear mixed models with multiple levels, based on pseudo-maximum-likelihood estimation via adaptive quadrature. They added that their method may be applied to longitudinal models, albeit without accounting for complex level 1 covariance structures. Skinner and Holmes (2003), in contrast, 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 methods proposed, once again, accommodate only two-level longitudinal data in which the individuals are the level 2 units and occasions are the level 1 units.

A survey such as the BLFS brings together the need for longitudinal analysis to explore income dynamics and the need to account for a complex stratified multistage sample design. This adds an extra level to the hierarchy in our data structure above the individual, with unequal selection probabilities and unequal non-response across successive waves of the survey. In Section 2 we outline the multivariate multilevel models that are increasingly applied to longitudinal data. Then in Section 3 we develop a weighting approach for multivariate multilevel models of the type that is introduced in Section 2 and we develop appropriate weights for the analysis of the BLFS. This approach extends the methods that were presented by Pfeffermann *et al.* (1998) to handle the extra level and complex correlation structures. In Section 4 we apply the method to data from the BLFS to investigate the labour income dynamics of employed heads of households with a particular focus on the differences between the informal and formal sectors. Section 5 presents some further discussion.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://www.blackwellpublishing.com/rss`

2. The multivariate multilevel model

Models with time as a continuous covariate are often used to analyse longitudinal data as they model change at the individual level across time. However, successive measurements within the

same individual are expected to be correlated. For example, the income of heads of households measured in month $t+1$ is expected to be correlated with their income measured in month t even after conditioning on their 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. Embedding these models within the multilevel framework allows the introduction of random coefficients at the individual level to capture the conditional correlation structure (see, for example, Singer and Willett (2003)). The simplest two-level random-coefficient model, where the level 2 units are the individuals and level 1 units are the measurements at various occasions, introduces a random intercept for each individual, imposing a constant correlation across the repeated measures. The model can be extended by introducing a random slope on the time variable but although this relaxes the constant correlation across repeated measures it imposes a correlation structure that is difficult to justify or explain. If t_{ij} is the time of measurement i for individual j, σ^2_{u0} and σ^2_{u1} are the variances for the random intercept and the random slopes respectively, with σ_{u01} the covariance between them, and σ^2_e is the within-individual variance, the components of the main diagonal for the error covariance matrix for the random-slope model are

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

and the off-diagonal components ($i \neq i'$) are

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

The issue is that although successive measurements within the same individual are expected to be more strongly correlated than those several months apart, even after conditioning on their characteristics, this more complex correlation structure may not follow the constrained quadratic structure, as presented above, that is implied by a random slope.

The multivariate multilevel approach extends the model with a random slope for time by assuming that the conditional correlation structure can be captured by correlated random effects at each time point within an individual rather than an intercept and slope at the individual level. Consider three-level data where y_{tij} is the response on occasion $t = 1, \ldots, T$ (T being the maximum number of occasions), for individual $i = 1, \ldots, n_j$ nested in cluster $j = 1, \ldots, n$. If $T = 2$ then the correlation structure from a random slope will be equivalent to the multivariate approach but with $T > 2$ the multivariate approach will offer more flexibility. Each response is considered as a component of a multivariate normally distributed random vector $\mathbf{y}_{ij}$. These $\mathbf{y}_{ij}$ are simultaneously modelled by the following two-level multivariate model (Goldstein, 2011), which is an extension of the two-level model that was considered by Pfeffermann *et al.* (1998):

$$\mathbf{y}_{ij} = X_{ij}\boldsymbol{\beta} + Z_{ij}\mathbf{v}_j + \mathbf{u}_{ij}. \tag{1}$$

The matrix X_{ij} contains the indicator variables for each occasion, the covariates that have a common fixed effect for each occasion and the interactions between the occasion indicator variables and the covariates that have a separate fixed effect for each occasion, and the vector $\boldsymbol{\beta}$ contains the corresponding coefficients. The matrix Z_{ij} contains the covariates that have a cluster-specific random effect, including a column of 1s for the random-cluster intercept, $\mathbf{v}_j \sim N(\mathbf{0}, \Sigma_v)$ is a vector of length q containing the corresponding random-cluster effects and $\mathbf{u}_{ij} \sim N(\mathbf{0}, \Sigma_u)$ is a vector of time-specific, individual random effects. As illustrated in the example below, to ensure identification, this model does not contain an overall intercept or the main effects for the covariates that have a separate fixed effect for each occasion. The result is still a model that reflects growth at the individual level, through the set of random effects for each

individual across occasions, with clustering of individuals in local areas. However, the variance components no longer have a simple interpretation as in traditional growth curve models (Snijders and Bosker, 1999) that use a random slope on time to generate the conditional correlation structure.

Example 1: if $T = 3$ and there is one covariate with a common fixed effect, $x^{(1)}$, one covariate with a separate fixed effect for each occasion, $x^{(2)}$, one covariate with a common random effect, $z^{(1)}$, and one covariate with a separate random effect for each occasion, $z^{(2)}$, then

$$X_{ij} = \begin{pmatrix} 1 & 0 & 0 & x_{1ij}^{(1)} & x_{1ij}^{(2)} & 0 & 0 \\ 0 & 1 & 0 & x_{2ij}^{(1)} & 0 & x_{2ij}^{(2)} & 0 \\ 0 & 0 & 1 & x_{3ij}^{(1)} & 0 & 0 & x_{3ij}^{(2)} \end{pmatrix},$$

$$Z_{ij} = \begin{pmatrix} 1 & z_{1ij}^{(1)} & z_{1ij}^{(2)} & 0 & 0 \\ 1 & z_{2ij}^{(1)} & 0 & z_{2ij}^{(2)} & 0 \\ 1 & z_{3ij}^{(1)} & 0 & 0 & z_{3ij}^{(2)} \end{pmatrix},$$

Σ_v is a 5×5 matrix and Σ_u is a 3×3 matrix.

Often $q = 1$ with the matrix Z_{ij} containing only a column of 1s and $\Sigma_v = \sigma_v^2$. The error covariance matrix, which is defined as $\Sigma_r = Z_{ij}\Sigma_v Z_{ij}^{\mathrm{T}} + \Sigma_u$, can be constrained to represent a desired structure to be tested. Linear or non-linear constraints can be imposed on its elements to express the different forms of covariance structures: the exponential, the uniform or the Toepliz, for example. (See Singer and Willett (2003), Diggle *et al.* (2002) or Fitzmaurice *et al.* (2004) for the forms of these structures and others.) If no constraints are imposed, the multivariate multilevel model is fitted with an unstructured error covariance matrix.

