Variance Estimation for Measures of Trends with Rotated Repeated Surveys

Yves G. BERGER
University of Southampton, S3RI, SO17 1BJ, Southampton, United Kingdom

Abstract

Measuring trend or change over time is a central problem for many users of social, economic and demographic data and is of interest in many areas of economics and social sciences. Smith *et al.* (2003) recognised that assessing change is one of the most important challenges in survey statistics. The primary interest of many users is often in trends rather than cross sectional estimates. Samples at different waves are not necessarily completely overlapping sets of units, because repeated surveys often use rotating samples which consist in selecting for each wave new units to replace old units that have been in the sample for a specified number of waves (Tam, 1984; Nordberg, 2000; Kalton, 2009). Moreover, surveys are usually stratified and units can be selected with unequal probabilities. In this paper, we propose a novel approach to estimate trends and its variance taking into account of rotations, stratification and unequal probabilities. The variance depends on covariances between estimates calculated from different waves. In a series of simulation based on the Swedish Labour Force Survey, Andersson *et al.* (2011) showed that the approach proposed by Berger & Priam (2010) can give more accurate estimates of covariance than standard estimators of covariance (Tam, 1984; Qualité & Tillé, 2008). In this paper, we show how the approach proposed by Berger & Priam (2010) can be used to estimate the variance of a trend parameter. The proposed method is a semi-parametric design-based approach which is based upon a multivariate linear regression (or general linear) model (Berger & Priam, 2011). This multivariate regression model captures the effect of rotations, unequal probabilities, stratification and unequal probabilities.

Key Words: Inclusion probabilities, Regression model, Rotation sampling designs.

1. Introduction

In this paper, we propose a novel approach to estimate a trend and its variance taking into account of the rotation, the stratification and the unequal probabilities. In Section 2, we define the class of rotation designs considered. In Section 3, we defined the trend parameter. In Section 3.1, we proposed a model-design unbiased estimator of the trend parameter. In Section 3.2, we show how Berger & Priam (2010) approach can be use to estimate the variance of a trend parameter.

2. Rotation Designs

Rotation designs consist in selecting, for each wave, new units to replace old units that have been in the sample for a specified number of waves (Tam, 1984; Nordberg, 2000, Kalton, 2009). In Sections 2.1, 2.2 and 2.3, we show how rotation designs can be used to select samples.

2.1 Unequal Probability Rotation Designs for Two Waves

Let s_1 and s_2 denote respectively the first and second wave samples. We consider that s_1 and s_2 have the same fixed sample size n. Assume that s_1 is a probability sample without replacement with first-order inclusion probabilities π_i. Suppose that s_2 is a simple random sample without replacement sample of $n_{1,2}$ units selected without replacement from s_1 combined with a sample of $n_{1|2} = n - n_{1,2}$ units selected without replacement from U / s_1 with probabilities $q_i = \pi_i (1 - n_{1,2} / n) / (1 - \pi_i)$; where U denotes the population and U / s_1 is the set of units not selected at wave 1. Tam (1984) studied this design when $\pi_i = n_1 / N$; where N denotes the population size. Note that the first-order inclusion probabilities of s_2 are also given by π_i (Berger & Priam, 2010).

2.2. Random Groups Rotation Designs for Two Waves

There are other rotation designs used in practice such that the rotation groups sampling design. Suppose that we have a single stratum randomly divided into G rotation groups. We assume N / G integer. At $t = g$, the first g groups are selected. At $t = g + 1$, group 1 rotates out and group $g + 1$ rotates in. By assuming that the G rotation groups are randomly constructed, the rotation group sampling design and the design described in Section 2.1 are equivalent when $\pi_i = n / N$. Rotation groups can be also constructed with systematic sampling.

2.3 Rotation Designs for More than Two Waves

Consider that we have T waves. Let { s_1, s_2, ..., s_t, ..., s_T } denote a series of samples selected at each waves, where the sample s_t and s_{t+1} are selected using a designs described in Section 2.1 or 2.2. Thus, any pair of sample s_ℓ and s_t can be overlapping sets of units. Let us consider that all the samples s_t $(t = 1, \cdots, T)$ have the same sample fixed size n. We denote by $n_{\ell,t}$ the number of units in $s_\ell \cap s_t$. Note that in practice $n_{t,(t+1)}$ is a fixed quantity which does not depend on t. For example, if 20% of units rotate in and out from the sample at each wave, we have that $n_{t,(t+1)} = 0.8 \times n$.

