Variance estimation in the analysis of clustered longitudinal survey data

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 3-12
Statistics Canada, Catalogue No. 12-001-XPB

Variance estimation in the analysis of clustered longitudinal survey data

Chris Skinner and Marcel de Toledo Vieira [1]

Abstract

We investigate the impact of cluster sampling on standard errors in the analysis of longitudinal survey data. We consider a widely used class of regression models for longitudinal data and a standard class of point estimators of a generalized least squares type. We argue theoretically that the impact of ignoring clustering in standard error estimation will tend to increase with the number of waves in the analysis, under some patterns of clustering which are realistic for many social surveys. The implication is that it is, in general, at least as important to allow for clustering in standard errors for longitudinal analyses as for cross-sectional analyses. We illustrate this theoretical argument with empirical evidence from a regression analysis of longitudinal data on gender role attitudes from the British Household Panel Survey. We also compare two approaches to variance estimation in the analysis of longitudinal survey data: a survey sampling approach based upon linearization and a multilevel modelling approach. We conclude that the impact of clustering can be seriously underestimated if it is simply handled by including an additive random effect to represent the clustering in a multilevel model.

Key Words: Clustering; Design effect; Misspecification effect; Multilevel model.

1. Introduction

It is well known that it is important to take account of sample clustering when estimating standard errors in the analysis of survey data. Otherwise, standard error estimators can be severely biased. In this paper we investigate the impact of clustering in the regression analysis of longitudinal survey data and compare it with the impact on corresponding cross-sectional analyses. Kish and Frankel (1974) presented empirical work which suggested that the impact of complex designs on variances decrease for more complex analytical statistics and so one might conjecture that the impact on longitudinal analyses might also be reduced. We shall argue that, in fact, the impact of clustering on longitudinal analyses can tend to be greater, at least for a number of common types of analysis and for some common practical settings. An intuitive explanation is that some common forms of longitudinal analysis of individual survey data 'pool' data over time and enable much temporal 'random' variation in individual responses to be 'extracted' in the estimation of regression coefficients. In contrast, it may only be possible to extract much less variation in the effects of clustering since such clustering, representing geography for example, often tends to generate more stable effects than repeated measurements of individual behaviour. As a consequence the relative importance of clustering in standard errors can increase the more waves of data are included in the analysis.

In addition to considering the impact of clustering on variance estimation, we shall also consider the question of how to undertake the variance estimation itself. It is natural for many analysts to represent clustering via multilevel models (Goldstein 2003, Chapter 9; Renard and Molenberghs 2002) and we shall consider how variance estimation methods based upon such models compare with survey sampling variance estimation procedures in the case of cluster sampling.

There is a well established literature on methods for taking account of complex sampling schemes in the regression analysis of survey data. See *e.g.*, Kish and Frankel (1974), Fuller (1975), Binder (1983), Skinner, Holt and Smith (1989) and Chambers and Skinner (2003). We restrict attention here to 'aggregate' regression analyses (Skinner *et al.* 1989), where regression coefficients at the 'population level' are the parameters of interest, where suitable estimates of these coefficients may be obtained by adapting standard model-based procedures using survey weights and where the variances of these estimated regression coefficients may be estimated by linearization methods (Kish and Frankel 1974; Fuller 1975). In this paper, we extend this work to the case when longitudinal survey observations are obtained, based upon an initial sample drawn according to a complex sampling scheme, focussing again on the case of a clustered design. We consider a standard class of linear regression models for such longitudinal data, as considered in the biostatistical literature (*e.g.*, Diggle, Heagerty, Liang and Zeger 2002), the multilevel modelling literature (*e.g.*, Goldstein 2003) and the econometric literature (*e.g.*, Baltagi 2001). We consider an established class of point estimators of a generalized least squares type, modified by survey weighting. For some applications of such methods to survey data, see Lavange, Koch and Schwartz (2001); Lavange, Stearns, Lafata, Koch and Shah (1996).

1. Chris Skinner, University of Southampton, United Kingdom; Marcel de Toledo Vieira, Universidade Federal de Juiz de Fora, Brazil.

The impact of a complex sampling scheme on variance estimation will be measured by the 'misspecification effect', denoted meff (Skinner 1989a), which is the variance of the point estimator of interest under the actual sampling scheme divided by the expectation of a specified variance estimator. This is a measure of the relative bias of the specified variance estimator. If it is unbiased then the meff will be one. If the actual sampling scheme involves clustering but the specified variance estimator is 'misspecified' by ignoring the clustering, then the expectation of the variance estimator will usually be less than the actual variance and the meff will be greater than one. This concept is closely related to that of the 'design effect' or deff of Kish (1965), defined as the variance of the point estimator under the given design divided by its variance under simple random sampling with the same sample size, a concept more relevant to the choice of design than to the choice of standard error estimator.

We shall illustrate our theoretical arguments with analyses of data from the British Household Panel Survey (BHPS) on attitudes to gender roles, where the units of primary analytic interest are individual women and the clusters consist of postcode sectors, used as primary sampling units in the selection of the first wave sample from an address register.

The framework, including the models and estimation methods, is described in Section 2. The theoretical properties of the variance estimation methods are considered in Section 3. Section 4 illustrates these properties numerically, using an analysis of BHPS data. Some concluding remarks are provided in Section 5.

