DESIGN-ADJUSTED ESTIMATION WITH REPEATED SURVEY DATA

R. L. CHAMBERS

1. THE BASIC PROBLEM

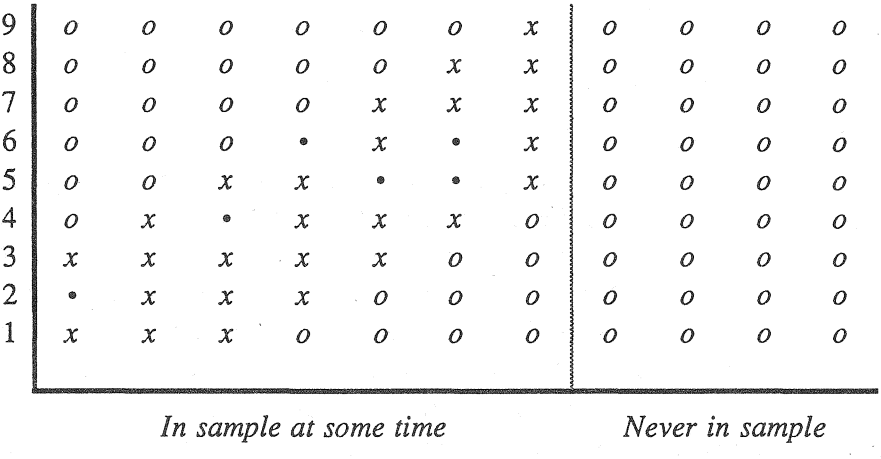
Consider a sequence of sample surveys of the same target population. The primary aim of these surveys is to provide 'snapshot' information on the status of this population at regular intervals (e.g. annual, quarterly, monthly). For example, the Australian Bureau of Statistics carries out monthly surveys of the Australian population to determine labour force participation rates, and quarterly surveys of Australian businesses to determine investment and expenditure patterns, while the Australian Bureau of Agricultural and Resource Economics carries out annual surveys of Australian farms to measure their economic performance.

A characteristic of all of these surveys is that they employ complex, highly stratified sample designs, along with partial sample rotation, both as a means of controlling sample response burden, as well as a means of increasing both efficiency of estimation at each time period as well as estimation of movement between time periods. Another characteristic of these surveys is that they are, to a greater or lesser extent, affected by sample non-response at each time period. Assuming a fixed underlying population over the time period of interest, the typical data structure generated by these repeated surveys can be set out as in Figure 1. Here x denotes an observed sample datum, • denotes a missing sample datum and o denotes an unobserved (non-sample) datum:

However, the most important characteristic of these repeated surveys is that they provide researchers with the opportunity to analyse a longitudinal data set for the target population, inasmuch as information on the same set of variables is obtained from each population unit surveyed at each point in time.

There are two basic types of modelling approaches that are usually considered when repeated survey data (also known as time series/cross-sectional data) are available.

Figure 1. A window on the typical data structure collected in repeated surveys. Vertical axis indicates time, horizontal axis indicates different population units.



Type A approach — Different models are needed to explain systematic changes in the population distribution of the survey variables over time.

Type B approach — The same model for the population distribution of the survey variables holds at each time point.

Both approaches emphasise that the aim is to model the underlying survey population, and not the survey sample. Given the complex sampling procedures underlying these surveys, and the presence of nonresponse, there is no guarantee that an analysis which is based on the sample data alone, and does not take into account available covariate information about relationships between the responding and nonresponding sample units, and between the sample and the population, will be adequate. A key requirement in any analysis of data obtained from these surveys therefore is that the inference be adjusted for possible biases induced by sample design, selection and non-response effects at each time point for which data are available.

The purpose of this paper is to indicate a strategy for carrying out this adjustment with repeated survey data. Following BRECKLING et al. [1] and CHAMBERS [2], the adjustment involves linking known population covariates (auxiliary information) to the observed sample data in order to explain systematic differences between the observed (respondent) sample

and the population. Although the actual problem considered will be that of fitting a linear regression model to these data on the basis of ignorable sampling and nonresponse, it should be noted that the strategy generalises to other types of models that are often considered with repeated survey data. For example, the method easily extends to the case where the variables of interest are categorical (for example, labour force status) and the interest is in fitting an appropriate nonlinear regression model, like a logistic regression model, to the repeated survey data. The approach also generalises to the case where either the sampling procedure or the nonresponse (or both) is non-ignorable. However, in such cases, the necessity to include models for these non-ignorable effects makes the development fairly complex.