To facilitate the presentation of the estimation procedure that is developed, the multivariate multilevel model (1) can be specified at the cluster level as

$$\mathbf{y}_j = X_j\boldsymbol{\beta} + \mathbf{r}_j, \tag{2}$$

where the vector $\mathbf{y}_j = (\mathbf{y}_{1j}, \ldots, \mathbf{y}_{n_j j})^{\mathrm{T}}$ contains the responses for the individuals in cluster j, the matrix X_j contains the X_{ij} for individuals in cluster j and $\mathbf{r}_j = (\mathbf{r}_{1j}, \ldots, \mathbf{r}_{n_j j})^{\mathrm{T}}$ is a vector of the composite errors with

$$\mathbf{r}_{ij} = Z_{ij}\mathbf{v}_j + \mathbf{u}_{ij}.$$

Here it is assumed that

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

with

$$V_j = Z_j\Sigma_v Z_j^{\mathrm{T}} + I_{n_j} \otimes \Sigma_u, \tag{3}$$

where the matrix Z_j contains the Z_{ij} for individuals in cluster j and I_{n_j} is the $n_j \times n_j$ identity matrix. The Kronecker product between I_{n_j} and the individual level error covariance matrix, Σ_u, gives a block diagonal matrix at the cluster level, where each block represents the repeated measures of one individual.

We first present the IGLS estimation method (Goldstein, 1986) for the multivariate multilevel model (2), following the notation of Pfeffermann *et al.* (1998). For this, the matrix V_j in equation (3) is expressed as a linear function of $\boldsymbol{\theta}$ such that

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

where $\boldsymbol{\theta}$ is the row vector that is formed with the s distinct elements of Σ_v and Σ_u, so that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s) = \{\text{vech}(\Sigma_v)^{\mathrm{T}}, \text{vech}(\Sigma_u)^{\mathrm{T}}\}$ and

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

The main differences between the IGLS for the multivariate multilevel model and the IGLS for the random-slope model that is described in the appendix of Pfeffermann *et al.* (1998) are the form of V_j and the definition of matrices H_{kj} and Δ_{kj}. As in Pfeffermann *et al.* (1998), H_{kj} are $q \times q$ matrices of 0s and 1s with $k = 1, \ldots, s$ and $j = 1, \ldots, n$, where s is the total number of parameters in $\boldsymbol{\theta}$ and q is the number of random effects at the cluster level. The Δ_{kj}-matrices are s $T \times T$ matrices of 0s and 1s determining the covariance structure being imposed in the multivariate model. This specification still allows for the various covariance structures to be imposed; see the example below.

Example 2: consider the Toeplitz structure for the error covariance matrix where $T = 3$ and $q = 1$, implying that $Z_j = \mathbf{1}$, a vector of 1s. Then

$$\Sigma_v = \sigma_v^2$$

and

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

Therefore, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^{\mathrm{T}} = (\sigma_v^2, \sigma^2, \sigma_1, \sigma_2)^{\mathrm{T}}$ and $s = 4$. The total variance is given as

$$V_j = \sum_{k=1}^{4} \theta_k G_{kj} = \sum_{k=1}^{4} \theta_k (\mathbf{1} H_{kj} \mathbf{1}^{\mathrm{T}} + I_{n_j} \otimes \Delta_{kj}),$$

where $H_{1j} = [1]$, $H_{2j} = [0]$, $H_{3j} = [0]$, $H_{4j} = [0]$,

$$\Delta_{1j} = \begin{pmatrix} 0 & & \\ 0 & 0 & \\ 0 & 0 & 0 \end{pmatrix},$$

$$\Delta_{2j} = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0 & 1 \end{pmatrix},$$

$$\Delta_{3j} = \begin{pmatrix} 0 & & \\ 1 & 0 & \\ 0 & 1 & 0 \end{pmatrix}$$

and

$$\Delta_{4j} = \begin{pmatrix} 0 & & \\ 0 & 0 & \\ 1 & 0 & 0 \end{pmatrix}.$$

The IGLS iterates between the estimation of $\hat{\boldsymbol{\beta}}$ (stage 1) and $\hat{\boldsymbol{\theta}}$ (stage 2) until convergence (with superscript r for the iteration 1, 2,...), given a set of initial values at the initial stage. The final iteration provides the estimates for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$. Following the equivalent stages to those presented by Pfeffermann *et al*. (1998), the IGLS algorithm for multivariate multilevel models is as follows.

In the initialization stage initial values for both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are calculated. The initial estimate for $\hat{\boldsymbol{\beta}}$ is, using ordinary least squares,

$$\hat{\boldsymbol{\beta}}^{(0)} = \left(\sum_{j=1}^{n} X_j^{\mathrm{T}} X_j \right)^{-1} \sum_{j=1}^{n} X_j^{\mathrm{T}} \mathbf{y}_j.$$

For the initial values of $\hat{\boldsymbol{\theta}}$ it is assumed that $\hat{\Sigma}_v^{(0)}$ is the zero matrix. Given the residuals

$$\hat{\mathbf{e}}_{ij}^{(0)} = \mathbf{y}_{ij} - X_{ij}\hat{\boldsymbol{\beta}}^{(0)}$$

and

$$\hat{\mathbf{u}}_j^{(0)} = \sum_i \hat{\mathbf{e}}_{ij}^{(0)} / n_j,$$

the initial estimate of Σ_u is

$$\hat{\Sigma}_u^{(0)} = \frac{\sum_j \left\{ \sum_i (\hat{\mathbf{e}}_{ij}^{(0)} - \hat{\mathbf{u}}_j^{(0)})(\hat{\mathbf{e}}_{ij}^{(0)} - \hat{\mathbf{u}}_j^{(0)})^{\mathrm{T}} \right\}}{\sum_j (n_j - 1)},$$

where Σ_j denotes the sum over the clusters ($j = 1, 2, \ldots, n$) in the sample and Σ_i denotes the sum over the level 1 units sampled in cluster j.

Step 1: at this stage $\hat{\boldsymbol{\beta}}^{(r)}$ is calculated by using the same formulation as in Pfeffermann *et al.* (1998), equation (5):

$$\hat{\boldsymbol{\beta}}^{(r)} = P^{(r)-1} Q^{(r)}, \tag{4}$$

where

$$P^{(r)} = \sum_j X_j^{\mathrm{T}} V_{jr}^{-1} X_j$$

and

$$Q^{(r)} = \sum_j X_j^{\mathrm{T}} V_{jr}^{-1} \mathbf{y}_j.$$

The matrix V_{jr}^{-1}, which is the inverse of $V_{jr} = V_j(\hat{\boldsymbol{\theta}}^{(r-1)})$, for the multivariate multilevel model can be written as

$$V_{jr}^{-1} = I_{n_j} \otimes \hat{\Sigma}_u^{-1} - (I_{n_j} \otimes \hat{\Sigma}_u^{-1}) Z_j A_j Z_j^{\mathrm{T}} (I_{n_j} \otimes \hat{\Sigma}_u^{-1}), \tag{5}$$

where

$$A_j = \{\hat{\Sigma}_v^{-1} + Z_j^{\mathrm{T}} (I_{n_j} \otimes \hat{\Sigma}_u^{-1}) Z_j\}^{-1}.$$

This follows from the results of Goldstein (1986) and Searle *et al*. (1992). (See Appendix A for details.) The fixed coefficients are thereby estimated for iteration (r). For simplicity of notation, in the above formula, the superscript $(r-1)$, for the previous iteration, was omitted from $\hat{\Sigma}_v$ and $\hat{\Sigma}_u$.