2. Regression model, data and inference procedures

Consider a finite population $U = \{1, ..., N\}$ of N units, assumed fixed across a series of occasions $t = 1, ..., T$. We shall refer to the units as individuals, although our discussion is applicable more generally. Let y_{it} denote the value of an outcome variable for individual $i \in U$ at occasion t and let $y_i = (y_{i1}, ..., y_{iT})'$ be the vector of repeated measurements. Let x_{it} denote a corresponding $1 \times q$ vector of values of covariates for individual i at occasion t and let $x_i = (x_{i1}', ..., x_{iT}')'$. We assume that the following linear model holds for the expectation of y_i conditional on $(x_1, ..., x_N)$:

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

where β is a $q \times 1$ vector of regression coefficients and the expectation is with respect to the model. We suppose that β is the target for inference, that is the regression coefficients are the parameters of primary interest to the analyst. Although we shall consider further features of this model, such as the covariance matrix of y_i, these will be assumed to be of secondary interest to the analyst.

The data available to make inference about β are from a longitudinal survey in which values of y_{it} and x_{it} are observed at each occasion (wave) $t = 1, ..., T$ for individuals i in a sample, s, drawn from U at wave 1 using a specified sampling scheme. For simplicity, we assume no non-response here, but return to this possibility in Section 4.

In order to formulate a point estimator of β, we extend the specification of (1) to the following 'working' model:

$$y_{it} = x_{it}\beta + u_i + v_{it}, \qquad (2)$$

where u_i and v_{it} are independent random effects with zero means and variances $\sigma_u^2 = \rho\sigma^2$ and $\sigma_v^2 = (1 - \rho)\sigma^2$ respectively, conditional on $(x_1, ..., x_N)$. This model may be called a uniform correlation model (Diggle *et al.* 2002, page 55) or a two-level model (Goldstein 2003). The parameter ρ is the intra-individual correlation.

The basic point estimator of β we consider is

$$\hat{\beta} = \left(\sum_{i \in s} w_i \, x_i' V^{-1} x_i \right)^{-1} \sum_{i \in s} w_i \, x_i' V^{-1} y_i, \qquad (3)$$

where w_i is a survey weight and V is a $T \times T$ estimated covariance matrix of y_i under the working model (2), *i.e.*, it has diagonal elements $\hat{\sigma}^2$ and off-diagonal elements $\hat{\rho}\,\hat{\sigma}^2$, where $(\hat{\rho}, \hat{\sigma}^2)$ is an estimator of (ρ, σ^2). (Note that in fact $\hat{\sigma}^2$ cancels out in (3) and hence σ^2 does not need to be estimated for $\hat{\beta}$). In the absence of the weight terms and survey considerations, the form of $\hat{\beta}$ is motivated by the generalized estimating equations (GEE) approach of Liang and Zeger (1986). The idea here is that $\hat{\beta}$, as a generalized least squares estimator of β, would be fully efficient if the working model (2) held. However, $\hat{\beta}$ remains consistent under (1) and may still be expected to combine within- and between-individual information in a reasonably efficient way even if the working model for the error structure does not hold exactly.

The survey weights are included in (3) following the pseudo-likelihood approach (Skinner 1989b) to ensure that $\hat{\beta}$ is approximately unbiased for β with respect to the model and the design, provided (1) holds.

There are a number of alternative ways of estimating ρ. In a non-survey setting, Liang and Zeger (1986) provide an iterative approach which alternates between estimates of β and ρ. Shah, Barnwell and Bieler (1997) describe how survey weights may be incorporated into this approach and implement this method in the REGRESS procedure of the software SUDAAN. By default, SUDAAN implements only one step of this iterative method and, in the non-survey setting, Lipsitz, Fitzmaurice, Orav and Laird (1994) conclude there is little to be lost by using only a single step.

For the working model in (2), the approach of Liang and Zeger (1986) to the estimation of β and ρ is virtually identical to the iterative generalized least squares (IGLS) estimation approach of Goldstein (1986). Both methods iterate between estimates of β and ρ and both use GLS to estimate β given the current estimate of ρ. The only slight difference is in the method used to estimate ρ. Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) show how to incorporate survey weights into the IGLS approach and their method may be expected to lead to very similar estimates of ρ to those in the SUDAAN REGRESS procedure. For the purposes of this paper, the precise form of $\hat{\rho}$ will not be critical and we may view $\hat{\beta}$ as either a weighted GEE or a weighted IGLS estimator.

We now turn to the estimation of the covariance matrix of $\hat{\beta}$ under the complex sampling scheme. We shall generally assume that a stratified multistage sampling scheme has been employed. We consider two main approaches to variance estimation.

Our first approach is the classical method of linearization (Skinner 1989b, page 78). The estimator of covariance matrix of $\hat{\beta}$ is

$$v(\hat{\beta}) = \left[\sum_{i\in s} w_i\, x_i' V^{-1} x_i\right]^{-1} \times \left[\sum_h n_h/(n_h-1)\sum_a (z_{ha}-\bar{z}_h)(z_{ha}-\bar{z}_h)'\right] \times \left[\sum_{i\in s} w_i\, x_i' V^{-1} x_i\right]^{-1} \quad (4)$$

where h denotes stratum, a denotes primary sampling unit (PSU), n_h is the number of PSUs in stratum h, $z_{ha} = \sum_i w_i\, x_i' V^{-1} e_i$, $\bar{z}_h = \sum_a z_{ha}/n_h$ and $e_i = y_i - x_i\hat{\beta}$. Similar estimators are considered by Shah *et al.* (1997, pages 8-9) and Lavange *et al.* (2001). 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).