Note that under a rotation groups sampling design in (see Section 2.2) with G groups, the quantities $n_{\ell,t}$ are fixed and given by $n_{\ell,t} =$ $n \max\{0, 1 - | t - \ell | / G\}$. Note that $n_{\ell,t} = 0$ if $| t - \ell | > G$, implying that $s_\ell \cap s_t = \varnothing$. This means that under rotation groups sampling, two samples s_ℓ and s_t will have no units in common when $| t - \ell | > G$. Under the rotation design described in Section 2.1, the quantities $n_{\ell,t}$ are random, and have a positive probability of being equal to zero when $| t - \ell | > 1$.

3. Estimating a Trend

Suppose that we would like to estimate the trend of a variable of interest y over time. This trend can be measured by the trend parameter $\boldsymbol{\beta}$ of the following model

$$y_{it} = \boldsymbol{\beta}' \mathbf{x}_{it} + e_{it} ; \tag{1}$$

where y_{it} and $\mathbf{x}_{it}$ denote respectively the values of unit i at wave t of the variable of interest y and p covariates. The residuals e_{it} are assumed to be correlated within subjects. We consider that $\mathbf{x}_{it}$ contains the wave number t and possible interactions of t with other covariates or even t^2 if the trend is not linear. For example, when $\mathbf{x}_{it} = (1, t)'$ we have that $\boldsymbol{\beta} = (\beta_0, \beta_t)'$ and $y_{it} = \beta_0 + \beta_t\, t + e_{it}$.

In Section 3.1, we propose an estimator for $\boldsymbol{\beta}$ and an estimator of its variance in Section 3.2. We show that both estimators are approximately design unbiased and take into account of the rotation design, the inclusion probabilities and stratification.

The classical model-based approach consists in estimating $\boldsymbol{\beta}$ using a fixed or random effect model from the sample data given by the values of y_{it} and $\mathbf{x}_{it}$ for the set of units selected at T consecutive waves. Hence, the sample data are given by $\{y_{it}, \mathbf{x}_{it} : i \in s\}$, where $s = \cup_{t=1}^{T} s_t$. Note that under a rotation groups sampling design, the sample data is empty ($s = \varnothing$) when $T - 1 > G$ (see Section 2.3). Nevertheless, the proposed estimator (2) and its variance estimator (5) can be still calculated even if $s = \varnothing$.

The model-based likelihood approach (Diggle et al. 1994) needs assumption about the correlations between the e_{it} in orders to obtain consistent estimates for the variance-covariance of the estimator of $\boldsymbol{\beta}$. The proposed point estimator of $\boldsymbol{\beta}$ in Section 3.1 does not depend on these correlations. However the proposed variance estimator (see Section 3.2) depends on between waves correlations generated by the rotation design. We propose to estimate these correlations using another multivariate regression model (6). We show that this approach gives design-based consistent estimator for the correlations even if model (6) does not fit the data.

3.1 The Proposed Estimator of $\boldsymbol{\beta}$

Let $\hat{\mathbf{B}}_U$ be the usual Ordinary Least Squares (OLS) estimator of $\boldsymbol{\beta}$ based on the population values $\{y_{it}, \mathbf{x}_{it} : i \in U\}$. It is well known that $\hat{\mathbf{B}}_U$ is model unbiased (Diggle *et al.* 1994 page 58); that is, $E_m(\hat{\mathbf{B}}_U) = \boldsymbol{\beta}$, where $E_m(\cdot)$ denotes the expectation with respect to the model (1). We proposed to predict $\hat{\mathbf{B}}_U$ by the following design-based weighted estimator

$$\hat{\boldsymbol{\beta}}_s = \left(\sum_{t=1}^{T} \sum_{i \in s_t} \frac{1}{\pi_i} \mathbf{x}_{it} \mathbf{x}_{it}^{'} \right)^{-1} \left(\sum_{t=1}^{T} \sum_{i \in s_t} \mathbf{x}_{it} \frac{y_{it}}{\pi_i} \right). \tag{2}$$

If the population model holds $\hat{\boldsymbol{\beta}}_s$ is also approximately model-design-unbiased, as

$$E_m(E_d(\hat{\boldsymbol{\beta}}_s)) \approx E_m(\hat{\mathbf{B}}_U) = \boldsymbol{\beta} ,$$

where $E_d(\cdot)$ denote the expectation with respect to the rotation design used.

