Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

On aligned composite estimates from overlapping samples for growth rates and totals

by Paul Knottnerus

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 265-283
Statistics Canada, Catalogue No. 12-001-X

On aligned composite estimates from overlapping samples for growth rates and totals

Paul Knottnerus[1]

Abstract

When monthly business surveys are not completely overlapping, there are two different estimators for the monthly growth rate of the turnover: (i) one that is based on the monthly estimated population totals and (ii) one that is purely based on enterprises observed on both occasions in the overlap of the corresponding surveys. The resulting estimates and variances might be quite different. This paper proposes an optimal composite estimator for the growth rate as well as the population totals.

Key Words: Business surveys; Coefficient of variation; General restriction estimator; Kalman equations; Panels; Variances.

1 Introduction

In many countries a monthly business survey is held for the major Standard Industrial Classification (SIC) codes to estimate the level of the monthly turnover and the change in that level compared to a month or a year ago. When repeatedly sampling a population, a complicating factor is that there are various methods for estimating the (relative) change from a panel with different outcomes especially when the samples on different occasions are not completely overlapping.

Kish (1965), Tam (1984), Laniel (1987), Hidiroglou, Särndal and Binder (1995), Nordberg (2000), Berger (2004), Qualité and Tillé (2008), Wood (2008) and Knottnerus and Van Delden (2012) examined various estimators for the parameter of change in different situations. The main aim of this paper is to derive estimators for a relative change as well as the corresponding population totals that are in line with each other and that have minimum variance property. The derivation of the aligned composite estimators is based on the general restriction (GR) estimator of Knottnerus (2003). Composite estimators for totals and (absolute) changes are also proposed by Särndal, Swensson and Wretman (1992, pages 370-378) but in separate steps. Moreover, this paper focuses on estimators for growth rates because: (i) users of figures from business surveys for a specific SIC code often are more interested in growth rates than in absolute changes, (ii) in practice there might be model-assisted reasons to look at growth rates (auxiliary variables in regression models often explain the different growth rates of the units rather than their different levels), and (iii) growth rates are needed for making an overall index for the (monthly) turnover for each of the major SIC codes. For instance, Smith, Pont and Jones (2003) describe the method of matched pairs to measure a change from month to month, using responses that are common to both periods. The authors use this method for deriving the monthly retail sales index (RSI).

The outline of the paper is as follows. Section 2 briefly describes two methods for estimating a growth rate of the total turnover for enterprises with a certain SIC code. Two examples illustrate the possibly substantial differences between the two approaches. Section 3 discusses the question of which estimation

1. Paul Knottnerus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. Email: pkts@cbs.nl.

method is to be preferred and explains as to why the difference between the variances of both estimators might be so large. For various situations Section 4 and Section 5 propose an optimal composite estimator. Section 6 discusses some extensions of the aligned composite (AC) estimator for growth rates and totals. Section 7 summarizes the main conclusions and issues to be further investigated.

2 Two estimators for the growth rate of the total turnover

Consider a population of N enterprises $U=\{1, ..., N\}$, and suppose there are no births and deaths in the population. Let Y_i denote the value of the turnover for the i-th enterprise in a given month (say t) and X_i the value of the turnover of that enterprise in month $t-12$. Hence, the variables y and x concern the same variable on two different occasions. Denote their population totals by Y and X, and their population means by $\bar{Y}$ and $\bar{X}$, respectively. That is, $Y=\sum_{i\in U} Y_i$, $X=\sum_{i\in U} X_i$, $\bar{Y}=Y/N$ and $\bar{X}=X/N$. Let s_1, s_2 and s_3 denote three mutually disjoint simple random samples from U without replacement (SRS). Define s_{12} and s_{23} by $s_{12}=s_1\cup s_2$ and $s_{23}=s_2\cup s_3$, respectively. Denote the size of s_k by n_k ($k=1, 2, 3, 12, 23$) and the corresponding sample means by $\bar{y}_k$ and $\bar{x}_k$. Let the variable x be observed in s_{12} on the first occasion and the variable y in s_{23} on the second occasion. Denote the overlap ratios by λ ($=n_2/n_{12}$) and μ ($=n_2/n_{23}$). The SRS estimators for the population totals Y and X are defined by $\hat{Y}_{SRS}=N\bar{y}_{23}$ and $\hat{X}_{SRS}=N\bar{x}_{12}$, respectively.

