Misspecification Effects in the Analysis of Longitudinal Survey Data

Marcel de Toledo Vieira
Departamento de Estatística, Universidade Federal de Juiz de Fora, Brasil
marcel.vieira@ufjf.edu.br

M. Fátima Salgueiro
ISCTE Business School and UNIDE, Lisbon University Institute, Portugal
fatima.salgueiro@iscte.pt

Peter W. F. Smith
S3RI and University of Southampton, United Kingdom
p.w.smith@soton.ac.uk

Abstract

Misspecification effects (*meffs*) measure the inflation of the sampling variance of an estimator as a result of the use of complex sampling schemes. Many longitudinal social survey designs employ multi-stage sampling, leading to some clustering of the sample and to *meffs* greater than one. For a model for panel data we consider methods for estimating parameters which allow for complex schemes. An empirical study using longitudinal data from the British Household Panel Survey is conducted, and a simulation study is performed.

Keywords: parametric models; longitudinal data; sampling impacts.

1 Introduction

Standard inferential methods are often not valid when analysing data obtained using a complex sampling scheme. The interest in fitting models to longitudinal complex survey data has been growing in the last decade. Skinner and Vieira (2007) presented evidence that the variance-inflating impacts of clustering may be higher for longitudinal analyses than for the corresponding cross-sectional analyses. We further investigate the impact of weighting, stratification and clustering in the regression analysis of longitudinal survey data, comparing it with the impact on cross-sectional analyses.

In Section 2 we introduce the longitudinal survey data under analysis. Section 3 presents the model, point and variance estimation procedures, and describes measures of misspecification effects (*meff*s). The motivating application and empirical results are presented in Section 4 and a simulation study is performed in Section 5. Section 6 contains a discussion.

2 Data and Sampling Design

The empirical evidence presented in this paper is based on data from the British Household Panel Survey (BHPS), a household panel survey of individuals in private domiciles in Great Britain. The BHPS follows longitudinally a sample of individuals selected in 1991 by a complex stratified two-stage sampling scheme, with clustering by area. Our analyses are based on a subsample of 2255 men and women aged 16 or more, who were original sample members, who gave a full interview in waves twelve to fifteen, and who were employed throughout the period. The following variables are considered: gender; age category; number of children in the household; qualification; social class; marital status; health status; hours normally worked per week; and logarithm of the household income.

In our sample, the relative frequency for both gender categories is approximately 50%. The distribution of the age category variable is negatively skewed, as the frequencies for the older categories are larger. Most of the respondents are either married or living as a couple in 2002. Approximately 80% of the respondents considered themselves in either good or excellent health condition. Furthermore, over 75% of the individuals worked at least 30 hours per week. About 55% of the individuals had a high level of education, and only 16.32% of them occupied a partly skilled or an unskilled position in their last job. Almost 62% of the respondents had no children in the household where they live. Moreover, the average household income of the sample members was approximately GBP 3365 in the month before the interview was made.

3 Model, Estimation Procedures and *Meff*s

Regression models have found a wide range of useful applications with longitudinal survey data (e.g., Diggle *et al.* 2002; Vieira and Skinner, 2008; Vieira, 2009). Let y_{it} denote the response of interest for individual *i* at time *t*. Let $y_i = (y_{i1}, ..., y_{iT})'$ be the vector of repeated measures.

We consider linear models of the following form to represent the expectation of y_i given the values of covariates:

$$E(y_i) = x_i \beta , \qquad (1)$$

where $x_i = (x_{i1}',\ldots,x_{iT}')'$, x_{it} is a $1\times q$ vector of specified values of covariates for woman *i* at wave *t*, β is the $q\times 1$ vector of regression coefficients, and the expectation is with respect to the model.

Following the pseudo-likelihood approach (Skinner, 1989; Skinner and Vieira, 2007), the most general estimator of β we consider is

$$\hat{\beta} = \left(\sum_{i\in s} w_i x_i' V^{-1} x_i \right)^{-1} \sum_{i\in s} w_i x_i' V^{-1} y_i , \qquad (2)$$

where w_i is a longitudinal survey weight, *V* is a $T\times T$ estimated 'working' variance matrix of y_i (Diggle *et al.*, 2002), taken as the exchangeable variance matrix with diagonal elements $\hat{\sigma}^2$ and off-diagonal elements $\hat{\rho}\hat{\sigma}^2$. Further discussion on the estimation of β and ρ is presented in Skinner and Vieira (2007).