Our second approach is more directly model-based. The model is first extended to represent the complex population underlying the sampling scheme and inference then takes place with respect to the extended model. We consider only the case of two-stage sampling from a clustered population, where the two-level model in (2) is extended to the three-level model (Goldstein 2003):

$$y_{ait} = x_{ait}\,\beta + \eta_a + u_{ai} + v_{ait}. \quad (5)$$

The additional subscript a denotes cluster and the additional random term η_a with variance σ_η^2 represents the cluster effect (assumed independent of u_{ai} and v_{ait}). We let σ_u^2 and σ_v^2 denote the variances of u_{ai} and v_{ait} respectively. Inference then takes place using IGLS, which may be weighted to avoid selection bias. This approach generates an estimated covariance matrix of the estimator of β directly. It should be noted, however that the estimator of β derived using weighted IGLS under model (5) may differ slightly from the estimator in (3) (although, for given estimates of the three variance components in (5), it will be the same as a weighted GEE estimator with a working covariance matrix based on this three-level model). Nevertheless, from our experience of social survey applications, such as in Section 4, and from theory (Scott and Holt 1982) the difference between these alternative point estimators will often be negligible.

Two broad approaches to deriving variance estimators from (5) are available. First, ignoring survey weights, the standard IGLS method (Goldstein 1986) may be employed, assuming that each random effect follows a normal distribution. Second, to avoid the assumption of normal homoscedastic random effects, a 'robust' variance estimation method (Goldstein 2003, page 80) may be employed. This approach is extended to handle survey weights in Pfeffermann *et al.* (1998). Leaving aside stratification, their variance estimator is identical to the linearization estimator in (4) for a given value of $\hat{\rho}$.

3. Properties of variance estimators

In this section we consider the properties of the estimators of the covariance matrix of $\hat{\beta}$ described in the previous section. We focus first on the linearization estimator $v(\hat{\beta})$ in (4).

The consistency of $v(\hat{\beta})$ for the covariance matrix of $\hat{\beta}$ follows established arguments in a suitable asymptotic framework (*e.g.*, Fuller 1975; Binder 1983). The one non-standard feature is the presence of V^{-1} in $\hat{\beta}$ and $v(\hat{\beta})$ and the dependence of V on $\hat{\rho}$. In fact, in large samples the covariance matrix of $\hat{\beta}$ depends on $\hat{\rho}$ only via its limiting value ρ^* (in a given asymptotic framework). To see this, write $\hat{\beta} - \beta = (\sum_s u_i)^{-1}\sum_s \tilde{z}_i$, where $u_i = w_i\, x_i' V^{-1} x_i$, $\tilde{z}_i = w_i\, x_i' V^{-1}\tilde{e}_i$ and $\tilde{e}_i = y_i - x_i\beta$. Note that, under weak regularity conditions (Fuller and Battese 1973, Corollary 3), the asymptotic distribution of $\hat{\beta} - \beta$ is the same as that of $\beta^* - \beta = (\sum_s u_i^*)^{-1}\sum_s z_i^*$, where $u_i^* = w_i\, x_i' V^{*-1} x_i$, $z_i^* = w_i\, x_i' V^{*-1}\tilde{e}_i$ and V^* takes the same form as V with $\hat{\rho}$ replaced by $\rho^* = p\lim(\hat{\rho})$, the probability limit of $\hat{\rho}$ in the asymptotic framework. Writing $\bar{z}^* = \sum_s z_i^*/n$ and $\bar{U} = p\lim(\sum_s u_i^*/n)$, we may thus approximate the covariance matrix of $\hat{\beta}$ asymptotically by $\mathrm{var}(\hat{\beta}) \approx \bar{U}^{-1}\,\mathrm{var}(\bar{z}^*)\bar{U}^{-1}$. If the working model (2) holds then $\rho^* = \rho$ and this covariance matrix will be the same for any consistent method of estimating ρ. Even if the working model does not hold, $v(\hat{\beta})$ will be consistent for $\bar{U}^{-1}\,\mathrm{var}(\bar{z}^*)\bar{U}^{-1}$ within the kinds of asymptotic frameworks considered by

Fuller (1975) and Binder (1983) and under the kinds of regularity conditions they and Fuller and Battese (1973) set out.

We next explore the impact on the linearization method of ignoring a complex sampling design. We denote by $v_0(\hat{\beta})$ the linearization estimator obtained from expression (4) by ignoring the design, *i.e.*, by assuming only a single stratum with PSUs identical to individuals so that $n_h = n$ is the overall sample size and z_{ha} is replaced by $z_i = w_i x_i' V^{-1} e_i$. We shall be concerned with the bias of $v_0(\hat{\beta})$ when in fact the design is complex. Let $\hat{\beta}_k$ denote the k^{th} element of $\hat{\beta}$ and let $v_0(\hat{\beta}_k)$ denote the k^{th} element of $v_0(\hat{\beta})$. Then, following Skinner (1989a, page 24), we shall measure the relative bias of the 'incorrectly specified' variance estimator $v_0(\hat{\beta}_k)$ as an estimator of $\text{var}(\hat{\beta}_k)$ by the *misspecification effect*, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \text{var}(\hat{\beta}_k)/E[v_0(\hat{\beta}_k)]$. Since $v(\hat{\beta}_k)$ is a consistent estimator of $\text{var}(\hat{\beta}_k)$, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ may be estimated by $v(\hat{\beta}_k)/v_0(\hat{\beta}_k)$ and is closely related to the idea of design effect.