3.2 An Estimator for the Variance-Covariance of $\hat{\boldsymbol{\beta}}_s$

Note that $\hat{\boldsymbol{\beta}}_s$ is a smooth function of Q totals where $Q = p(p+1)/2$; that is,

$$\hat{\boldsymbol{\beta}}_s = f(\hat{\tau}_1, \cdots, \hat{\tau}_Q) = f(\hat{\boldsymbol{\tau}});$$

where

$$\hat{\tau}_q = \sum_{t=1}^{T} \hat{\tau}_t^{(q)}, \qquad (q = 1, \cdots, Q) \tag{3}$$

with

$$\hat{\tau}_t^{(q)} = \sum_{i \in s_t} \frac{w_{it}^{(q)}}{\pi_i} \qquad (q = 1, \cdots, Q)$$

and $w_{it}^{(q)}$ is defined by the totals involved in (2). For example, under a random groups rotation design, we have that $Q = 3$, $w_{it}^{(1)} = 1$, $w_{it}^{(2)} = y_{it}$ and $w_{it}^{(3)} = ty_{it}$, when $\mathbf{x}_{it} = (1, t)'$.

As $\hat{\boldsymbol{\beta}}_s$ is a smooth function of totals, we can use the delta method to derive the following design-based linearised estimator for the variance-covariance

$$\hat{\text{var}}_d(\hat{\boldsymbol{\beta}}_s) = \nabla(\hat{\boldsymbol{\tau}})' \, \hat{\text{var}}_d(\hat{\boldsymbol{\tau}}) \, \nabla(\hat{\boldsymbol{\tau}}), \tag{4}$$

where $\nabla(\hat{\boldsymbol{\tau}}) = \partial \hat{\boldsymbol{\beta}}_s / \partial \hat{\boldsymbol{\tau}}$ is the $Q \times 1$ gradient vector of $f(\hat{\boldsymbol{\tau}})$ at $\hat{\boldsymbol{\tau}}$. Using (3), we have that

$$\hat{\boldsymbol{\tau}} = \sum_{t=1}^{T} \hat{\boldsymbol{\tau}}_t,$$

where $\hat{\boldsymbol{\tau}}_t = (\hat{\tau}_t^{(1)}, \cdots, \hat{\tau}_t^{(Q)})'$. Thus, (4) can be re-written as

$$\hat{\text{var}}_d(\hat{\boldsymbol{\beta}}_s) = \nabla(\hat{\boldsymbol{\tau}})' \sum_{\ell=1}^{T} \sum_{t=1}^{T} \hat{\boldsymbol{\Sigma}}_{\ell t} \, \nabla(\hat{\boldsymbol{\tau}}); \tag{5}$$

where $\hat{\boldsymbol{\Sigma}}_{\ell t} = \hat{\text{cov}}_d(\hat{\boldsymbol{\tau}}_\ell, \hat{\boldsymbol{\tau}}_t)$ is the $Q \times Q$ matrix block (ℓ, t) of the covariance matrix $\hat{\boldsymbol{\Sigma}}$ between the estimators $\{\hat{\tau}_1^{(1)}, \cdots, \hat{\tau}_1^{(Q)}, \cdots, \hat{\tau}_t^{(1)}, \cdots, \hat{\tau}_t^{(Q)}, \cdots, \hat{\tau}_T^{(1)}, \cdots, \hat{\tau}_T^{(Q)}\}$. We propose to use a multivariate regression approach (Berger & Priam, 2010) to calculate the covariance matrix $\hat{\boldsymbol{\Sigma}}$.