Under (1), $\hat{\beta}$ is approximately unbiased with respect to the model and the survey design and may still be expected to combine both within and between individual information in a reasonably efficient manner, even if the working model for the error structure does not hold exactly (Skinner and Vieira, 2007).

Without the weight terms and survey sampling considerations, the form of $\hat{\beta}$, given by (2), is motivated by the generalized estimating equations (GEE) approach of Liang and Zeger (1986), which we denote by $\hat{\beta}_n$.

The following estimator of the covariance matrix of $\hat{\beta}$ allows for a stratified multistage sampling scheme and it is based upon the classical method of linearization (Skinner, 1989; Skinner and Vieira, 2007)

$$v(\hat{\beta}) = \left[\sum_{i \in s} w_i x_i ' V^{-1} x_i\right]^{-1} \left[\sum_h n_h / (n_h - 1) \sum_a (z_{ha} - \overline{z}_h)(z_{ha} - \overline{z}_h)'\right] \left[\sum_{i \in s} w_i x_i ' V^{-1} x_i\right]^{-1}$$

where h denotes stratum, a denotes primary sampling unit (PSU), n_h is the number of PSUs in stratum h, $z_{ha} = \sum_i w_i x_i ' V^{-1} e_i$, $\overline{z}_h = \sum_a z_{ha} / n_h$ and $e_i = y_i - x_i \hat{\beta}$. If the weights, the sampling scheme and the difference between $n/(n - 1)$ and 1 are ignored, this estimator reduces to the 'robust' variance estimator presented by Liang and Zeger (1986).

We consider three further alternatives for estimating the covariance matrix of $\hat{\beta}$: (*i*) $v_a(\hat{\beta})$, which considers h =1 and therefore ignores stratification; (*ii*) $v_h(\hat{\beta})$, which considers a =1 and therefore ignores clustering; and (*iii*) $v_n(\hat{\beta})$, which considers h =1 and a =1 and therefore ignores both stratification and clustering. We also perform variance estimation for $\hat{\beta}_n$.

We are concerned with the potential bias of $v_a(\hat{\beta})$, $v_h(\hat{\beta})$, and $v_n(\hat{\beta})$, when in fact the design is complex. Skinner (1989) has proposed the *misspecification effect* (*meff*), which is designed to measure the effects of incorrect specification of both the sampling scheme and the considered model.

The effect of the complex sampling scheme on $v_a(\hat{\beta})$ and $v_h(\hat{\beta})$ can be evaluated if we examine the *meff's* distribution. We consider $meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)] = v(\hat{\beta}_k)/v_a(\hat{\beta}_k)$; $meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)] = v(\hat{\beta}_k)/v_h(\hat{\beta}_k)$; and $meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)] = v(\hat{\beta}_k)/v_n(\hat{\beta}_k)$, where $\hat{\beta}_k$ denote the k^{th} element of $\hat{\beta}$. The $meff_a$, $meff_h$, and $meff_n$ measure the impact of stratification, clustering, and both stratification and clustering, respectively. We also calculate all the considered versions of the *meff* measure for $\hat{\beta}_n$. Furthermore, $meff_g = v(\hat{\beta}_k)/v_n(\hat{\beta}_{nk})$ is calculated in order to access the bias caused by ignoring all the sampling scheme features.

4 Application

The paper is motivated by a regression analysis of four waves of BHPS data, which considers logarithm of the household income as the dependent variable. We first estimate *meffs* for the linearization estimator, considering $\hat{\beta}$, as discussed in Section 3.

Using data from just the first wave and setting $x_i = 1$, the estimated $meff_n$ for this cross-sectional mean is given in Table 1 as about 1.3. In order to evaluate the impact of the longitudinal aspect of the data, we estimated a series of each type of the *meffs* discussed above, using data for waves 12 to 15.