Define the growth rate g of the total turnover between the two occasions by $g=G-1$ with $G=Y/X$. For estimating G there are two options. One of the standard (STN) options is based on the estimated totals on both occasions, that is

$$\hat{G}_{STN}=\frac{\hat{Y}_{SRS}}{\hat{X}_{SRS}}=\frac{\bar{y}_{23}}{\bar{x}_{12}}; \tag{2.1}$$

see Nordberg (2000), Qualité and Tillé (2008) and Knottnerus and Van Delden (2012). Note that the estimator $\hat{g}_{STN}=\hat{G}_{STN}-1$ for g has the same variance as $\hat{G}_{STN}$. For sufficiently large n this variance can be approximated by using a first-order Taylor series expansion of $\hat{G}_{STN}$. That is,

$$\begin{aligned}\operatorname{var}\left(\hat{G}_{STN}\right) &\approx \frac{1}{\bar{X}^2}\operatorname{var}\left(\bar{y}_{23}-G\bar{x}_{12}\right)\\ &=\frac{1}{\bar{X}^2}\left\{\operatorname{var}\left(\bar{y}_{23}\right)+G^2\operatorname{var}\left(\bar{x}_{12}\right)-2G\operatorname{cov}\left(\bar{y}_{23},\bar{x}_{12}\right)\right\}\\ &=\frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n_{23}}-\frac{1}{N}\right)S_y^2+G^2\left(\frac{1}{n_{12}}-\frac{1}{N}\right)S_x^2-2G\left(\frac{\lambda\mu}{n_2}-\frac{1}{N}\right)S_{xy}\right\},\end{aligned} \tag{2.2}$$

where $S_y^2=\sum_U\left(Y_i-\bar{Y}\right)^2/\left(N-1\right)$ is the adjusted population variance of the Y_i and S_x^2 that of the X_i while $S_{xy}=\sum_U\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)/\left(N-1\right)$ is their adjusted population covariance. Cochran (1977, page 153) suggests as *working rule* to use the large-sample result if the sample size exceeds 30 and the coefficients of variation of the numerator and denominator are less than 10%. For (different) derivations

of the expression for $\text{cov}(\bar{y}_{23}, \bar{x}_{12})$ used in (2.2), see Tam (1984) and Knottnerus and Van Delden (2012). The adjusted population (co)variances can be estimated unbiasedly by the sample (co)variances; recall sample (co)variances s_{yk}^2 and s_{yxk} from sample s_k $(k = 1, 2, 3, 12, 23)$ are defined by

$$s_{yk}^2 = \frac{1}{n_k - 1} \sum_{i \in s_k} (Y_i - \bar{y}_k)^2$$

$$s_{yxk} = \frac{1}{n_k - 1} \sum_{i \in s_k} (Y_i - \bar{y}_k)(X_i - \bar{x}_k).$$

An alternative option for estimating G and g is based on enterprises observed on both occasions in overlap s_2 (OLP). That is,

$$\hat{G}_{OLP} = \frac{\bar{y}_2}{\bar{x}_2} \tag{2.3}$$

For sufficiently large n_2, the well-known approximation for the variance of this estimator is

$$\begin{aligned} \text{var}\left(\hat{G}_{OLP}\right) &\approx \frac{1}{\bar{X}^2} \text{var}\left(\bar{y}_2 - G\bar{x}_2\right) \\ &= \frac{1}{\bar{X}^2}\left(\frac{1}{n_2} - \frac{1}{N}\right) S_{y-Gx}^2, \end{aligned} \tag{2.4}$$

where S_{y-Gx}^2 stands for $S_y^2 + G^2 S_x^2 - 2GS_{xy}$; see Cochran (1977, page 31). In order to get some more insight into the merits of both $\hat{g}_{STN}$ and $\hat{g}_{OLP}$, consider the following examples.