Step 2: at this stage, $\hat{\boldsymbol{\theta}}^{(r)}$ is also calculated by using the formulation in Pfeffermann *et al.* (1998), equation (5):

$$\hat{\boldsymbol{\theta}}^{(r)} = R^{(r)-1} S^{(r)}, \tag{6}$$

where $R^{(r)}$ is an $s \times s$ matrix and $S^{(r)}$ is a vector of length s. The $[k, l]$ element of $R^{(r)}$ is

$$R^{(r)}[k,l] = \sum_j \mathrm{tr}(V_{jr}^{-1} G_{kj} V_{jr}^{-1} G_{lj})$$

and the kth element of $S^{(r)}$ is

$$S^{(r)}[k] = \sum_j \mathrm{tr}(\hat{\mathbf{e}}_j^{\mathrm{T}} V_{jr}^{-1} G_{kj} V_{jr}^{-1} \hat{\mathbf{e}}_j),$$

where $\hat{\mathbf{e}}_j = \mathbf{y}_j - X_j \hat{\boldsymbol{\beta}}^{(r)}$ are the raw residuals from step 1.

At convergence, the covariance estimators for the IGLS estimates are calculated as

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

and, by analogy with Goldstein (2011),

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\theta}}_{\mathrm{IGLS}}) = 2R^{-1}.$$

3. Estimation method

3.1. Probability-weighted iterative generalized least squares

When working with a sample that is generated by a complex sample design, weights are often used within regression analysis to estimate the model as if it were fitted to a census of the population rather than the sample. Often, this is referred to as pseudo-maximum-likelihood and can be viewed as estimating census sums in ordinary least squares with weighted sample sums (Skinner, 1989b; Binder, 1983). It was developed with the motivation of correcting the bias that usually occurs when ordinary least squares regression analysis is performed on samples with an informative design, i.e. when the unequal selection probabilities result in weights that are related to the outcome variable. The PWIGLS estimation method, as presented by Pfeffermann *et al.* (1998), incorporates the ideas of pseudo-maximum-likelihood estimation to correct for bias resulting from using IGLS with samples from an informative design. The method allows for the estimation of two-level random-coefficients models for continuous outcomes, producing estimators of standard errors that are robust to model failure within the PSUs but assume independence between PSUs (Rabe-Hesketh and Skrondal, 2006).

Pseudo-maximum-likelihood estimation is not as straightforward for multilevel models as it is for single-level models because the covariance structure of the population values is modelled. For this reason the census log-likelihood function cannot be expressed as a simple sum but as a sum across all levels of the data hierarchy. For a two-level model, for example, this may be written as

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

where $\boldsymbol{\varphi}$ is the vector of parameters in the model, $\phi(\mathbf{u}_j)$ is the multivariate normal probability density of the level 2 random effects, $\log\{ f_{ij}(y_{ij}, \boldsymbol{\varphi} | \mathbf{u}_j) \}$ is the log-likelihood contribution of the level 1 units conditioned on the level 2 random effects (Rabe-Hesketh and Skrondal, 2006), N

is the number of clusters (level 2 units) in the population and N_j is the number of level 1 units in cluster j in the population. The log-likelihood for the sample units cannot be expressed as a simple weighted sum of the sample contributions either, but rather as

$$\hat{l}(\boldsymbol{\varphi}) = \sum_{j \in S} w_j \log\Big\{ \int \Big(\exp\Big[\sum_{i \in S_j} w_{i|j} \log\{ f_{ij}(y_{ij}, \boldsymbol{\varphi} | \mathbf{u}_j) \} \Big] \Big) \phi(\mathbf{u}_j) \, \mathrm{d}\mathbf{u}_j \Big\},$$

where S contains the indices of the sampled clusters and S_j the indices of the sample level 1 units in cluster j. The main difference is that the sum for each of the levels requires the respective conditional probabilities of selection. The log-likelihood contributions of the level 1 units are weighted by $w_{i|j}$, i.e. the inverse of the selection probability of unit i in cluster j given that cluster j has been selected; and log-likelihood contributions of the level 2 units are weighted by w_j, i.e. the inverse of the selection probability of cluster j.

Pfeffermann *et al.* (1998) followed the same underlying idea. Firstly, the IGLS estimation procedure for the census parameters was specified. Secondly, weighted sample estimates were substituted for the census estimators. Therefore, each population sum over the level 2 units was replaced by the weighted sample sum using w_j, and each population sum over the level 1 units was replaced by a weighted sample sum using $w_{i|j}$. The selection probabilities, expressed as weights, are then incorporated in the estimation for both the fixed and the variance–covariance components of the model.

Like the pseudo-maximum-likelihood 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 sampled level 2 units n and the number of sampled level 1 units n_j within cluster j increase (i.e. $n \to \infty$ and $n_j \to \infty$). (See Pfeffermann *et al.* (1998), pages 28 and 29, for further discussion.) In Pfeffermann *et al.* (1998), three scenarios were tested in a simulation study. 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.

Most of the discussion that is presented in Pfeffermann *et al.* (1998) was concerned with methods of scaling the weights, w_j and $w_{i|j}$, to reduce the bias that is generated by having small samples from sampled clusters when implementing their method. This issue was also discussed by Rabe-Hesketh and Skrondal (2006). Pfeffermann *et al.* (1998) showed that different scaling methods affect the estimates in different ways. Their so-called 'method two' was recommended as the best potential scaling method under weights from informative designs. This particular method multiplies the original level 1 weights $w^*_{i|j}$ by a constant λ_j, the inverse of the average weight in cluster j, so that the sum of the scaled level 1 weights equals the cluster sample size n_j. Following this specification, the scaled weights may be written as

$$w_{i|j} = \lambda_j w^*_{i|j} = w^*_{i|j} n_j \Big/ \sum_i w^*_{i|j}. \tag{7}$$

For a two-level multilevel model, scaling is required for only the level 1 weights. The multiplication of level 2 weights by a constant only rescales the pseudolikelihood. Therefore, it has no effects on the estimates (Rabe-Hesketh and Skrondal, 2006). The scaling of level 1 weights, however, may have an effect on the estimates of the fixed part of the model and a bigger effect on the estimation of the variance components.