TABLE 1. *Meff* estimates for longitudinal means

Meff	Waves			
	12	12 and 13	12 to 14	12 to 15
$meff_a\left[\hat{\beta}_k, v_a\left(\hat{\beta}_k\right)\right]$	0.971	0.965	0.965	0.963
$meff_h\left[\hat{\beta}_k, v_h\left(\hat{\beta}_k\right)\right]$	1.490	1.653	1.699	1.695
$meff_n\left[\hat{\beta}_k, v_n\left(\hat{\beta}_k\right)\right]$	1.282	1.431	1.474	1.458
$meff_a\left[\hat{\beta}_{nk}, v_a\left(\hat{\beta}_{nk}\right)\right]$	0.969	0.963	0.961	0.960
$meff_h\left[\hat{\beta}_{nk}, v_h\left(\hat{\beta}_{nk}\right)\right]$	1.572	1.795	1.830	1.870
$meff_n\left[\hat{\beta}_{nk}, v_n\left(\hat{\beta}_k\right)\right]$	1.343	1.504	1.575	1.653
$meff_g$	1.494	1.598	1.778	1.706

Although these estimated *meffs* are subject to sampling error, there is a tendency for $meff_h$, $meff_n$, and $meff_g$ to increase with the number of waves. It therefore seems that it becomes more important to allow for clustering and for the complex sampling design in general when the number of waves in the analysis increases.

Furthermore, stratification effects appear to be constant with increases in the number of waves. When we included educational level as a covariate, we also noticed some evidence for $meff_h$, $meff_n$, and $meff_g$ to increase with the number of waves.

The model has been further elaborated by adding time, gender, age category, marital status, number of children in the household, social class, health status, and numbers of hours normally worked as covariates. Once more, we observed some evidence of a tendency for those *meffs* to diverge from one as the number of waves increases, at least for the coefficients of some of the covariates.

We also confirmed the observation of Skinner and Vieira (2007) that *meffs* for regression coefficients tend not to be greater than meffs for the means of the dependent variable.

5 Simulation Study

As results reported in Section 4 are subject to sampling error we have conducted a simulation study to evaluate the behaviour of the *meff* measures. Each of the d =1, …, D replicate samples is based on the BHPS data subset described above which is considered as the 'target population'. We evaluated the properties of variance estimators for unweighted point estimators and assessed only different impacts of clustering. We studied the *meff* when the number of waves in the analysis is increased. Note that we did not assess the impact of either stratification or unequal probability sampling.

Let y_{iat} be the value for the study variable for unit $i = 1,2,\ldots,n_d^{sim}$, in PSU $a = 1,\ldots,m_d^{sim}$, at wave t of the survey, where n_d^{sim} and m_d^{sim} are the sample size and the number of PSUs for the replicate sample d. For generating the values of y_{iat} for the simulation study, we used the following uniform correlation model, which allows for the impact of clustering:

$$y_{iat} = x_{iat}\beta + \eta_a + u_{ia} + v_{iat}, \qquad (3)$$

with $\eta_a \sim N(0,\sigma_\eta^2)$, $u_{ia} \sim N(0,\sigma_u^2)$, and $v_{iat} \sim N(0,\sigma_v^2)$. We consider the logarithm of the household income as the dependent variable and the remaining variables listed in Section 2 as covariates. We have held the values of the covariates as fixed.

The adopted the values for β, σ_η^2, σ_u^2, and σ_v^2 have been obtained by maximum likelihood estimation considering the 'target population'. In particular, we have considered different realistic choices for σ_η^2, $\sigma_\eta^2 = 0.06$ (actual value estimated from fitting (3)), $\sigma_\eta^2 = 0.12$, and $\sigma_\eta^2 = 0.18$ to enable the evaluation of effects of different impacts of clustering on the considered variance estimation procedures.

Let

$$\hat{E}(\,m\hat{e}ff\,) = \frac{1}{D}\sum_{d=1}^{D} m\hat{e}ff^{(d)} \,,$$

be the mean of our parameter of interest estimated over repeated simulation,

$$var(m\hat{e}ff) = \frac{1}{D-1}\sum_{d=1}^{D}\left[m\hat{e}ff^{(d)} - \hat{E}(m\hat{e}ff)\right]^2 ,$$

be a simulation estimator of $\mathrm{VAR}(m\hat{e}ff)$, the population variance of the misspecification effect measure, and

$$se[\,\hat{E}(m\hat{e}ff)\,] = \sqrt{var(m\hat{e}ff)/D}$$

the simulation standard error of $\hat{E}(m\hat{e}ff)$.