Example 2.1. The data used in this example are panel observations on the turnover of Dutch supermarkets in February 2011 and 2012 from stratum 3 (size class 3). The stratum size is $N = 386$. Furthermore, $n_1 = 15$, $n_2 = 57$ and $n_3 = 17$. For the different samples we have (in thousand euros)

$$\bar{y}_{23} = 97.2,\ \bar{x}_{12} = 89.8,\ s_{y23}^2 = 3,781,\ \text{and}\ s_{x12}^2 = 2,232.$$

The population correlation coefficient $\rho_{xy}\left(= S_{xy}/S_x S_y\right)$ between the Y_i and the X_i is estimated from overlap s_2 by $\hat{\rho}_{xy2} = s_{xy2}/s_{y2}s_{x2} = 0.876$. To avoid negative variance estimates, Knottnerus and Van Delden (2012) propose estimating S_{xy} in (2.2) by $\hat{S}_{xy} = \hat{\rho}_{xy2} s_{x12} s_{y23} = 2,545$. Substituting the above outcomes into (2.1) and (2.2), we obtain $\hat{g}_{STN} = 0.082\ (= 8.2\%)$ and $\hat{\text{var}}(\hat{g}_{STN}) = 0.00324$. Assuming normality and using $u_{0.975} = 1.96$, the 95%-confidence interval is approximately $I_{STN}^{95} \approx (-3.0\%, 19.4\%)$. In contrast, from overlap s_2 we get the estimates

$$\bar{y}_2 = 102.2,\ \bar{x}_2 = 97.3\ \text{and}\ \hat{g}_{OLP} = 0.050\ (= 5.0\%).$$

Substituting the same estimates as before for $\bar{X}$ and the (co)variances of the X_i and Y_i into (2.4) yields $\hat{\text{var}}(\hat{g}_{OLP}) = 0.00166$. Under the normality assumption this yields a smaller 95%-confidence interval $I_{OLP}^{95} \approx (-3.0\%, 13.0\%)$.

Example 2.2. Among the data of Example 2.1 there were three enterprises with extreme g-values of -50%, 133% and -91%. It is beyond the scope of this paper to further analyse or correct these outliers. But

to illustrate the difference between the estimators $\hat{g}_{STN}$ and $\hat{g}_{OLP}$ once more, we simply omit these enterprises so that $n_2 = 54$ instead of $n_2 = 57$. A first result is that estimate $\hat{\rho}_{xy2}$ increases from 0.876 to 0.970. The latter is fairly high in spite of the fact that the coefficient of variation of the growth rates $g_i = (Y_i / X_i - 1)$ is $cv_{g2} = s_{g2} / \bar{g}_2 = 4.1$ which still indicates a rather high volatility among the growth rates in this example. Furthermore, in analogy with the previous example, we get $\hat{g}_{STN} = 0.074\ (= 7.4\%)$ with $\text{var}(\hat{g}_{STN}) = 0.00251$ and $\hat{g}_{OLP} = 0.039\ (= 3.9\%)$ with $\text{var}(\hat{g}_{OLP}) = 0.00039$. The corresponding 95%-confidence intervals in this slightly modified example are approximately $I^{95}_{STN} \approx (\text{-}2.4\%, 17.2\%)$ and $I^{95}_{OLP} \approx (0.1\%, 7.7\%)$. Compared to Example 2.1 the interval I^{95}_{OLP} decreased relatively stronger than I^{95}_{STN}.

In addition, Example 2.2 may serve as a warning to be cautious when using sample means as $\bar{y}_{23}$ and $\bar{x}_{12}$ for estimating growth rates because these estimates may lead to unnecessarily large confidence interval around a suboptimal estimate. In the next section we look more closely at the question of what kind of circumstances may lead to a large interval I^{95}_{STN}.