For a two-level multivariate longitudinal model, such as those presented in Section 2, the PWIGLS estimation follows from the same idea. Attention needs to be given to the set of weights to be included in the analysis. It is necessary that a set of weights for each level of the

data hierarchy is available. For a three-level multivariate balanced data set, where occasions define the multivariate structure, the sets of weights that are needed are the same as before: $w_{i|j}$ and w_j. These, however, are rarely provided by data producers but are essential for the PWIGLS to be implemented. If not available, they can be calculated from the longitudinal weights, which are usually provided with longitudinal survey data to reflect the sample design and to compensate for sample dropout between successive occasions. Skinner and Holmes (2003) showed how to split the longitudinal weights into individual weights, to reflect the sample design, and occasion level weights, to adjust for dropout.

For the framework that is presented here there is no need to have distinct occasion level weights; level 2 weights w_j are still the cluster level weights, i.e. the inverse of the probability of selecting cluster j, and the level 1 weights $w_{i|j}$ are the individual weights for completers, adjusted for non-response. Hereafter, let

$$W_j = \mathrm{diag}(w_{1|j}, \ldots, w_{n_j|j}) \tag{8}$$

be the diagonal matrix with the n_j individual longitudinal weights on the diagonal.

3.2. Probability-weighted iterative generalized least squares for multivariate multilevel models

The PWIGLS estimation method for multivariate multilevel models modifies the IGLS algorithm that was presented in Section 2 by estimating census versions of P, Q, R and S as follows.

(a) Replace every sum over j by a weighted sum with cluster weights w_j.
(b) Replace every sum over i by a weighted sum with individual weights $w_{i|j}$, noting that for the implicit sums over i in P, Q, R and S this is equivalent to replacing the identity matrix by W_j in V_{jr}^{-1}, A_j and G_{kj}, and for those in the initialization stage to using weighted least squares with weight matrix W_j.
(c) In the initialization stage, replace n_j by $\hat{N}_j = \Sigma_i w_{i|j}$.

For example, in the initialization stage, $\hat{\boldsymbol{\beta}}^{(0)} = \Sigma_j w_j (X_j^{\mathrm{T}} W_j X_j)^{-1} \Sigma_j w_j (X_j^{\mathrm{T}} W_j \mathbf{y}_j)$ and, in step 1, $\hat{P}^{(r)} = \Sigma_j w_j (X_j^{\mathrm{T}} V_{jr}^{-1} X_j)$, where the I_{n_j} in V_{jr}^{-1} and A_j have been replaced by W_j.

Pfeffermann *et al.* (1998) used the Taylor series linearization method, as described in Skinner (1989b), to estimate the variance of the PWIGLS estimates. This method is based on the randomization variance (Pfeffermann, 1993), assuming that the level 2 units were selected with replacement and that the contributions to the pseudolikelihood are independent. The Taylor series linearization method provides robust estimates for the standard errors in the form of the sandwich estimator (Huber, 1967; White, 1982; Freedman, 2006). The variance estimator for the PWIGLS estimate of the β-coefficients is

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{P}^{-1} \frac{n}{n-1} \Big(\sum_{j \in S} w_j^2 \mathbf{c}_j \mathbf{c}_j^{\mathrm{T}} \Big) \hat{P}^{-1}, \tag{9}$$

where $\mathbf{c}_j = (X_j \hat{V}_j^{-1} \hat{\mathbf{e}}_j)$. Given that $\hat{\boldsymbol{\theta}}$ is estimated by equation (6), an estimate for $\mathrm{var}(\hat{\boldsymbol{\theta}})$ is

$$\begin{aligned} \widehat{\mathrm{var}}(\hat{\boldsymbol{\theta}}) &= \widehat{\mathrm{var}}(\hat{R}^{-1} \hat{S}) \\ &= \hat{R}^{-1} \hat{V}_L(\hat{S}) \hat{R}^{-1} \\ &= \hat{R}^{-1} \frac{n}{n-1} \Big\{ \sum_{j \in S} w_j^2 (\hat{S} - \hat{R}\hat{\boldsymbol{\theta}})(\hat{S} - \hat{R}\hat{\boldsymbol{\theta}})^{\mathrm{T}} \Big\} \hat{R}^{-1}, \end{aligned} \tag{10}$$

with $\hat{V}_L$ determined following the linearization principles that are presented in Skinner (1989b) and Vieira and Skinner (2008). Here the lack of a subscript (r) indicates the value of the vector or matrix when the PWIGLS algorithm has converged. The variances in equations (9) and (10) can be used to calculate Wald confidence intervals and test statistics for the β-coefficients and the $\boldsymbol{\theta}$.

This extended method can be applied to both longitudinal data sets and cross-sectional data sets. Computer routines for the implementation of such a method were also developed and are available from

`http://www.blackwellpublishing.com/rss`

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 directly but can be approximated via the Toeplitz form that constrains the elements of each diagonal to be constant, but does not directly impose the auto-regressive decay down the columns.

4. Application to the Brazilian Labour Force Survey

We now apply the methodology that was developed in Section 3 to investigate income dynamics in Brazil at the individual level with a focus on the differences between the formal and informal sectors of the economy. Multivariate multilevel models are fitted to longitudinal data on monthly labour income of employed heads of households in the BLFS from January 2004 to December 2005. There are 8592 heads of households who meet this inclusion criterion at occasion 1. Of these, 6524 respond on all eight occasions and the completers form the data set that was used in the substantive modelling. Although restricting the analysis to completers loses 24% of the heads of households, it discards only 16% of the 62 335 observations in the individual-by-occasion data. Models for labour income are usually estimated on the basis of the Mincer equation, which was proposed by Mincer and Polachek (1974), which uses the log-transformation of the response variable. Previous attempts to fit a model based on the Mincer equation to the logarithm of income resulted in a model including the covariates listed below. In addition, the hierarchical structure of the data that are used in this section allows the inclusion of covariates at three different levels:

(a) occasion level covariates—*Occasion* (eight indicator variables); *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*; *Working Hours* (on the log-scale); *Proxy Respondent*; *Number of Household Members*;
(b) individual level covariates—*Gender* (females as the baseline); *Age* at the first interview and a squared term; *Race* (whites against all other categories collapsed); *Education* in years of schooling from 0 to 17; *Month of First Entering the Sample* from January 2004 to September 2004;
(c) cluster level covariates—*Metropolitan Region* (Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo and Porto Alegre); cluster means for Age and Education, and proportions for Race (proportion of white), for the categories of Type of Worker and for Proxy Respondent.