For the models that have been fitted to each generated replicate sample, we have set x_i = 1 and therefore we have still studied only the behavior of the *meff* for longitudinal means. Let n_a be the sample size for PSU *a* in the 'target population' and n_{da}^{sim} be the sample size for PSU *a* in the replicate sample *d*.

Table 2 presents results for three scenarios: (i) ($m_d^{sim} = 200$, $n_{da}^{sim} = n_a$, and $\sigma_a^2 = 0.35$); (ii) ($m_d^{sim} = 200$, $n_{da}^{sim} = n_a$, and $\sigma_a^2 = 0.70$); and (iii) ($m_d^{sim} = 200$, $n_{da}^{sim} = n_a$, and $\sigma_a^2 = 1.35$). Note that $m = 234$ in the 'target population'.

TABLE 2. $\hat{E}(\,m\hat{e}ff\,)$ and $se[\,\hat{E}(m\hat{e}ff)\,]$ (in brackets), for three scenarios.

n_j^{sim*}	σ_η^2	Waves			
		12	12 and 13	12 to 14	12 to 15
n_j	0.06	1.1901 (0.0044)	1.2077 (0.0046)	1.2115 (0.0047)	1.2143 (0.0047)
	0.12	1.2766 (0.0054)	1.3014 (0.0057)	1.3106 (0.0058)	1.3157 (0.0058)
	0.18	1.3624 (0.0066)	1.3933 (0.0069)	1.4061 (0.0070)	1.4118 (0.0070)

D=1000

The simulation results also give evidence that there is a tendency for the *meff* to increase as the number of waves in the analysis increases, at least for longitudinal means. This tendency seems to be stronger for larger clustering impacts. *Meff's* increase when the clustering impacts are increased, as expected from the survey sampling literature

(Vieira, 2009). Simulation standard errors of $\hat{E}(m\hat{e}ff)$ appear to increase when number of waves and clustering impacts are increased.

6 Discussion

We have presented evidence that clustering impacts may be stronger for longitudinal studies than for cross-sectional studies, and that *meffs* for the regression coefficients may increase with the number of waves considered in the analysis. The main implication of these findings is that standard errors in analysis of longitudinal survey data may be misleading if the initial sample was clustered and if this clustering is ignored. We have also observed that *meffs* for regression coefficients tend not to be greater than *meffs* for the means of the dependent variable.

Acknowledgments: The research of the first author was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) grant CEX-APQ-00467-2008. The research of the second author was supported by the Fundação para a Ciência e a Tecnologia grant PTDC/GES/72784/2006.

References

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

Liang, K. and Zeger, S. L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73: (1) 13-22.

Salgueiro, M. F. R. F., Smith, P. W. F. e Vieira, M. D. T. (2010) A Multi-Process Second-Order Latent Growth Curve Model for Subjective Well-Being. Submmitted to *Multivariate Behavioral Research*.

Skinner, C.J. (1989) Domain means, regression and multivariate analysis. In Skinner, C. J., Holt, D. and Smith, T. M. F. eds. *Analysis of Complex Surveys*. Chichester: Wiley, pp. 59-87.

Skinner, C.J. and Holmes, D. (2003). *Random Effects Models for Longitudinal Survey Data. Analysis of Survey Data*, R.L. Chambers and C.J. Skinner (eds). Chichester: Wiley.

Skinner, C. and Vieira, M. D. T. (2007) Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology*. 33: (1), 3-12.

Vieira, M. D. T. (2009). *Analysis of Longitudinal Survey Data*. 1. ed. Saarbrücken: VDM Verlag Dr. Müller.

Vieira, M. D. T. and Skinner, C. J. (2008) Estimating Models for Panel Survey Data under Complex Sampling. *Journal of Official Statistics*, 24, 343-364.