3 Reasons for a large interval I^{95}_{STN}

In order to get more insight into the difference between $\text{var}(\hat{g}_{OLP})$ and $\text{var}(\hat{g}_{STN})$, we assume $n_{12} = n_{23} = n$ and $G,\ S_{xy} > 0$; hence, $\lambda = \mu = n_2 / n$. Then subtracting (2.4) from (2.2) yields

$$\begin{aligned}
\text{var}(\hat{g}_{STN}) - \text{var}(\hat{g}_{OLP}) &\approx \frac{1}{\bar{X}^2}\left\{2G\left(\frac{1}{n_2} - \frac{\lambda}{n}\right)S_{xy} - \left(\frac{1}{n_2} - \frac{1}{n}\right)\left(S_y^2 + G^2 S_x^2\right)\right\} \\
&= \frac{1}{\lambda n \bar{X}^2}\left\{2G(1-\lambda^2)S_{xy} - (1-\lambda)\left(S_y^2 + G^2 S_x^2\right)\right\} \\
&= \frac{1-\lambda}{\lambda n \bar{X}^2}\left(2G\lambda S_{xy} - S^2_{y-Gx}\right). \qquad (3.1)
\end{aligned}$$

In other words, $\text{var}(\hat{g}_{OLP})$ is smaller than $\text{var}(\hat{g}_{STN})$ when $\lambda > S^2_{y-Gx} / 2GS_{xy}$ provided $S_{xy} > 0$. Assuming $S_y^2 = S_x^2$, Qualité and Tillé (2008) derive a similar result for the parameter of *absolute* change when $\lambda > (1 - \rho_{xy}) / \rho_{xy}$. An anonymous referee pointed out that $\lambda < (1 - \rho_{xy}) / \rho_{xy}$ is a sufficient condition for $\text{var}(\hat{g}_{OLP}) > \text{var}(\hat{g}_{STN})$ because (3.1) can be rewritten as

$$\frac{(1-\lambda)GS_xS_y}{\lambda n \bar{X}^2}\left(2\lambda\rho_{xy} + 2\rho_{xy} - \frac{S_y^2 + G^2S_x^2}{GS_xS_y}\right) \le \frac{(1-\lambda)GS_xS_y}{\lambda n \bar{X}^2}\left(2\lambda\rho_{xy} + 2\rho_{xy} - 2\right) < 0,$$

provided that $\lambda < (1 - \rho_{xy}) / \rho_{xy}$.

If N is sufficiently large, a weaker condition can be derived under some standard model assumptions. Suppose that the data satisfy the model $Y_i = BX_i + u_i$ with $E(u_i) = 0$, $E(u_i^2) = \sigma^2 X_i^{\delta}$ and $E(u_i u_j) = 0$ $(i \neq j)$; recall X_i is not random in this context. Under this model, we make the (weak) assumptions (i) $G = S_{yx} / S_x^2$ and (ii) $S^2_{y-Gx} = S_y^2(1 - \rho_{xy}^2)$. To justify these assumptions, recall from regression theory that

$\hat{B} = S_{yx}/S_x^2$ can be seen as the unbiased, consistent estimator for B from an ordinary least squares (OLS) regression of Y_i on X_i and a *constant* $(i = 1,...,N)$. Furthermore, the corresponding OLS estimator $\left(\bar{Y} - \hat{B}\bar{X}\right)$ for the *constant* has zero expectation under the above model while its variance is of order $1/N$. Hence, $0 = \text{plim}\left(\bar{Y} - \hat{B}\bar{X}\right) = \text{plim}\left\{\bar{X}\left(G - \hat{B}\right)\right\}$ as $N \to \infty$ and provided $\bar{X} > c > 0$ for all N, we get the somewhat counterintuitive result $\text{plim}\left(G - \hat{B}\right) = 0$. In fact, it can be shown that

$$G = \bar{Y}/\bar{X} = \hat{B}\left[1 + O_p\left(1/\sqrt{N}\right)\right] = \left(S_{yx}/S_x^2\right)\left[1 + O_p\left(1/\sqrt{N}\right)\right]$$