In a series of simulation based on the Swedish Labour Force Survey, Andersson *et al.* (2011) showed that the method proposed by Berger & Priam (2010) gives more accurate estimates of covariance than standard estimators of covariance (Tam, 1984; Qualité & Tillé, 2008) for estimate of strata domains. This is not a surprise since the correlations are implicitly calculated within each stratum (Andersson *et al.* 2011). This property is important when the trend parameter is an interaction parameter with strata domains. Furthermore, the approach proposed by Berger & Priam (2010) can accommodate temporal stratification and unequal probabilities. Temporal stratification means that the stratification at t differs from the stratification at $t+1$; i.e., new strata are created and units move between strata. However, the approach proposed by Berger & Priam (2010) relies on the assumption that the sampling fractions are negligible. Berger (2004) proposed a more general method based on the same principle, which account for large sampling fractions. For large sampling fractions, it is recommended to use the more general estimator proposed by Berger (2004). In the rest of this section, we show how the approach proposed by Berger & Priam (2010) can be used to calculate the covariance matrix $\hat{\mathbf{\Sigma}}$, when we have one stratum.

Let $\tilde{w}_{it}^{(q)}$ be defined by $\tilde{w}_{it}^{(q)} = 0$ if $i \notin s_t$, and $\tilde{w}_{it}^{(q)} = w_{it}^{(q)} / \pi_i$ if $i \in s_t$; where $i \in s = \cup_{t=1}^{T} s_t$. Let $\tilde{n} = \# s$. Consider the following $\tilde{n} \times QT$ matrix of the $\tilde{w}_{it}^{(q)}$: $\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_t, \cdots, \tilde{\mathbf{w}}_T)$; where $\tilde{\mathbf{w}}_t = (\tilde{\mathbf{w}}_t^{(1)}, \cdots, \tilde{\mathbf{w}}_t^{(Q)})$ and $\tilde{\mathbf{w}}_t^{(q)} = (\tilde{w}_{1t}^{(q)}, \cdots, \tilde{w}_{\tilde{n}t}^{(q)})'$.

$$\tilde{\mathbf{w}} = \mathbf{Z}_s \boldsymbol{\alpha} + \boldsymbol{\varepsilon}\,; \qquad (6)$$

where $\boldsymbol{\alpha}$ is a $L \times QT$ matrix of regression parameters and $\mathbf{Z}_s$ is a $\tilde{n} \times L$ design matrix which specifies the fixed sizes constraints of the rotation design. The residuals $\boldsymbol{\varepsilon}$ have a $QT \times QT$ covariance matrix $\mathbf{S}$.

For the rotation design described in Section 2.1, $L = 2T - 1$, as we have T design variables $z_{t;i}$, and $T - 1$ interactions $z_{t-1;i} z_{t;i}$; where $z_{t;i} = 1$ if $i \in s_t$, and $z_{t;i} = 0$ if $i \notin s_t$. It can be shown that $\Sigma_{i \in s} z_{t;i} = n$ and $\Sigma_{i \in s} z_{t-1;i} z_{t;i} = n_{(t-1),t}$ are fixed. For the rotation scheme describe in Section 2.2, we have additional interactions $z_{\ell;i} z_{t;i}$, because the sums $\Sigma_{i \in s} z_{\ell;i} z_{t;i} = n_{\ell,t}$ are fixed. The fact that these sums are fixed justifies the covariate $\mathbf{Z}_s$ used in (6) (Berger & Priam, 2010). More design variables are needed for stratified designs (Berger & Priam, 2010). The matrix $\tilde{\mathbf{y}}$ can be modified to accommodate two-stage designs (Berger & Priam, 2010).