Two sets of estimates were computed for a multivariate multilevel model of the logarithm of income for a balanced longitudinal data set with eight occasions for every head of household: one weighted, using the proposed extension to the PWIGLS method, and one unweighted, using

standard IGLS. The weights that were used were the w_j, for the PSU level, and an adjusted $w_{i|j}$, for the heads-of-household level. The cluster level weights w_j were computed as the inverse of the PSU selection probabilities. The adjusted weights are

$$w^*_{i|j} = w_{ij(8)}/w_j, \tag{11}$$

where the longitudinal weights $w_{ij(8)}$ represent the inverse of the probability of responding up to occasion 8. These were obtained by fitting logistic regression models utilizing all 62 335 observations in the individual-by-occasion data, considering data from neighbouring pairs of occasions, to estimate the response probabilities as in Lepkowski (1989). This method therefore makes use of all the available covariate information at each occasion and conditionally adjusts the weights of each occasion given the adjustment for the previous occasion. All the covariates that are listed above were considered for inclusion in the logistic regression models but only those that were significant at the 5% level were retained for calculating the weights. The following covariates were retained in one or more of the models: Month of First Entering the Sample, Age, Education, Proxy Respondent, Number of Household Members and Metropolitan Region. Note that the level 1 weights (heads of household) were not necessarily constant within the level 2 units (PSUs). Therefore, the scaling method that is applied to the level 1 weights to obtain $w_{i|j}$ is expected to have some effect on the parameter estimates.

In a complete-case analysis, the longitudinal weights for the heads of households within a cluster compensate for the losses of data between each pair of occasions by reweighting to the first occasion. As the non-response is reflected in the weights this will naturally be reflected in the variance estimator that was outlined in Section 3.2 conditional on the weight. However, an alternative would be to impute the missing data and to take into account the imputation in the variance estimation. Carillo *et al.* (2011) took an imputation approach but in the context of fitting models by using generalized estimating equations and with missing responses. Here we have missing occasion-specific covariates as well as the missing response and are directly estimating the conditional correlation structure in the data. Hence, a donor-based imputation approach (see, for example, Andridge and Little (2010) and Little and Rubin (2002), section 4.3.2) would be considerably more difficult and the generalized estimating equation method would not be appropriate. With the unweighted data, the loss is considered to be missing at random and the clusters can be considered as a simple random sample given only the covariates in the model. When the IGLS method was used for the unweighted estimators, the same robust estimation methods for the standard errors were applied.

An important feature of the BLFS sampling design is its rotation scheme. The models fitted in this section incorporate this feature. As mentioned, there is an 8-month gap between the fourth and the fifth interviews for every head of household. This is a similar issue to that when modelling unequally spaced time data and is accounted for by constraining the parameters of the error covariance matrix Σ_u. Starting with a completely unstructured covariance matrix and the unweighted data, on the basis of the Akaike information criterion AIC and the Bayesian information criterion BIC preliminary model selection exercises considered the different forms of covariance structures. The model imposing an unstructured error covariance matrix had the smallest values for AIC (26 691) and BIC (26 893). However, this was not the preferred structure, since it is the least parsimonious model and the estimation of additional nuisance parameters can cause loss of efficiency in the inference for the fixed part of the model. The traditional random-slope model had AIC (27 534) and BIC (27 561) considerably larger. Thus, a modified lag-dependent structure, similar to the Toeplitz structure, but also incorporating the structural gap, was considered (AIC = 26 950 and BIC = 27 016). Here, this structure is referred to as a general linear lag-dependent covariance structure and for the BLFS data has the form

Table 2. Means or frequency distributions and estimated β-coefficients for the BLFS data†

Covariate	*% or means*		*Estimates*	
	Unweighted	*Weighted*	*Unweighted*	*Weighted*
Occasion				
1			4.401 (0.139)	4.311 (0.163)
2			4.406 (0.139)	4.314 (0.162)
3			4.404 (0.139)	4.306 (0.163)
4			4.410 (0.139)	4.318 (0.162)
13			4.439 (0.139)	4.343 (0.162)
14			4.434 (0.139)	4.339 (0.162)
15			4.433 (0.139)	4.336 (0.162)
16			4.438 (0.139)	4.343 (0.162)
White	54.68%	59.35%	0.121 (0.018)	0.102 (0.023)
Males	76.99%	77.79%	0.599 (0.137)	0.530 (0.164)
Age (at occasion 1)	42.27	42.35	3.485 (1.429)‡	2.062 (1.767)‡
squared term			−0.450 (0.056)‡	−0.474 (0.070)‡
Education (at occasion 1)	8.44	8.74	−0.069 (0.012)	−0.083 (0.014)
squared term			0.009 (0.001)	0.010 (0.001)
Type of Worker (formal as baseline)	47.30%	45.83%		
Informal	14.04%	14.87%	−0.091 (0.014)	−0.077 (0.018)
Employer	7.86%	8.43%	0.044 (0.029)	0.018 (0.032)
Military service	7.73%	7.35%	0.038 (0.020)	0.060 (0.023)
Self-employed	23.07%	23.53%	−0.131 (0.019)	−0.151 (0.023)
Type of activity (manufacturing as baseline)	19.14%	18.84%		
Building	9.92%	9.73%	−0.007 (0.014)	−0.015 (0.016)
Commerce	18.30%	18.20%	−0.041 (0.009)	−0.044 (0.011)
Financial	14.18%	15.24%	−0.006 (0.012)	−0.009 (0.012)
Social services	14.12%	14.19%	0.033 (0.015)	0.028 (0.016)
Domestic services	5.38%	4.76%	−0.045 (0.023)	−0.033 (0.023)
Other services	18.38%	18.57%	−0.029 (0.011)	−0.035 (0.014)
Other activities	0.58%	0.46%	−0.028 (0.034)	−0.011 (0.024)
Duration of Employment (× 120)	103.52	102.09	0.053 (0.006)	0.050 (0.009)
squared term			−0.014 (0.003)	−0.014 (0.004)
Working Hours (logged)	3.77	3.77	0.251 (0.029)	0.257 (0.036)
Proxy Respondent	1.55%	1.54%	0.004 (0.008)	0.009 (0.008)
Metropolitan Region (Recife as baseline)	6.32%	3.04%		
Salvador	10.53%	5.14%	0.037 (0.033)	0.013 (0.040)
Belo Horizonte	14.85%	7.57%	0.244 (0.029)	0.254 (0.032)
Rio de Janeiro	29.71%	36.03%	0.211 (0.028)	0.217 (0.029)
São Paulo	24.77%	41.11%	0.360 (0.030)	0.368 (0.032)
Porto Alegre	13.83%	7.12%	0.226 (0.033)	0.239 (0.035)
Interaction: White and the following				
Type of Worker (formal as baseline)	46.37%	45.49%		
Informal	12.34%	13.18%	0.043 (0.020)	0.028 (0.024)
Employer	22.20%	22.40%	0.105 (0.038)	0.114 (0.043)
Military service	8.21%	7.66%	−0.041 (0.027)	−0.066 (0.030)
Self-employed	10.88%	11.27%	0.090 (0.028)	0.101 (0.035)
Interaction: Male and the following				
Age (at occasion 1)	41.77	41.82	0.004 (0.002)	0.006 (0.002)
Education (at occasion 1)	8.44	8.71	0.057 (0.014)	0.076 (0.016)
squared term			−0.003 (0.001)	−0.005 (0.001)
Working Hours (logged)	3.80	3.81	−0.085 (0.034)	−0.088 (0.041)
Proxy Respondent	1.61%	1.60%	−0.042 (0.009)	−0.040 (0.010)