as $N \to \infty$. This justifies assumption (i); for further details, see the end of this section. Furthermore, $S_y^2\left(1 - \rho_{xy}^2\right)$ can be seen as the (unexplained) variance of the residuals from the OLS regression. However, under the above model assumptions, these residuals are asymptotically equal to $Y_i - GX_i$ from which the *approximate* validity of (ii) follows. In addition, noting that $S_y^2\rho_{xy}^2$ is the so-called *explained* variance of the above OLS regression, it follows from assumption (i) that $S_y^2\rho_{xy}^2 = \hat{B}^2 S_x^2 \approx G^2 S_x^2$. Combining this with assumptions (i) and (ii), we can rewrite (3.1) as

$$\begin{aligned} \text{var}\left(\hat{g}_{STN}\right) - \text{var}\left(\hat{g}_{OLP}\right) &\approx \frac{1-\lambda}{\lambda n\bar{X}^2}\left\{2G^2\lambda S_x^2 - \left(1-\rho_{xy}^2\right)S_y^2\right\} \\ &\approx \frac{(1-\lambda)S_y^2}{\lambda n\bar{X}^2}\left(2\lambda\rho_{xy}^2 - 1 + \rho_{xy}^2\right) \\ &= \frac{(1-\lambda)S_y^2}{\lambda n\bar{X}^2}\left\{\rho_{xy}^2\left(1+2\lambda\right) - 1\right\}. \end{aligned} \tag{3.2}$$

Hence, $\text{var}\left(\hat{g}_{OLP}\right)$ is larger than $\text{var}\left(\hat{g}_{STN}\right)$ when

$$\lambda < \left(1-\rho_{xy}^2\right)/2\rho_{xy}^2 \quad \left[> \left(1-\rho_{xy}\right)/\rho_{xy}\right]. \tag{3.3}$$

Thus for say $\rho_{xy} = 0.9$, $\text{var}\left(\hat{g}_{OLP}\right)$ is under the above model for sufficiently large N larger than $\text{var}\left(\hat{g}_{STN}\right)$ when $\lambda < 0.117$, and for say $\rho_{xy} = 0.75$ when $\lambda < 0.389$. In addition, applying (3.2) to the data in Example 2.1 with $\lambda \approx 57/73 = 0.78$ and $\rho_{xy} = 0.876$ yields as approximation for the difference between both variances 0.0017 which is not very different from the actual difference of 0.0016 (=0.00324-0.00166) in the example. For Example 2.2, taking $\lambda = 54/70 = 0.77$ and $\rho_{xy} = 0.970$, applying (3.2) yields 0.00226 instead of 0.00212 (=0.00251-0.00039) in the example.

Under the above assumptions, it can also be shown that the ratio, say Q, of $\text{var}\left(\hat{g}_{OLP}\right)$ and $\text{var}\left(\hat{g}_{STN}\right)$ can be approximated by

$$Q = \frac{\text{var}\left(\hat{g}_{OLP}\right)}{\text{var}\left(\hat{g}_{STN}\right)} \approx \left(\lambda^{-1} - f\right)\left(1 - f + 2\left(1-\lambda\right)\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}, \tag{3.4}$$

irrespective of the values of S_y^2 and S_x^2; f stands for n/N. For a proof of (3.4), see Appendix A.1. From (3.4) it can be seen that Q and $\text{var}\left(\hat{g}_{OLP}\right)$ tend to zero as ρ_{xy}^2 tends to unity, provided N is sufficiently large and $\lambda < 1$.