Berger & Priam (2010) showed that

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{D}}' \hat{\mathbf{S}} \hat{\mathbf{D}} \qquad (7)$$

is an approximately design unbiased estimator for the covariance matrix between the $\hat{\tau}_t^{(q)}$ when the finite population corrections are negligible. The matrix $\hat{\mathbf{S}}$ is the OLS residual covariance matrix estimate of the model (6) and $\hat{\mathbf{D}}$ is a diagonal matrix with diagonal elements $\{\hat{\mathrm{var}}(\hat{\tau}_t^{(q)}) \hat{S}_{jj}^{-1}\}^{1/2}$ where $\hat{\mathrm{var}}(\hat{\tau}_t^{(q)})$ is a standard design-based variance estimator of $\hat{\tau}_t^{(q)}$ where $(t-1)Q + q = j$ and $\hat{S}_{qq}$ is the q-th diagonal component of $\hat{\mathbf{S}}$. The estimator $\hat{\mathbf{\Sigma}}$ is a design-based consistent estimator for the covariance matrix between the $\hat{\tau}_t^{(q)}$ even when model (6) does not fit the data. The estimator for the variance of $\hat{\boldsymbol{\beta}}_s$ is obtained by substituting (7) into (5).

Note that the overall variance (model & design) of $\hat{\boldsymbol{\beta}}_s$ is given by

$$\begin{aligned} \mathrm{var}(\hat{\boldsymbol{\beta}}_s) &= E_m(\mathrm{var}_d(\hat{\boldsymbol{\beta}}_s)) + \mathrm{var}_m(E_d(\hat{\boldsymbol{\beta}}_s)) \\ &\approx E_m(\mathrm{var}_d(\hat{\boldsymbol{\beta}}_s)) + \mathrm{var}_m(\hat{\mathbf{B}}_U) \\ &\approx E_m(\mathrm{var}_d(\hat{\boldsymbol{\beta}}_s)) \end{aligned}$$

Thus, the proposed variance estimator is also approximately unbiased for the overall variance of $\hat{\boldsymbol{\beta}}_s$, as

$$E(\hat{\mathrm{var}}_d(\hat{\boldsymbol{\beta}}_s)) = E_m(E_d(\hat{\mathrm{var}}_d(\hat{\boldsymbol{\beta}}_s))) \approx E_m(\mathrm{var}_d(\hat{\boldsymbol{\beta}}_s)) \approx \mathrm{var}(\hat{\boldsymbol{\beta}}_s).$$

4. Conclusion

In this paper, we propose a novel approach to estimate a trend parameter and its variance taking into account of rotations, stratification and unequal probabilities. The point estimator of the trend parameter is a standard design-based estimator of the population trend estimator. We show that the proposed estimator is also model-design unbiased. For variance estimation, we used the delta method as the point estimator is a function of totals. However, it is necessary to estimate covariances between cross-sectional totals measured at different waves. We propose a semi-parametric design-based approach (Berger & Priam, 2011) to estimate these covariances. This model captures the effect of rotations, unequal probabilities and stratification.

Acknowledgements

This work was supported by grant RES-000-22-3045 of the Economic and Social Research Council (UK) and the British Academy.

References

Andersson C., Andersson K. and Lundquist P. (2011) *Variansskattningar avseende förändringsskattningar i panelundersökninga*r (Variance Estimation of change in panel surveys.). Methodology reports from Statistics Sweden.

Berger Y.G. (2004). Variance estimation for measures of change in probability sampling. Canadian Journal of Statistics, 32, 4, 451-467.

Berger Y.G. and Priam R. (2010), Estimation of correlations between cross-sectional Estimates from Repeated Surveys - an Application to the Variance of Change. Proceeding of the 2010 Symposium of Statistics Canada.

Diggle P. J., Liang K-Y. and Zeger S. L. (1994). The Analysis of Longitudinal Data. Oxford: Clarendon Press.

Kalton G. (2009), Design for surveys over time. Handbook of Statistics: Design, Method and Applications: D. Pfeffermann and C.R. Rao. (editors). Elsevier.

Nordberg L. (2000). On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. Journal of Official Statistics, 16, 363-378.

Qualité L. and Tillé, Y. (2008). Variance estimation of change in repeated surveys and its application to the Swiss survey of value added. Survey Methodology, 34, 2, 173-181.

Smith P., Pont M. and Jones T. (2003). Developments in business survey methodology in the Office for National Statistics, Journal of the Royal Statistical Society, Series D, 52, 1-30.

Tam S.M. (1984), On covariances from overlapping samples. The American Statistician, 38, 288-289.