To investigate the nature of $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$, we first write:

$$v_0(\hat{\beta}) = \left(\sum_s u_i\right)^{-1} [n/(n-1)] \times \left[\sum_s (z_i - \bar{z})(z_i - \bar{z})'\right]\left(\sum_s u_i\right)^{-1}, \qquad (6)$$

where $\bar{z} = \sum_s z_i / n$. Then, as an asymptotic approximation, we have $E[v_0(\hat{\beta})] \approx \bar{U}^{-1}[n^{-1}S_z^*]\bar{U}^{-1}$, where S_z^* is the probability limit of the finite population covariance matrix of z_i^*. Using the fact that the numerator of $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ may be approximated by $\bar{U}^{-1}\,\text{var}(\bar{z}^*)\bar{U}^{-1}$, we can thus write:

$$\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \frac{(\bar{U}^{-1})_k \,\text{var}(\bar{z}^*)\,(\bar{U}^{-1})_k'}{(\bar{U}^{-1})_k [n^{-1}S_z^*](\bar{U}^{-1})_k'}, \qquad (7)$$

where $(\bar{U}^{-1})_k$ is the k^{th} row of $\bar{U}^{-1}$. This simplifies in the case $q = 1$ to:

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = \text{var}(\bar{z}^*)/[n^{-1}S_z^*]. \qquad (8)$$

We may explore more specific forms of these expressions under different models and assumptions about the weights and the sampling scheme. We focus here on the impact of clustering, assuming equal weights and no stratification. Consider the three-level model in (5) and, to simplify matters, suppose that $q = 1$ and $x_{ait} \equiv 1$ and β is the mean of y_{ait}. Then, straightforward algebra shows that the value of z_i^* for individual i within cluster a is $[1 + \rho^*(T-1)]^{-1}\sum_t(\eta_a + u_{ai} + v_{ait})$. Now suppose that two-stage sampling is employed with a common sample size m per cluster. Then, evaluating the variance $\text{var}(\bar{z}^*)$ and probability limit S_z^* in (8) with respect to the model in (5), we find, in a similar manner to Skinner (1989a, page 38):

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = 1 + (m-1)\tau, \qquad (9)$$

where $\tau = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2/T)$ is the intracluster correlation of z_i^*. We see that, under this model, the meff increases as T increases (provided $\sigma_v^2 > 0$) and thus the impact of clustering on variance estimation is greater in the longitudinal case than for the cross-sectional problem (where $T = 1$).

This finding depends on the rather strong assumption that the cluster effects η_a are constant over time. In fact, (9) still holds if we replace η_a by a time-varying effect η_{at} provided we replace τ by $\tau = \text{var}(\bar{\eta}_a)/[\text{var}(\bar{\eta}_a) + \sigma_u^2 + \sigma_v^2/T]$, where $\bar{\eta}_a = \sum_t \eta_{at}/T$. Now, the meff will increase as T increases if (and only if) $\sigma_u^2 + \sigma_v^2/T$ decreases faster with T than $\text{var}(\bar{\eta}_a)$. Whether this is the case will depend on the particular application. However, we suggest that for many longitudinal surveys of individuals with area-based clusters (the kind of setting we have in mind), this condition is plausible. In such applications we may often expect σ_v^2 to be large relative to σ_u^2 (*i.e.*, for the cross-sectional intracluster correlation to be small) in particular as a result of wave-specific measurement error and thus for $\sigma_u^2 + \sigma_v^2/T$ to decrease fairly rapidly as T increases. The socio-economic characteristics of areas may often be expected to be more stable and only in unusual situations might we expect measurement error to lead to much occasion-specific variance in η_{at}. Thus, we suggest that the ratio of $\text{var}(\bar{\eta}_a)$ for $T = 5$, say, compared to $T = 1$ may in such applications usually be expected to be greater than $(\sigma_u^2 + \sigma_v^2/5)/(\sigma_u^2 + \sigma_v^2)$ which will approach 1/5 as σ_u^2/σ_v^2 approaches 0. We thus suggest that in many practical circumstances it will be more important to allow for clustering in longitudinal analyses than in corresponding cross-sectional analyses. An empirical illustration is provided in Section 4.

We now consider the properties of variance estimators based upon the three-level model in (5). We consider only the approach based upon the assumption of normally distributed homoscedastic random effects, ignoring survey weights, given the (virtual) equivalence of the 'robust' multilevel approach and linearization.

If model (5) is correct and we can indeed ignore survey weights then the model-based variance estimator will be consistent (Goldstein 1986). However, as discussed in Skinner (1989b, page 68) and supported by theory in Skinner (1986), the main feature of clustering likely to impact on the standard errors of estimated regression coefficients is the variation in regression coefficients between clusters. This is not allowed for in model (5).

To see how model (5) may fail to capture the effects of clustering adequately, consider the cross-sectional case ($T = 1$) where x is scalar. Then, if the three-level model (5) holds, an approximate expression for the meff of the variance estimator of $\hat{\beta}$ based upon the two-level model (2) is:

$$\text{meff} = 1 + (m-1)\tau_1\tau_x, \qquad (10)$$

where $\tau_1 = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2)$ and τ_x is the intracluster correlations for x (Scott and Holt 1982; Skinner 1989b, page 68). This result extends in the longitudinal case, to:

$$1 \leq \text{meff} \leq 1 + (m-1)\tilde{\tau}\tau_z, \qquad (11)$$

where $\tilde{\tau}$ is the long-run ($T = \infty$) version of τ (see Appendix) and τ_z is an intracluster correlation coefficient for $z_{ai} = \sum_t x_{ait}/T$. The proof of this result and the simplifying assumptions required are sketched in the Appendix. The main point is that both $\tilde{\tau}$ and τ_z will often be small in which case $\tilde{\tau}\tau_z$ will be very small and thus meff may be implausibly close to one with the model-based variance estimator being subject to downward bias. We explore this empirically in Section 4. Of course, random coefficients could be introduced into model (5) and we consider this also in Section 4. However, given the difficulty of specifying a correct random coefficient model, this approach does not seem likely to be very robust.