It should be noted that in practice the correlations ρ_{xy} often are rather high by the very nature of the data (Y_i, X_i). That is, a large (small) enterprise in period $(t-12)$ is in most cases still large (small) after 12 months; Knottnerus and Van Delden (2012, page 47) found for various strata an overall mean correlation of 0.90 and a variance of 0.0074. So it appears that $\text{var}(\hat{g}_{STN})$ is more affected by a decrease of λ than $\text{var}(\hat{g}_{OLP})$ unless λ is extremely low because (i) $\text{var}(\hat{g}_{OLP}) = \text{var}(\hat{g}_{STN})$ when $\lambda = 1$ and (ii) Q is large when ρ_{xy}^2 is large. For example, when $\rho_{xy} = 0.9$ and $f = 0.1$ a decrease of λ from 0.9 to 0.5 leads to a decrease of Q from 0.58 to 0.37; recall $Q = 1$ when $\lambda = 1$. This emphasizes once more the importance of avoiding panel attrition when using estimator $\hat{g}_{STN}$ while N is large.

A natural question that remains to be answered is when is N sufficiently large. To answer this question, consider the difference $\Delta \equiv \hat{B} - G$ and its variance, say σ_Δ^2. The difference Δ can be written as

$$\begin{aligned}\Delta &= \frac{S_{xy}}{S_x^2} - \frac{\overline{Y}}{\overline{X}} = \frac{1}{N-1}\sum_{i\in U}\frac{X_i - \overline{X}}{S_x^2}Y_i - \frac{1}{N}\sum_{i\in U}\frac{Y_i}{\overline{X}} \\ &\approx \frac{1}{N}\sum_{i\in U}\left(\frac{X_i - \overline{X}}{S_x^2} - \frac{1}{\overline{X}}\right)Y_i \\ &= \frac{1}{N}\sum_{i\in U} M_i U_i \quad \left(M_i = \frac{X_i - \overline{X}}{S_x^2} - \frac{1}{\overline{X}}\right).\end{aligned}$$

In the second line we assumed $N >> 1$ and in the last line we used the model assumption $Y_i = BX_i + U_i$. Next, assuming $\text{var}(U_i) = \sigma^2 X_i^{\delta}$, we get

$$\sigma_\Delta^2 \equiv \text{var}\left(\hat{B} - G\right) = \frac{\sigma^2}{N^2}\sum_{i\in U} M_i^2 X_i^{\delta}.$$

This variance can be estimated by

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}^2}{Nn_2}\sum_{i\in s_2}\hat{m}_i^2 X_i^{\hat{\delta}},$$

where

$$\hat{m}_i = \frac{X_i - \overline{x}_2}{s_{x2}^2} - \frac{1}{\overline{x}_2}, \qquad \hat{\sigma}^2 = \frac{1}{n_2 - 1}\sum_{i\in s_2}\left(Y_i - \frac{\overline{y}_2}{\overline{x}_2}X_i\right)^2 \Big/ X_i^{\hat{\delta}}$$

and $\hat{\delta}$ is an estimate from the OLS regression

$$\ln\left(Y_i - \frac{\overline{y}_2}{\overline{x}_2}X_i\right)^2 = \alpha + \delta \ln X_i + w_i \qquad (i = 1,...,n_2);$$

units with $Y_i = \overline{y}_2 X_i / \overline{x}_2$ are omitted. Based on $\hat{\sigma}_\Delta^2$, one may call N sufficiently large if the outcome of (3.1) will not severely be affected by replacing G by $G + \hat{\sigma}_\Delta$. In addition, it should be borne in mind that

relationships for very large N are probably still a reasonably appropriate indication for what may occur when N is not very large.

4 Composite estimator for the growth rate

Examining a composite estimator (COM) of the form

$$\hat{g}_{COM} = k\hat{g}_{STN} + (1-k)\hat{g}_{OLP}, \tag{4.1}$$

it follows from minimizing $\text{var}(\hat{g}_{COM})$ with respect to k that

$$k = \frac{\text{var}(\hat{g}_{OLP}) - \text{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}{\text{var}(\hat{g}_{OLP}) + \text{var}(\hat{g}_{STN}) - 2\,\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}; \tag{4.2}$$

see also Särndal et al. (1992, page 372). Note that, by construction, $\text{var}(\hat{g}_{COM})$ can not exceed $\min\{\text{var}(\hat{g}_{STN}), \text{var}(\hat{g}_{OLP})\}$.