(*continued*)

Table 2 (*continued*)

Covariate	*% or means*		*Estimates*	
	Unweighted	*Weighted*	*Unweighted*	*Weighted*
Contextual effects				
Proportion of Informal			−0.330 (0.133)	−0.289 (0.150)
Proportion of Employer			1.845 (0.234)	1.690 (0.237)
Proportion of Military			−0.225 (0.209)	−0.395 (0.250)
Proportion of Self-employed			−0.435 (0.129)	−0.273 (0.155)
Average Education			0.072 (0.007)	0.086 (0.008)

†Estimated standard errors are given in parentheses.
‡Values $\times 10^3$.

$$\Sigma_r = \sigma_v^2 \mathbf{1}\mathbf{1}^{\mathrm{T}} + \begin{pmatrix} \gamma 0 & & & & & & & \\ \gamma 1 & \gamma 0 & & & & & & \\ \gamma 2 & \gamma 1 & \gamma 0 & & & & & \\ \gamma 3 & \gamma 2 & \gamma 1 & \gamma 0 & & & & \\ \gamma 12 & \gamma 11 & \gamma 10 & \gamma 9 & \gamma 0 & & & \\ \gamma 13 & \gamma 12 & \gamma 11 & \gamma 10 & \gamma 1 & \gamma 0 & & \\ \gamma 14 & \gamma 13 & \gamma 12 & \gamma 11 & \gamma 2 & \gamma 1 & \gamma 0 & \\ \gamma 15 & \gamma 14 & \gamma 13 & \gamma 12 & \gamma 3 & \gamma 2 & \gamma 1 & \gamma 0 \end{pmatrix}, \tag{12}$$

where the subscripts for the covariance terms γ represent the time distance between the pairs of occasions. Because σ_v^2 is added to every entry of Σ_u, the resulting structure of Σ_r is the same as that imposed on Σ_u. This covariance structure has a total of 12 parameters to be estimated as components of $\hat{\boldsymbol{\theta}}$ and is a more parsimonious alternative to the unstructured matrix. The same model was then fitted by using the weights. Hence, a total of 12 H_{kj}-matrices need to be specified for the PWIGLS estimation.

The second column of Table 2 presents the unweighted percentages of heads of household in each category of the categorical covariates and the means for the continuous covariates at occasion 1. The third column presents these summary statistics weighted by the longitudinal weights $w_{ij(8)}$. The percentage of heads of households in the informal sector is relatively large and the difference between the unweighted and weighted percentages indicates differential non-response at occasion 1 and subsequent drop-out, with those in the informal sector being more likely to have a non-response at some occasion than those in the formal sector. Similarly, white heads of households are more likely to have a non-response. There are also differences across the regions with, as expected, the larger metropolitan areas (Rio de Janeiro and São Paulo) experiencing more non-response. The block of percentage labelled 'Interaction: White and Type of Worker' provides the distribution of whites across the various types of worker. The percentage of whites in the formal and informal sectors are similar to those for heads of households as is the effect of weighting.

Cross-tabulations (which are not presented) reveal that heads of household working in the formal and informal sectors are distributed across all categories of the other covariates. Also, there is some movement between the sectors over time with 81% and 49% of heads of household in respectively the formal and informal sector at occasion 1 remaining in the formal and informal sector for all eight occasions.

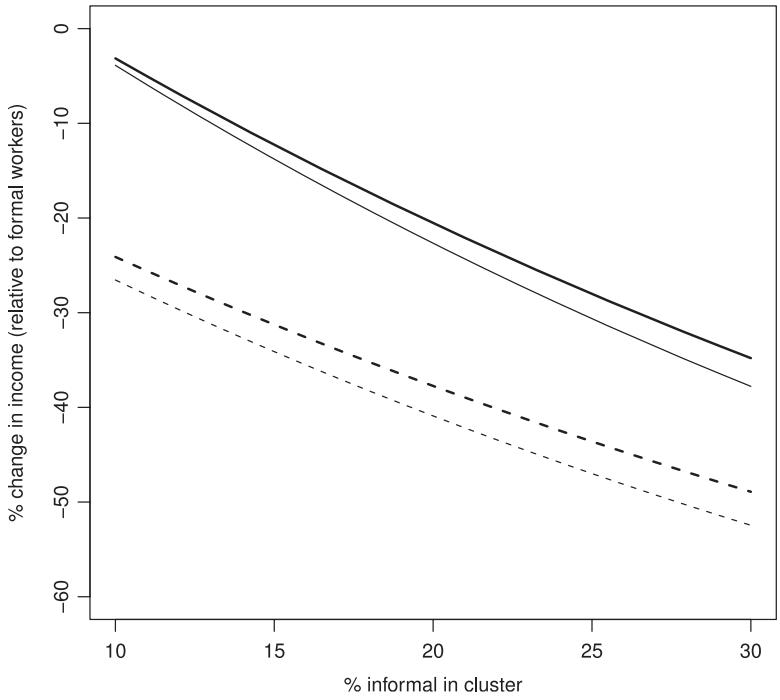

Fig. 1. Estimated relative effects of working in the informal sector: ———, PWIGLS, white; ———, IGLS, white; – – –, PWIGLS, non-white, ------, IGLS, non-white

The fourth and fifth columns of Table 2 present two sets of estimates for the fixed part of the selected multivariate multilevel models: one estimated via IGLS and the other via PWIGLS. The model includes variables for all three levels of the data hierarchy and interaction terms. It also includes the stratum variable (Metropolitan Region). The use of weights is, therefore, expected to correct for biases due to the non-response and the complex sample design that are not accounted for by the covariates which are already in the model, such as the selection of PSUs with probability proportional to the number of households. Comparing the two sets of estimated β-coefficients for the model, small differences relative to their estimated standard errors are found, except for the estimates of the effects of White, Type of Worker and the contextual variables for the proportion of self-employed and average education. Some of these differences can be explained by the differential drop-out that was noted earlier, for which the weights are correcting. The estimated standard errors of the PWIGLS estimated β-coefficients are noticeably larger than those for IGLS, reflecting the effect of the weighting on the accuracy of the estimates.