Our focus in this section has so far been on the potential bias (or inconsistency) of variance estimation methods. It is also desirable to consider their efficiency. In particular, the linearization method may be expected to be less efficient than model-based variance estimation if the model is correct. The relative importance of efficiency vs. bias may be expected to increase as the number of clusters decreases. Wolter (1985, Chapter 8) summarises a number of simulation studies investigating both the bias and variance of the linearization variance estimator and these studies suggest that the linearization method performs well even with few clusters. Possible degrees of freedom corrections to confidence intervals for regression coefficients based upon the linearization method with small numbers of clusters are discussed by Fuller (1984). A simulation study of estimators for multilevel models in Maas and Hox (2004) does not suggest that the linearization method performs noticeably worse than the model-based approach, in terms of the coverage of confidence intervals for coefficients in β, even with as few as 30 clusters.

4. Example: Regression analysis of BHPS data on attitudes to gender roles

We now present an application to BHPS data to illustrate some of the theoretical properties discussed in the previous section.

Recent decades have witnessed major changes in the roles of men and women in the family in many countries. Social scientists are interested in the relation between changing attitudes to gender roles and changes in behaviour, such as parenthood and labour force participation (*e.g.*, Morgan and Waite 1987; Fan and Marini 2000). A variety of forms of statistical analysis are used to provide evidence about these relationships. Here, we consider estimating a linear model of form (1), with a measure of attitude to gender roles as the outcome variable, y, following an analysis of Berrington (2002).

The data come from waves 1, 3, 5, 7 and 9 (collected in 1991, 1993, 1995, 1997, and 1999 respectively) of the BHPS and these waves are coded $t = 1, ..., T = 5$ respectively. Respondents were asked whether they 'strongly agreed', 'agreed', 'neither agreed nor disagreed', 'disagreed' or 'strongly disagreed' with a series of statements concerning the family, women's roles, and work out of the household. Responses were scored from 1 to 5. Factor analysis was used to assess which statements could be combined into a gender role attitude measure. The attitude score, y_{it}, considered here is the total score for six selected statements for woman i at wave t. Higher scores signify more egalitarian gender role attitudes. Berrington (2002) provides further discussion of this variable. A more sophisticated analysis might include a measurement error model for attitudes (*e.g.*, Fan and Marini 2000), with each of the five-point responses to the six statements treated as ordinal variables. Here, we adopt a simpler approach, treating the aggregate score y_{it} and the associated coefficient vector β as scientifically interesting, with the measurement error included in the error term of the model.

Covariates for the regression analysis were selected on the basis of discussion in Berrington (2002) but reduced in number to facilitate a focus on the methodological issues of interest. The covariate of primary scientific interest is economic activity, which distinguishes in particular between women who are at home looking after children (denoted 'family care') and women following other forms of activity in relation to the labour market. Variables reflecting age and education are also included since these have often been found to be strongly related to gender role attitudes (*e.g.*, Fan and Marini 2000). All these covariates may change values between waves. A year variable (scored 1, 3, …, 9) is also included. This may reflect both historical change and the general ageing of the women in the sample.

The BHPS is a household panel survey of individuals in private domiciles in Great Britain (Taylor, Brice, Buck and Prentice-Lane 2001). The initial (wave one) sample in 1991 was selected by a stratified multistage design in which households had approximately equal probabilities of inclusion. The households were clustered into 250 primary

sampling units (PSUs), consisting of postcode sectors. All resident members aged 16 or over were selected in sample households. All adults selected at wave one were followed from wave two onwards and represent the longitudinal sample. The survey is subject to attrition and other forms of wave non-response. To handle this non-response, we have simply replaced s in (3) by the 'longitudinal sample' of individuals for which observations are available for each of $t = 1, ..., T$ and have chosen not to apply any survey weighting since our aim is to study potential misspecification effects associated with clustering and we wish to avoid confounding these with weighting effects. We also ignore the impact of stratification in the numerical work in this section (but see Section 5 for some comments on the effect of weights and stratification).

Given the analytic interest in whether women's primary labour market activity is 'caring for a family', we define our study population as women aged 16-39 in 1991. Thus our data consist of the longitudinal sample of women in the eligible age range for whom full interview outcomes (complete records) were obtained in all five waves, a sample of $n = 1{,}340$ women. These women are spread fairly evenly across 248 postcode sectors. The small average sample size of around five per postcode sector combined with the relatively low intra-postcode sector correlation for the attitude variable of interest leads to relatively small impacts of the design, as measured by meffs. Since our aims are methodological ones, we have chosen to group the postcode sectors into 47 geographically contiguous clusters, to create sharper comparisons, less blurred by sampling errors which can be appreciable in variance estimation. The meffs in the tables we present therefore tend to be greater than they are for the actual design. The latter results tend to follow similar patterns, although the patterns are less clear-cut as a result of sampling error.

We first estimate meffs for the linearization estimator, as discussed at the beginning of Section 3. Using data from just the first wave and setting $x_{ait} \equiv 1$, the estimated meff for this cross-sectional mean is given in Table 1 as about 1.5. This value is plausible since, if we make the usual approximation of (9) for unequal sample cluster sizes by replacing m by $\bar{m}$, the average sample size per cluster, we find that $1 + (\bar{m} - 1)\tau = 1.5$ and $\bar{m} = 1{,}340/47 \approx 29$ imply a value of τ of about 0.02 and such a small value is in line with other estimated values of τ found for attitudinal variables in British surveys (Lynn and Lievesley 1991, Appendix D).