2. LINEAR MODELLING OF REPEATED SURVEY DATA

To minimise notational complexity, the following assumptions are made:

- The underlying population is fixed, with units indexed by $I = 1, 2, \ldots, N$, and the survey is carried out at time points $t = 1, 2, \ldots, T$.
- The population level covariate information that is available consists of the values of a scalar variable Z. These values determine the sample design of the survey, and they remain unchanged throughout the time period of interest (that is, there is no change in sample design over this period).
- Probability (that is, ignorable given Z) methods are used to select the sample at each time point, and the sample nonresponse at each time period is also ignorable.
- The aim is to model the linear regression over the population of a (univariate) survey variable Y on another (univariate) survey variable X. Typically, the population distributions of both Y and X are related to that of the sample design covariate Z. For example, Y could be the quantity of a particular manufactured good produced by a manufacturing business, X could be the current (depreciated) value of the manufacturing plant and machinery of the business, and Z could be the work force of the business.

Under a type A approach, it is assumed that the regression of Y on X varies with time in some systematic way:

$$E(Y_{It}|X_{It}) = \alpha_t + \beta_t X_{It}$$
$$var(Y_{It}|X_{It}) = \sigma_t^2. \qquad (1)$$

The aim is then to 'explain' changes in the regression parameters over time, by reference to corresponding changes in a 'global' variable G, say by a model of the form:

$$\Delta\beta_t = f(\Delta G_t; \xi).$$

Here Δ is the first order differencing operator, f is a given functional form and ξ is a vector of parameters to be estimated. Typically, ξ is estimated by minimising the squared deviations between the estimated values of $\Delta\beta_t$ and the corresponding values of f.

Under a type B approach, however, the same regression model is assumed to hold for all NT population data values:

$$E(Y|X) = \alpha + \beta X$$

$$var(Y|X) = \sigma^2. \tag{2}$$

In this case, the aim is to use the survey data collected over the entire time period of interest to estimate the parameters of (2).

In order to fit either (1) or (2), we develop two basic identities for the regression coefficients in these models. Put

$$\mu_Y = E(Y) \; ; \; \mu_X = E(X)$$

and

$$\sigma_{YY} = var(Y) \; ; \; \sigma_{XX} = var(X) \; ; \; \sigma_{YX} = cov(Y,X).$$

Then

$$\beta = \sigma_{XX}^{-1}\sigma_{YX} \; ; \; \alpha = \mu_Y - \beta\mu_X.$$

Let a subscript of Z denote expectation conditional on the known population Z-values. It follows

$$\beta = \frac{E(cov_Z(Y,X)) + cov(E_Z(Y), E_Z(X))}{E(var_Z(X)) + var(E_Z(X))} \tag{3a}$$

and

$$\alpha = E(E_Z(Y)) - \beta E(E_Z(X)). \tag{3b}$$

The identities (3a) and (3b) hold generally, but their interpretation differs somewhat between type A and type B approaches. Under a type A approach the expectations in these identities are with respect to the distribution of Y and X across spatial units (I). That is, we effectively add a time index to all quantities in (3a) and (3b). Under a type B approach, however, these expectations are across both spatial units (I) and time units (t), so (3a) and (3b) hold as they stand.

It can readily be seen that the real problem in analysis of longitudinal survey data is therefore one of fitting an appropriate model for behaviour conditional on Z. Estimates of the parameters of the linear model linking Y and X follow on substitution of the estimated parameters of this conditional model in (3a) and (3b).

The simplest specification for this conditional behaviour (perhaps after appropriate transformation of the survey data) is where different population units are uncorrelated, given Z, and where the conditional expectations of Y and X given Z vary linearly with Z, but their corresponding second order conditional moments are constant. If one also allows this behaviour to vary across time, this implies a model of the form

$$E_Z(Y_{It}) = a_Y(t) + b_Y(t)Z_I \tag{4a}$$

$$E_Z(X_{It}) = a_X(t) + b_X(t)Z_I \tag{4b}$$

and

$$cov_Z(Y_{It}, Y_{Is}) = \chi_{YY}(t,s)\ ;\ cov_Z(X_{It}, X_{Is}) = \chi_{XX}(t,s) \tag{4c}$$

$$cov_Z(Y_{It}, X_{Is}) = \chi_{YX}(t,s)\ ;\ cov_Z(X_{It}, Y_{Is}) = \chi_{XY}(t,s). \tag{4d}$$