Having addressed the methodological aspects of the application, we now turn our attention to the substantive side of the results, focusing on the income dynamics and differences between the formal and informal sectors. Both sets of estimates for the underlying growth in income indicate little growth across successive months but do suggest a 3% increase in real income when comparing occasions that are 1 year apart. Fig. 1 helps to assess how income dynamics differ between the formal and informal sectors. For each set of estimates, it presents the difference in income for informal workers, relative to formal workers, allowing for the ethnicity effect and the strong contextual effect of informal workers within the cluster. Both sets of estimates show

Table 3. Estimated residual covariance matrices†

Estimator	*General linear lag-dependent matrix*
Unweighted $\hat{\sigma}_v^2 = 0.0107$ (0.0027)	$\hat{\Sigma}_u =$
	0.2957 (0.0067)
	0.2543 (0.0065) 0.2957 (0.0067)
	0.2489 (0.0065) 0.2543 (0.0065) 0.2957 (0.0067)
	0.2441 (0.0065) 0.2489 (0.0065) 0.2543 (0.0065) 0.2957 (0.0067)
	0.2117 (0.0064) 0.2131 (0.0064) 0.2172 (0.0065) 0.2174 (0.0066) 0.2957 (0.0067)
	0.2095 (0.0064) 0.2117 (0.0064) 0.2131 (0.0064) 0.2172 (0.0064) 0.2543 (0.0065) 0.2957 (0.0065)
	0.2076 (0.0063) 0.2095 (0.0064) 0.2117 (0.0064) 0.2131 (0.0064) 0.2489 (0.0065) 0.2543 (0.0065) 0.2957 (0.0067)
	0.2051 (0.0064) 0.2076 (0.0063) 0.2095 (0.0064) 0.2117 (0.0064) 0.2441 (0.0065) 0.2489 (0.0065) 0.2543 (0.0065) 0.2957 (0.0067)
Weighted $\hat{\sigma}_v^2 = 0.0096$ (0.0032)	$\hat{\Sigma}_u =$
	0.3053 (0.0084)
	0.2663 (0.0082) 0.3053 (0.0084)
	0.2615 (0.0081) 0.2663 (0.0082) 0.3053 (0.0084)
	0.2574 (0.0081) 0.2615 (0.0081) 0.2663 (0.0082) 0.3053 (0.0084)
	0.2243 (0.0080) 0.2263 (0.0081) 0.2290 (0.0083) 0.2297 (0.0085) 0.3053 (0.0084)
	0.2222 (0.0079) 0.2243 (0.0080) 0.2263 (0.0081) 0.2290 (0.0083) 0.2663 (0.0082) 0.3053 (0.0084)
	0.2199 (0.0079) 0.2222 (0.0079) 0.2243 (0.0080) 0.2263 (0.0081) 0.2615 (0.0081) 0.2663 (0.0082) 0.3053 (0.0084)
	0.2176 (0.0079) 0.2199 (0.0079) 0.2222 (0.0079) 0.2243 (0.0080) 0.2574 (0.0081) 0.2615 (0.0081) 0.2663 (0.0082) 0.3053 (0.0084)

†Estimated standard errors are given in parentheses.

clearly the disadvantage of working in the informal sector. This is particularly pronounced for non-whites and those living in clusters with a high proportion of informal workers. However, this is a mixture of the effect of moving sectors and the structural differences of the informal sector in the economy. Separating these effects would require a respecification of the Type of Worker covariate as discussed in Singer and Willett (2003), page 176. Again using PWIGLS tends to moderate the effects relative to using IGLS.

Table 3 and Fig. 2 present the weighted and unweighted estimates of σ_v^2 and Σ_u, and associated auto-correlation functions for the model selected. These data are strongly correlated. However, the differences between estimation methods are again not large, reinforcing the small effect of applying the weights in this example. This could be an indication that the sampling and non-response are not informative at either level given the covariates in the model. The estimated standard errors of the PWIGLS estimated $\boldsymbol{\theta}$ are again larger than those for IGLS.

The model that was fitted in this section explored the potential of the multilevel approach for the analysis of a complex longitudinal data set. The labour income of heads of households was found to be relatively stable over the short time window that is covered by the BLFS design, but with large and significant differences between the formal and informal sectors. The use of the

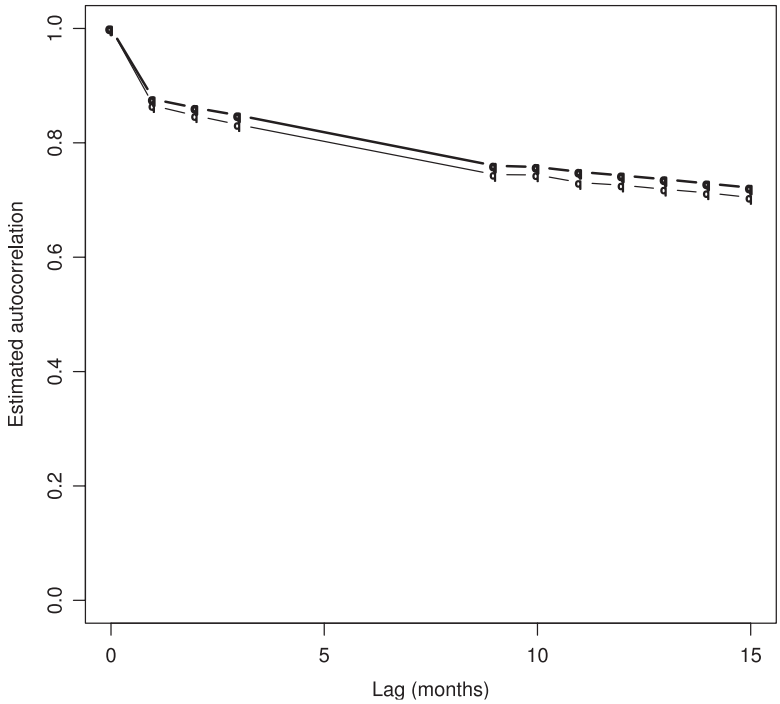

Fig. 2. Estimated auto-correlation functions: ———, PWIGLS; ———, IGLS

sampling weights when fitting the model tended to increase the estimated standard errors of the parameter estimates.

5. Conclusions

The BLFS data originating from a multistage sampling design, which involves unequal probabilities of selection and the clustering of basic units in higher level units, have a hierarchical structure. These data are collected through rotating sampling schemes that substitute part of the sample in successive waves. In addition the data display a complex error covariance structure due to the dependence over time between the responses within the same individual. Pfeffermann *et al.* (1998) offered a framework to account for the sampling design in multilevel models. This framework can be applied to two-level random-coefficient models without accounting for the complex error correlation structure.

This paper extended this framework to multivariate multilevel models and to handle different data complexities: the hierarchical data structure; the complex residual correlation structure; the features of the sampling design which included the sampling weights and the rotating panels; panel non-response.