Using the linearized forms of the estimators $\hat{g}_{OLP}$ and $\hat{g}_{STN}$, we get for their covariance

$$\begin{aligned}\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN}) &\approx \text{cov}\left(\frac{\bar{y}_2 - G\bar{x}_2}{\bar{X}}, \frac{\bar{y}_{23} - G\bar{x}_{12}}{\bar{X}}\right) \\ &= \frac{1}{\bar{X}^2}\left\{\text{cov}(\bar{y}_2, \bar{y}_{23}) - G\,\text{cov}(\bar{y}_2, \bar{x}_{12}) - G\,\text{cov}(\bar{x}_2, \bar{y}_{23}) + G^2\,\text{cov}(\bar{x}_2, \bar{x}_{12})\right\}.\end{aligned}$$

Now using some results from Knottnerus (2003, page 377)

$$\text{cov}(\bar{y}_2, \bar{y}_{23}) = \text{var}(\bar{y}_{23}) \quad \left[= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) S_y^2\right]$$

$$\text{cov}(\bar{x}_2, \bar{y}_{23}) = \text{cov}(\bar{x}_{23}, \bar{y}_{23}) \quad \left[= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) S_{xy}\right],$$

we obtain

$$\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN}) \approx \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n_{23}} - \frac{1}{N}\right)\left(S_y^2 - GS_{yx}\right) + \left(\frac{1}{n_{12}} - \frac{1}{N}\right)\left(G^2 S_x^2 - GS_{yx}\right)\right\}. \tag{4.3}$$

In practice k can be estimated by replacing all (co)variances in (4.2) by their sample estimates, yielding

$$\hat{k} = \frac{\hat{\text{var}}(\hat{g}_{OLP}) - \hat{\text{cov}}(\hat{g}_{OLP}, \hat{g}_{STN})}{\hat{\text{var}}(\hat{g}_{OLP}) + \hat{\text{var}}(\hat{g}_{STN}) - 2\,\hat{\text{cov}}(\hat{g}_{OLP}, \hat{g}_{STN})} \tag{4.4}$$

To illustrate this approach, consider the following example.

Example 4.1. The data are the same as for Example 2.1. Applying formulas (2.1) - (2.4) and (4.3) to these data yields

$$\hat{g}_{STN} = 0.082 \ (0.00254), \ \ \hat{g}_{OLP} = 0.050 \ (0.00134), \ \text{and}$$
$$\hat{\text{cov}}\left(\hat{g}_{STN}, \hat{g}_{OLP}\right) = 0.00097.$$

The variances are mentioned between parentheses. Substituting these estimates into (4.4) yields $\hat{k} = 0.191$ and subsequently, $\hat{g}_{COM} = 0.056$ (0.00127). For the ease of exposition, all (co)variances in (4.4) are estimated from overlap s_2, including the estimates of G and $\bar{X}$ in (2.2), (2.4) and (4.3). Furthermore, using these estimates, we found that $\text{var}\left(\hat{g}_{STN}\right) < \text{var}\left(\hat{g}_{OLP}\right)$ and $k > 0.5$ only if $n_2 \leq 12 \ \left(\lambda \leq 0.167\right)$.

For the sake of completeness, we also give an example for the composite estimator for the parameter of absolute change (i.e., $\bar{D} = \bar{Y} - \bar{X}$).

Example 4.2. We use the same data as in Example 2.1. As before all estimates for the (co)variances are based on s_2. Define D_i by $D_i = Y_i - X_i$. Then we have two estimators for the parameter of absolute change

$$\hat{\bar{D}}_{STN} = \bar{y}_{23} - \bar{x}_{12} = 7.35 \quad \text{and} \quad \hat{\bar{D}}_{OLP} = \bar{d}_2 = \bar{y}_2 - \bar{x}_2 = 4.89.$$