Table 1 Estimates for longitudinal means

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3,5	1-7	1-9
	19.83	0.12	1.51	1.50	1.68	1.81	1.84

To assess the impact of the longitudinal aspect of the data, we estimated a series of meffs using data for waves $1, \ldots, t$ for $t = 2, 3, \ldots, 5$. Although these estimated meffs are subject to sampling error, there seems clear evidence in Table 1 of a tendency for the meff to increase with the number of waves. This trend might be anticipated from the theoretical discussion in Section 3 if the average level of egalitarian attitudes in an area varies less from year to year than the attitude scores of individual women. This seems plausible since the latter will be affected both by measurement error and genuine changes in attitudes, so that $\text{var}(\bar{\eta}_a)$ may be expected to decline more slowly with T than $\text{var}(\bar{u}_a + \bar{v}_a)$. We may therefore expect τ, and consequently the meff, to increase as T increases, as we observe in Table 1.

We next elaborate the analysis by including indicator variables for economic activity as covariates. The resulting regression model has an intercept term and four covariates representing contrasts between women who are employed full-time and women in other categories of economic activity. The estimated meffs are presented in Table 2. The intercept term is a domain mean and standard theory for a meff of a mean in a domain cutting across clusters (Skinner 1989b, page 60) suggests that it will be somewhat less than the meff for the mean in the whole sample, as indeed is observed with the meff for the cross-section domain mean of 1.13 in Table 2 being less than the value 1.51 in Table 1. As before, there is some evidence in Table 2 of tendency for the meff to increase, from 1.13 with one wave to 1.50 with five waves, albeit with lower values of the meffs than in Table 1. The meffs for the contrasts in Table 2 vary in size, some greater than and some less than one. These meffs may be viewed as a combination of the traditional variance inflating effect of clustering in surveys together with the variance reducing effect of blocking in an experiment. Such variance reduction arises if the domains being contrasted share a common cluster effect (of the form η_a in model (5)) which tends to cancel out in the contrasts, implying that the actual variance of the contrast is lower than the expectation of the variance estimator which assumes independence between domains. The latter expectation will be inflated by common cluster effects. The main feature of these results of interest here is that there is again no tendency for the meffs to converge to one as the number of waves increases. If there is a trend, it is in the opposite direction. For the contrast of particular scientific interest, that between women who are full-time employed and those who are 'at home caring for a family', the meff is consistently well below one.

We next refine the model further by including, as additional covariates, age group, year and qualifications. The estimated meffs are given in Table 3. The meffs for the regression coefficients corresponding to categories of

economic activity again vary, some being above one and some below one, for the same reasons as for the contrasts (which may also be interpreted as regression coefficients) in Table 2. There is again some evidence of a tendency for these meffs to diverge away from one as the number of waves increases. A comparison of Tables 1 and 3 confirms the observation of Kish and Frankel (1974) that meffs for regression coefficients tend not to be greater than meffs for the means of the dependent variable.

Table 2 Estimates for regression with covariates defined by economic activity

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3, 5	1-7	1-9
Intercept	20.58	0.11	1.13	1.01	1.09	1.38	1.50
Contrasts for							
PT employed	-1.03	0.10	0.93	0.91	0.93	1.00	0.89
Other inactive	-0.80	0.15	0.60	0.96	0.68	0.76	0.81
FT student	0.41	0.24	1.10	1.32	1.14	1.48	1.44
Family care	-2.18	0.10	0.72	0.49	0.58	0.66	0.60

Note: a) intercept is mean for women full-time employed
b) contrasts are for other categories of economic activity relative to full-time employed

Table 3 Estimates for regression coefficients with additional covariates in model

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3, 5	1-7	1-9
Intercept	20.20	0.30	0.95	0.87	0.87	1.04	1.07
Year, *t*	-0.04	0.01	-	0.86	0.69	0.59	0.96
Age Group							
16-21	0.00	-					
22-27	-0.71	0.25	1.22	1.37	1.44	1.73	1.64
28-33	-0.89	0.27	1.38	1.40	1.46	1.68	1.59
34+	-1.03	0.27	0.94	1.10	1.13	1.26	1.34
Economic Activity							
FT employed	0.00	-					
PT employed	-0.93	0.10	0.97	0.95	0.96	1.06	0.91
Other inactive	-0.75	0.15	0.60	0.96	0.68	0.77	0.81
FT student	0.17	0.24	0.93	1.32	1.23	1.39	1.32
Family care	-2.09	0.10	0.77	0.59	0.70	0.78	0.67
Qualification							
Degree	0.00	-					
QF	-0.52	0.21	0.77	0.64	0.75	0.87	0.85
A-level	-0.61	0.24	0.98	0.87	0.94	0.94	1.01
O-level	-0.44	0.20	0.62	0.62	0.59	0.69	0.73
Other	-1.16	0.22	0.83	0.83	0.78	0.80	0.82

We next consider model-based standard errors obtained from the three level model in (5), as discussed in section 2. The results are given in Table 4 in the column headed '3 level model-based'. For comparison, we also estimate the standard errors under the two level model in (2) - the results are in the column headed '2 level model-based'. The estimates in the two columns are virtually identical. There is a single digit difference in the third decimal place for some coefficients and slightly greater difference for the intercept term. We suggest that this is evidence that simply adding in a random area effect term can seriously understate the impact of clustering on the standard errors of the estimated regression coefficients. This evidence is in line with the theoretical upper bound for the meff in (11). The estimated value of $\tilde{\tau}$ in (11) is 0.019 and none of the covariates may be expected to display important intra-area correlation so the expected values of the variance estimators for the two-level and three-level models would be expected to be very close.