Fitting the type A model (1) under this setup is direct. Design-adjusted estimates of β_t and α_t are computed by substituting appropriate estimates of the parameters of (4) into (3) at each time point and then averaging over the population Z-values. Thus (3a) implies an estimate for β_t in (2) of the form

$$\hat{\beta}_t = \frac{\hat{\chi}_{YX}(t,t) + \hat{b}_Y(t)\hat{b}_X(t)var(Z)}{\hat{\chi}_{XX}(t,t) + \hat{b}_X^2(t)var(Z)}. \tag{5}$$

Under joint normality of Y, X and Z, (5) is the 'Pearson adjusted' maximum likelihood estimate of β (SKINNER, HOLT and SMITH [3]). In general, however, if N is large (as it usually is) and if sample-based estimates of all time specific Z-conditional quantities are also maximum likelihood, then the simple moment estimators for β_t and α_t generated under the above approach will closely approximate the corresponding maximum likelihood estimates for these parameters under a much broader class of distributions for Y, X and Z.

The parameters of the type B model (2) do not vary with t, while the opposite is true of the parameters of (4). In order to use (4) to fit (2), therefore, one needs to use (4) to define a corresponding set of time invariant conditional moments for Y and X. Since the target population does not change over time, these time invariant conditional moments are obtained by

averaging over corresponding t-specific moments (defined by setting $s = t$ for second order moments) in (4). For example

$$E_Z(Y) = a_Y + b_Y Z = T^{-1}\sum_{t=1}^{T} a_Y(t) + b_Y(t)Z \tag{6a}$$

while

$$var_Z(X) = \chi_{XX}(Z) = T^{-1}\sum_{t=1}^{T}\chi_{XX}(t,t) + T^{-1}\sum_{t=1}^{T}\{(a_X(t)-a_X)+(b_X(t)-b_X)Z\}^2 \tag{6b}$$

and

$$cov_Z(Y,X) = \chi_{YX}(Z) = T^{-1}\sum_{t=1}^{T}\chi_{YX}(t,t) + T^{-1}\sum_{t=1}^{T}\left\{\begin{array}{l}\left[(a_Y(t)-a_Y)+(b_Y(t)-b_Y)Z\right] \\ \times\left[(a_X(t)-a_X)+(b_X(t)-b_X)Z\right]\end{array}\right\}. \tag{6c}$$

Given sample-based estimates of the time specific quantities in (4), their time invariant counterparts are easily estimated by substitution in (6), and design-adjusted estimates of β and α in (2) then obtained by substitution in (3) and averaging over the population Z-values. Thus

$$\hat{\beta} = \frac{\overline{\hat{\chi}_{YX}(Z)} + \hat{b}_Y\hat{b}_X\, var(Z)}{\overline{\hat{\chi}_{XX}(Z)} + \hat{b}_X^2\, var(Z)} \tag{7}$$

where a 'hat' denotes an estimate, and a 'bar' denotes averaging over the population Z-values. That is

$$\overline{\hat{\chi}_{YX}(Z)} = N^{-1}\sum_{I=1}^{N}\hat{\chi}_{YX}(Z_I)$$

while

$$var(Z) = N^{-1}\sum_{I=1}^{N}(Z_I - \bar{Z})^2.$$

3. GLS ESTIMATION OF THE PARAMETERS OF (4)

The development in the previous section indicates that the real problem in fitting either (1) or (2) is in using the longitudinal survey data to compute efficient estimates of the parameters of the conditional model (4). In this section a method of estimating these parameters based on a generalised least squares (GLS) approach is briefly outlined. Without loss of generality, it will be assumed that units that have ever appeared in sample over the period of interest are indexed by $I = 1, 2, \ldots, n$. The set of time points when unit I actually appears in sample will be denoted T_I. Thus $T_I = \{t_1, t_2, \ldots, t_{k(I)}\}$, where $k(I)$ denotes the number of times unit I appears in sample over the period of interest. The sample data D_s can then be expressed in vector form as

$$D_{It} = \begin{pmatrix} Y_{It} \\ X_{It} \end{pmatrix} ; \; D_I = \begin{pmatrix} D_{It_1} \\ D_{It_2} \\ \vdots \\ D_{It_{k(I)}} \end{pmatrix} ; \; D_S = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix}.$$

