A multivariate regression approach for estimating variances of measures of population change over time from rotating repeated surveys

Yves G. Berger and Rodolphe Priam

University of Southampton, UK

Abstract

Measuring 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 changes or trends from one time period to another. We propose a multivariate linear regression approach to estimate the variance of change.

A common problem is to compare two cross-sectional estimates for the same study variable taken on two different waves or occasions, and to judge whether the observed change is statistically significant. Assessing the significance of a change plays an important part in preliminary exploratory data analysis which helps to model a trend or a change.

Suppose, we wish to estimate the absolute change $\Delta = \tau_2 - \tau_1$ between two population totals $\tau_1 = \sum_{i \in U} y_{i1}$ and $\tau_2 = \sum_{i \in U} y_{i2}$ at wave 1 and wave 2 respectively. The quantities y_{i1} and y_{i2} are respectively the value of the variable of interest measure at wave 1 and 2. Suppose that Δ is estimated by $\hat{\Delta} = \hat{\tau}_2 - \hat{\tau}_1$, where $\hat{\tau}_1$ and $\hat{\tau}_2$ are the Horvitz-Thompson estimates. A primary interest is to test if an observed change is due to an actual change in the population or simply due to sampling errors. The variance of the change $\hat{\Delta}$ is given by

$$var(\hat{\Delta}) = var(\hat{\tau}_1) + var(\hat{\tau}_2) - 2 \times cor(\hat{\tau}_1, \hat{\tau}_2)\sqrt{var(\hat{\tau}_1)var(\hat{\tau}_2)}$$

Standard estimators can be used to estimate the variances $var(\hat{\tau}_1)$ and $var(\hat{\tau}_2)$. The correlation $cor(\hat{\tau}_1, \hat{\tau}_2)$ is the most difficult part to estimate.

The estimation of the correlation would be relatively straightforward if $\hat{\tau}_1$ and $\hat{\tau}_2$ were based upon the same sample, or if the sample remained the same from one wave to the next. Unfortunately, samples at different waves are usually not completely overlapping sets of units, because repeated surveys use rotating samples which consist in

selecting new units at wave 2 to replace old units that have been in the sample for a specified number of waves. Therefore, y_{i1} is known and y_{i2} is unknown for the units being replaced, on the other hands, y_{i1} is unknown and y_{i2} is known for the new units. For most of the units sampled at both waves, y_{i1} and y_{i2} are known.

Several methods can be used to estimate the variance (e.g. Kish 1965; Tam 1984, Holmes & Skinner 2000, Berger 2004). We propose to use a multivariate linear regression approach to estimate the correlation. Consider the multivariate model

$$\begin{pmatrix} \breve{y}_{i1} \\ \breve{y}_{i2} \end{pmatrix} = \begin{pmatrix} \beta_1^1 z_{i1} + \beta_2^1 z_{i2} + \beta_{12}^1 z_{i1} z_{i2} \\ \beta_1^2 z_{i1} + \beta_2^2 z_{i2} + \beta_{12}^2 z_{i1} z_{i2} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

Where $\breve{y}_{i1} = y_{i1}/\pi_{i1}$ and $\breve{y}_{i2} = y_{i2}/\pi_{i2}$. The quantities π_{i1} and π_{i2} are respectively the inclusion probabilities for wave 1 and 2. The quantities $z_{ik} = 1$ if $i \in$ wave k sample and $z_{ij} = 0$ otherwise. The 2×1 vector residuals $(\epsilon_1, \epsilon_2)^{'}$ can be assumed random with a bivariate normal distribution $N(0, \Sigma)$ where Σ is the 2×2 residual covariance matrix. Note that this model includes interactions between the variable z_{i1} and z_{i2}. We will show that these interactions capture the rotation of the repeated survey.

Finally, the estimator proposed for the correlation will be based upon the residuals covariance matrix Σ. Let $\hat{\Sigma}$ be the estimator of Σ. The proposed estimator for the correlation between $\hat{\tau}_1$ and $\hat{\tau}_2$ is given by

$$\hat{cor}(\hat{\tau}_1, \hat{\tau}_2) = \frac{\hat{\Sigma}_{1,2}}{\sqrt{\hat{\Sigma}_{1,1}\hat{\Sigma}_{2,2}}}$$

where $\hat{\Sigma}_{k,\ell}$ is the component (k, ℓ) of the matrix $\hat{\Sigma}$.

We will show why the multivariate linear regression model is suitable for estimating the correlation, and how our result could be extended to include stratification.

Keywords

Change, Correlation, Horvitz-Thompson estimator, Multivariate regression, Repeated surveys, Rotation, Survey sampling.

References

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Canad. J. Statist. 32*, 451–467.

Holmes, D.J. and Skinner, C.J. (2000). *Variance Estimation for Labour Force Survey Estimates of Level and Change*. GSS Methodology Series 21.

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

Smith, P., Pont, M., and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000. *The Statistician 52(3)*, 257–295.

Tam, S.M. (1984). On covariances from overlapping samples. *Amer. Statist. 38*, 288–289.