We suggested in Section 3 that the main feature of clustering likely to impact on the covariance matrix of $\hat{\beta}$ is the variation in regression coefficients between clusters. We have explored this idea by introducing random coefficients in the model. Treating the elements of β now as the expected values of the random coefficients, we found that the estimates of β were hardly changed. We found that the estimated standard errors of these estimates were indeed inflated, much more so than from the introduction of the extra cluster random effect in model (5), and that the inflation was of an order similar to those of the meffs in Tables 2 and 3. Nevertheless, the IGLS method did lead to several negative estimates of the variances of the random coefficients, raising issues of which coefficients to allow to vary or more generally the issue of model specification. This problem is accentuated with increasing numbers of covariates, as the number of parameters in the covariance matrix of the coefficient vector increases with the square of the number of covariates. Overall, the inclusion of random coefficients seems to raise at least as many problems as it solves, if the clustering is not of intrinsic scientific interest, and thus does not seem a very satisfactory way to allow for clustering in variance estimation. It is simpler to change the method of variance estimation.

As mentioned at the end of Section 2, one alternative is a 'robust' variance estimation method based on the model in (5) (Goldstein 2003, page 80). Values of such robust standard error estimates are also included in Table 4. As anticipated in Section 2, the robust standard error estimator for the two level model performs very similarly to the linearization estimator which ignores clustering. The robust standard error estimator for the three level model performs very similarly to the linearization estimator which allows for two stage sampling. The slight differences reflect the differences between the methods of estimating V.

Table 4 Estimated standard errors of regression coefficients

	Linearization		Multilevel modelling			
	SRS	complex	2 level model-based	2 level robust	3 level model-based	3 level robust
Intercept	0.287	0.296	0.253	0.288	0.259	0.293
Year, *t*	0.014	0.014	0.013	0.014	0.013	0.014
Age Group						
16-21						
22-27	0.191	0.245	0.155	0.192	0.155	0.243
28-33	0.214	0.270	0.187	0.215	0.187	0.266
34+	0.237	0.275	0.218	0.238	0.218	0.271
Economic Activity						
FT employed						
PT employed	0.103	0.098	0.098	0.103	0.098	0.096
Other inactive	0.166	0.150	0.146	0.166	0.146	0.148
FT student	0.207	0.238	0.199	0.207	0.199	0.236
Family care	0.125	0.102	0.112	0.125	0.112	0.101
Qualification						
Degree						
QF	0.228	0.210	0.207	0.228	0.208	0.211
A-level	0.238	0.239	0.209	0.240	0.210	0.237
O-level	0.234	0.199	0.217	0.235	0.218	0.199
Other	0.247	0.224	0.229	0.249	0.230	0.223

The linearization method in the presence of two-stage sampling is thus very close to robust variance estimation methods used in the literature on multilevel modeling. The distinction between the methods becomes stronger if we allow also for stratification and weighting. Another distinction is that in the multilevel modeling approach, differences between model-based and the robust standard errors might be used as a diagnostic tool to detect departures from the model (Maas and Hox 2004). For example, the large differences in the three-level standard errors for the coefficients of age group in Table 4 might lead to consideration of the inclusion of random coefficients for age group. This contrasts with the survey sampling approach where the error structure in model (5) is only treated as a working model and it is not necessarily expected that standard errors based upon this model will be approximately valid.

5. Discussion

We have presented some theoretical arguments and empirical evidence that the impact of ignoring clustering in standard error estimation for certain longitudinal analyses can tend to be larger than for corresponding cross-sectional analyses. The implication is that it is, in general, at least as important to allow for clustering in standard errors for longitudinal analyses as for cross-sectional analyses and that the findings of, for example, Kish and Frankel (1974), should not be used as grounds to ignore complex sampling in the former case.

The longitudinal analyses considered in this paper are of a certain kind and we should emphasise that the patterns observed for meffs in these kinds of analyses may well not extend to all kinds of longitudinal analyses. To speculate about the class of models and estimators for which the patterns observed in this paper might apply, we conjecture that increased meffs for longitudinal analyses will arise when the longitudinal design enables temporal 'random' variation in individual responses to be extracted from between-person differences and hence to reduce the component of standard errors due to these differences, but provides less 'explanation' of between cluster differences, so that the relative importance of this component of standard errors becomes greater.

The empirical work presented in this paper has also been restricted to the impact of clustering. We have undertaken corresponding work allowing for weighting and stratification and found broadly similar findings. Stratification tends to have a smaller effect than clustering. The sample selection probabilities in the BHPS do not vary greatly and the impact of weighting by the reciprocals of these probabilities on both point and variance estimates tends not to be large. There is rather greater variation among the longitudinal weights which are provided with BHPS data for analyses of sets of individuals who have responded at each wave up to and including a given year T. The impact

of these weights on point and variance estimates is somewhat greater. As T increases and further attrition occurs, the longitudinal weights tend to become more variable and lead to greater inflation of variances. This tends to compound the effect we have described of meffs increasing with T.

Leaving aside consideration of stratification and weighting, we have compared two approaches to allowing for cluster sampling. We have treated the survey sampling approach as a benchmark. We have also considered a multilevel modelling approach to allow for clustering. We have suggested that the use of a simple additive random effect to represent clustering can seriously understate the impact of clustering and may lead to underestimation of standard errors. If the clustering is of scientific interest, one solution would be to consider including random coefficients. Another would be to use the 'GEE2' approach (Liang, Zeger and Qaqish 1992) and specify an additional parametric model for $E(y_i\, y_i')$. If the clustering is treated as a nuisance, simply reflecting administrative convenience in data collection, we suggest the survey sampling approach has a number of practical advantages. This is discussed further by Lavange *et al.* (1996, 2001) in relation to other applications to repeated measures data.