It follows from (4) that

$$E_Z(D_s) = \left\{ \underset{1 \le I \le n}{rbind} \left((1 \; Z_I) \otimes \mathbf{I}_{2T} [blk\_select(2T, 2, T_I),] \right) \right\} \begin{pmatrix} a \\ b \end{pmatrix}$$

$$= U_s(Z) \begin{pmatrix} a \\ b \end{pmatrix} \tag{8}$$

where $\mathbf{I}_{2T}$ denotes the identity matrix of order $2T$, *rbind* denotes the function that 'stacks' row vectors (indexed by I) to form a matrix, *blk_select(M, n, p)* is another function that divides the integers between 1 and M into M/n labelled blocks each of size n and returns the integers corresponding to blocks with labels in the set p. For example, *blk_select*(6, 2, {1, 3}) = {1, 2, 5, 6}, and

$$a' = (a_Y(1) \; a_X(1) \; a_Y(2) \; a_X(2) \cdots a_Y(T) \; a_X(T))$$

$$b' = (b_Y(1) \; b_X(1) \; b_Y(2) \; b_X(2) \cdots b_Y(T) \; b_X(T)).$$

Similarly, put

$$cov_Z(D_{It}, D_{Is}) = \chi(t,s) = \begin{bmatrix} \chi_{YY}(t,s) & \chi_{YX}(t,s) \\ \chi_{XY}(t,s) & \chi_{XX}(t,s) \end{bmatrix}$$

and

$$\chi = \begin{bmatrix} \chi(1,1) & \chi(1,2) & \cdots & \chi(1,T) \\ \chi(2,1) & \chi(2,2) & \cdots & \chi(2,T) \\ \vdots & \vdots & \ddots & \vdots \\ \chi(T,1) & \chi(T,2) & \cdots & \chi(T,T) \end{bmatrix}.$$

Then

$$cov_Z(D_I) = \chi_I = \chi[blk\_select(2T,2,T_I), blk\_select(2T,2,T_I)]$$

and, since different population units are uncorrelated under (4),

$$cov_Z(D_s) = \chi_s = blk\_diag\{\chi_I, I = 1,\ldots,n\} \tag{9}$$

where *blk_diag* denotes a function whose value is a square matrix with block diagonal structure defined by the matrices provided as arguments to this function.

The GLS estimator of the parameters a and b in (8) is therefore

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left[U_s'(Z)\hat{\chi}_s^{-1}U_s(Z)\right]^{-1}\left[U_s'(Z)\hat{\chi}_s^{-1}D_s\right]. \tag{10}$$

Note that (10) depends on knowledge of χ, and, in turn, this requires that χ be modelled. Given such a model, the estimation procedure is iterative, cycling between the GLS estimates of a and b defined by (10) and the estimate of χ.

Choice of an appropriate model for χ will depend on the particular population being surveyed and the nature of the survey design. Under the model (4), different population units are uncorrelated given their values of Z. In this case a simple autoregressive structure for χ should suffice to capture the time dependence in the survey data. For example

$$\chi(t,s) = \begin{bmatrix} \omega_{YY}\rho_{YY}^{|t-s|} & \omega_{YX}\rho_{YX}^{|t-s|} \\ \omega_{YX}\rho_{XY}^{|t-s|} & \omega_{XX}\rho_{XX}^{|t-s|} \end{bmatrix}. \tag{11}$$

Standard moment estimates for the parameters of (11) are straightforward to specify.

4. SUMMARY

The main contribution of this paper has been to describe a strategy for the analysis of repeated survey data which incorporates adjustments compensating for the effect of the complex sample design that is characteristic of such surveys. For the purpose of exposition, the strategy has been developed in the context of a standard linear regression analysis of

these data, and has assumed ignorable sampling and nonresponse. It is technically feasible, though notationally complex, to relax these assumptions.

A key feature of the approach is that incorporation of sample design effects inevitably requires modelling of the relationship between the survey variables of interest and the sample design information. Since the paper is expository, it has assumed a simple linear regression structure for this relationship. The basic approach, however, can easily be extended to cover more complex relationships, including those where the underlying population and sample structures are clustered, so that variance component models are appropriate for describing the relationship between the survey variables and the sample design information.

REFERENCES

[1] BRECKLING, J. U., CHAMBERS, R. L., DORFMAN, A. H., TAM, S. M. and WELSH, A. H. (1990). Maximum likelihood inference from sample survey data. *Canberra Statistics Technical Reports - 015 - 90*, Statistics Research Section, SMS, Australian National University.

[2] CHAMBERS, R. L. (1986). Design adjusted parameter estimation. *Journal of the Royal Statistical Society*, A **149**, 161-173.

[3] SKINNER, C. J., HOLT, D. and SMITH, T. M. F. (1989) (Editors). *Analysis of complex surveys*. Chichester: Wiley.

Australian Bureau of Agricultural and Resource Economics
GPO Box 1563
Canberra ACT 2601
Australia