An analysis of the BLFS accounting for the sampling design features was compared with an analysis using IGLS. We found significance income differentials and that allowing for the weights has some effect on the estimated coefficient and standard errors, although a strong effect of the weights was not observed. This could be because the sample design is non-informative conditional on the covariates in the model. Given that we have used similar covariates in our logistic regression model, this is highly likely. However, many surveys provide longitudinal

weights to adjust for dropout that are informative even after controlling for covariates, since they are based on information that is not publicly released. In such situations, the parameter estimates from using IGLS would be biased, which would be alleviated by using the PWIGLS method that is proposed in this paper.

Acknowledgements

Veiga was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, which is an agency of the Brazilian Ministry of Education, under grant BEX 2392-04-6. This research was undertaken while Veiga was a doctoral student at the University of Southampton.

Appendix A: Deriving V_j^{-1}

By using the result

$$(A + BCB^{\mathrm{T}})^{-1} = A^{-1} - A^{-1}B(C^{-1} + B^{\mathrm{T}}A^{-1}B)^{-1}B^{\mathrm{T}}A^{-1}$$

and writing $V_j = I_{n_j} \otimes \Sigma_u + Z_j \Sigma_v Z_j^{\mathrm{T}}$ as $V_2 + Z_j \Sigma_v Z_j^{\mathrm{T}}$, we have

$$V_j^{-1} = {V_2}^{-1} - {V_2}^{-1} Z_j (\Sigma_v^{-1} + Z_j^{\mathrm{T}} {V_2}^{-1} Z_j)^{-1} Z_j^{\mathrm{T}} {V_2}^{-1}.$$

Since ${V_2}^{-1} = I_{n_j} \otimes \Sigma_u^{-1}$,

$$V_j^{-1} = (I_{n_j} \otimes \Sigma_u^{-1}) - (I_{n_j} \otimes \Sigma_u^{-1}) Z_j A_j Z_j^{\mathrm{T}} (I_{n_j} \otimes \Sigma_u^{-1}),$$

where

$$A_j = \{\Sigma_v^{-1} + Z_j^{\mathrm{T}} (I_{n_j} \otimes \Sigma_u^{-1}) Z_j\}^{-1}.$$

References

Andridge, R. R. and Little, R. J. A. (2010) A review of hot deck imputation for survey non-response. *Int. Statist. Rev.*, **78**, 40–64.

Antonaci, G. de Abreu and Silva, D. B. do Nascimento (2007) Analysis of alternative rotation patterns for the Brazilian system of integrated household surveys. *Proc. 56th Sessn Int. Statist. Inst.*

de Barros, R. P., Corseuil, C. H. and Leite, P. G. (2000) Labor market and poverty in Brazil. *Texto Para Discussão 723*. Instituto de Pesquisa Econômica Aplicada, Brasília.

Binder, D. (1983) On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.*, **51**, 279–292.

Carillo, I., Chen, J. and Wu, C. (2011) Pseudo-gee approach to analyzing longitudinal surveys under imputation for missing responses. *J. Off. Statist.*, **2**, 255–277.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002) *The Analysis of Longitudinal Data*, 2nd edn. Oxford: Oxford University Press.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004) *Applied Longitudinal Analysis*. New York: Wiley-Interscience.

Freedman, D. A. (2006) On the so-called "Huber-sandwich estimator" and "robust standard errors". *Am. Statistn*, **60**, 299–302.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

Goldstein, H. (2011) *Multilevel Statistical Models*, 4th edn. Chichester: Wiley.

Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* (ed. J. N. LeCam), pp. 221–233. Berkeley: University of California Press.

International Labour Organisation (2012) Measuring informality: a statistical manual on the informal sector and informal employment. International Labour Organisation, Geneva. (Available from `http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_182300.pdf`.)

Kish, L. and Frankel, M. R. (1974) Inference from complex samples. *J. R. Statist. Soc.* B, **36**, 1–22.

LaVange, L. M., Koch, G. G. and Schwartz, T. A. (2001) Applying sample survey methods to clinical trials data. *Statist. Med.*, **20**, 2609–2623.

Lee, E. S. and Forthofer, R. N. (2005) *Analyzing Complex Survey Data*. Thousand Oaks: Sage.

Lepkowski, J. M. (1989) The treatment of wave nonresponse in panel surveys. In *Panel Surveys* (eds D. Kasprzyk, G. Duncan, G. Kalton and M. Singh), ch. 5. New York: Wiley.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. Hoboken: Wiley.
Longford, N. T. (1993) *Random Coefficient Models*. Oxford: Clarendon.
Mincer, J. and Polachek, S. (1974) Family investment in human capital: earnings of women. *J. Polit. Econ.*, **82**, S76–S108.
de Passos, A. F., Ansiliero, G. and Paiva, L. H. (2005) Mercado de trabalho evolução recente e perspectivas. *Technical Report*. Instituto de Pesquisa Econômica Aplicada, Brasília.
Pessoa, D. and Silva, P. (1998) *Análise de Dados Amostrais Complexos*. São Paulo: Associação Brasileira de Estatística.
Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 317–337.
Pfeffermann, D. and La Vange, L. (1989) Regression models for stratified multi-stage cluster samples. In *Analysis of Complex Surveys* (eds C. Skinner, D. Holt and T. Smith), ch. 12. Chichester: Wiley.
Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc.* B, **60**, 23–40.
Rabe-Hesketh, S. and Skrondal, A. (2006) Multilevel modelling of complex survey data. *J. R. Statist. Soc.* A, **169**, 805–827.
Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. New York: Wiley.
Singer, J. D. and Willett, J. B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Occurrence*. New York: Oxford University Press.
Skinner, C. J. (1986) Design effects of two-stage sampling. *J. R. Statist. Soc.* B, **48**, 89–99.
Skinner, C. (1989a) Introduction to part A. In *Analysis of Complex Surveys* (eds C. Skinner, D. Holt and T. Smith), ch. 2. Chichester: Wiley.
Skinner, C. (1989b). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (eds C. Skinner, D. Holt and T. Smith), ch. 3. Chichester: Wiley.
Skinner, C. and Holmes, D. (2003) Random effects models for longitudinal survey data. In *Analysis of Survey Data* (eds R. Chambers and C. Skinner). Chichester: Wiley.
Skinner, C. and Vieira, M. (2007) Variance estimation in the analysis of clustered longitudinal survey data. *Surv. Methodol.*, **33**, 3–12.
Snijders, T. A. B. and Bosker, R. J. (1999) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks: Sage.
Vieira, M. and Skinner, C. (2008) Estimating models for panel survey data under complex sampling. *J. Off. Statist.*, **24**, 343–364.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.
Yang, M., Goldstein, H., Browne, W. and Woodhouse, G. (2002) Multivariate multilevel analyses of examination results. *J. R. Statist. Soc.* A, **165**, 137–153.