Appendix

Justification for (11)

For simplicity, x and β are taken to be scalar, $\hat{\beta}$ is taken to be the ordinary least squares estimator and it is assumed that the sample sizes within clusters are all equal to m. The meff in (11) is defined as $\mathrm{var}_3(\hat{\beta})/E_3[v_2(\hat{\beta})]$, where E_3 and var_3 are moments with respect to the three-level model in (5) and $v_2(\hat{\beta})$ is a variance estimator based upon the two-level model in (2). Under (5) we obtain

$$\mathrm{var}_3(\hat{\beta}) = \left(\sum_{cit} x_{cit}^2\right)^{-2}\left(\sigma_\eta^2\sum_c x_{c++}^2 + \sigma_u^2\sum_{ci} x_{ci+}^2 + \sigma_v^2\sum_{cit} x_{cit}^2\right),$$

where + denotes summation across a suffix, $\sigma_\eta^2, \sigma_u^2$ and σ_v^2 are the respective variances of η_a, u_{ai} and v_{ait} and x_{cit} is centred at 0. We further suppose that $v_2(\hat{\beta})$ is defined so that $E[v_2(\hat{\beta})] \approx (\sum_{cit} x_{cit}^2)^{-2}[(\sigma_\eta^2+\sigma_u^2)\sum_{ci} x_{ci+}^2 + \sigma_v^2\sum_{cit} x_{cit}^2]$. After some algebra we may show that

$$\mathrm{meff} = 1+(m-1)\tilde{\tau}\tau_z\rho[1+(T-1)\tau_x]/[1+(T-1)\rho\tau_x], \quad (12)$$

where $\tilde{\tau} = \sigma_\eta^2/(\sigma_\eta^2+\sigma_u^2)$, $\rho = (\sigma_\eta^2+\sigma_u^2)/(\sigma_\eta^2+\sigma_u^2+\sigma_v^2)$, $\tau_x = \sigma_{xB}^2/\sigma_x^2$, $\sigma_x^2 = \sum_{cit} x_{cit}^2/(nT)$, $\sigma_{xB}^2 = [\sum_{ci}(x_{ci+}/T)^2/n - \sigma_x^2/T]/[1-1/T]$, $\tau_z = \sigma_{zB}^2/\sigma_z^2$, $\sigma_z^2 = \sum_{ci} z_{ci}^2/n$, $\sigma_{zB}^2 = [\sum_c (z_{c+}/m)^2/C - \sigma_z^2/m]/[1-1/m]$ and $n = Cm$ is the sample size. Note that $\tilde{\tau}\rho = \tau_1$ and, when $T = 1$, $\tau_z = \tau_x$ so that (12) reduces to (10). In general $\rho \le 1$ and (11) follows from (12). In fact, we estimate ρ as 0.59 in our application so the bound in (11) is not expected to be very tight.

Acknowledgements

The research of the second author was supported by grant 20.0286/01.3 from the Brazilian National Council for Scientific and Technological Development (CNPq).

References

Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*. 2nd Ed. Chichester: John Wiley & Sons, Inc.

Berrington, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study. In *Meaning and Choice*: *Value Orientations and Life Course Decisions*, (Ed., R. Lesthaeghe) Brussels: NIDI.

Chambers, R.L., and Skinner, C.J. Eds. (2003). *Analysis of Survey Data*. Chichester: John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-92.

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

Fan, P.-L., and Marini, M.M. (2000). Influences on gender-role attitudes during the transition to adulthood. *Social Science Research*, 29, 258-283.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*. Vol. 37, Series C, 117-132.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.

Fuller, W.A., and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 74, 430-431.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Ed. London: Arnold.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

Lavange, L.M., Koch, G.G. and Schwartz, T.A. (2001). Applying sample survey methods to clinical trials data. *Statistics in Medicine*, 20, 2609-23.

Lavange, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G. and Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, 5, 311-329.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society*, Series B, 54, 3-40.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 50, 270-278.

Lynn, P., and Lievesley, D. (1991). *Drawing General Population Samples in Great Britain*. London: Social and Community Planning Research.

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 and Data Analysis*, 46, 427-440.

Morgan, S.P., and Waite, L.J. (1987). Parenthood and the attitudes of young adults. *Am. Sociological Review*, 52, 541-547.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-56.

Renard, D., and Molenberghs, G. (2002). Multilevel modelling of complex survey data. In *Topics in Modelling Clustered Data* (Eds., M. Aerts, H. Geys, G. Molenberghs and L.M. Ryan). Boca Raton: Chapman and Hall/CRC. 263-272.

Scott, A.J., and Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

Shah, B.V., Barnwell, B.G. and Bieler, G.S. (1997). SUDAAN User's manual, release 7.5. Research triangle park, NC: Research Triangle Institute.

Skinner, C.J. (1986). Design effects of two stage sampling. *Journal of the Royal Statistical Society*, Series B, 48, 89-99.

Skinner, C.J. (1989a). Introduction to Part A. In *Analysis of Complex Surveys*. (Eds., C.J. Skinner, D. Holt and T.M.F. Smith) Chichester: John Wiley & Sons, Inc. 23-58.

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

Skinner, C.J., Holt, D. and Smith, T.M.F. Eds. (1989). *Analysis of Complex Surveys*. Chichester: John Wiley & Sons, Inc.

Taylor, M.F. ed, Brice, J., Buck, N. and Prentice-Lane, E. (2001). *British Household Panel Survey - User Manual - Volume A*: *Introduction*, *Technical Report and Appendices*. Colchester, University of Essex.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.