For the (co)variances of $\hat{\bar{D}}_{STN}$ and $\hat{\bar{D}}_{OLP}$ we get

$$\begin{aligned}
\hat{\text{var}}\left(\hat{\bar{D}}_{STN}\right) &= \hat{\text{var}}\left(\bar{y}_{23}\right) + \hat{\text{var}}\left(\bar{x}_{12}\right) - 2\,\hat{\text{cov}}\left(\bar{y}_{23}, \bar{x}_{12}\right) \\
&= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) s_{y2}^2 + \left(\frac{1}{n_{12}} - \frac{1}{N}\right) s_{x2}^2 - 2\left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right) s_{xy2} = 23.58 \\
\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) &= \left(\frac{1}{n_2} - \frac{1}{N}\right) s_{y-x,2}^2 = 13.11
\end{aligned}$$

$$\begin{aligned}
\hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right) &= \hat{\text{cov}}\left(\bar{y}_{23} - \bar{x}_{12}, \bar{y}_2 - \bar{x}_2\right) \\
&= \left(\frac{1}{n_{23}} - \frac{1}{N}\right)\left(s_{y2}^2 - s_{xy2}\right) - \left(\frac{1}{n_{12}} - \frac{1}{N}\right)\left(s_{xy2} - s_{x2}^2\right) = 9.46.
\end{aligned}$$

In analogy with (4.4) we now obtain

$$\hat{k} = \frac{\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) - \hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)}{\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) + \hat{\text{var}}\left(\hat{\bar{D}}_{STN}\right) - 2\,\hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)} = 0.206$$

and consequently, $\hat{\bar{D}}_{COM} = 5.40\,(12.37)$.

Note that $\hat{g}_{COM}$ can be rewritten as

$$\begin{aligned}
\hat{g}_{COM} &= \hat{g}_{OLP} + \hat{k}\left(\hat{g}_{STN} - \hat{g}_{OLP}\right) \\
&\approx \hat{g}_{OLP} + k\left(\hat{g}_{STN} - \hat{g}_{OLP}\right),
\end{aligned}$$

where we used a first-order Taylor series approximation of $\hat{g}_{COM}$. Therefore, the random character of estimator $\hat{k}$ can be neglected for estimating $\text{var}\left(\hat{g}_{COM}\right)$. The error thus introduced is of order $1/n_2$ as

$n_2 \to \infty$ and $\hat{g}_{COM}$ is asymptotically unbiased. Recall that the standard procedure for estimating the variance of the ratio estimator or the regression estimator is based on a first-order Taylor series approximation as well.

In addition, under the same assumptions as (3.4), it can be shown that for sufficiently large N,

$$k = \left(1 + \frac{2\lambda\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}; \tag{4.5}$$

for a proof of (4.5), see Appendix A.1. From (4.5) it can be seen that k is decreasing in λ. So we have the somewhat counterintuitive result that k is decreasing in λ whereas according to (3.1), ratio Q in (3.4) is a convex function of λ; recall that $\text{var}\left(\hat{g}_{STN}\right) = \text{var}\left(\hat{g}_{OLP}\right)$ and, consequently, $Q = 1$ for $\lambda = 1$ and $\lambda = S_{y-Gx}^2 \big/ 2GS_{xy}$.

5 Aligned composite estimators for growth rates and totals

So far we only looked at growth rates because in practice the estimate $\hat{X}_{SRS}$ for the turnover of 12 months ago can be considered more or less as fixed (i.e., can not be changed anymore). When X refers to the total turnover in month $(t-1)$, it is likely that the figures for the preceding month can still be improved and modified. In such a situation the initial estimate $\hat{X}_{SRS}$ might be revised as well.

Before examining a multivariate composite estimator for growth rates and totals, we first look at a multivariate composite estimator for the parameter of absolute change and the corresponding population means or totals; also see Example 4.2. Define the initial vector estimator $\hat{\theta}_0$ by $\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$. Denote the underlying parameter vector to be estimated by $\theta = \left(\theta_1, \theta_2, \theta_3\right)'$. Let V_0 denote the covariance matrix of $\hat{\theta}_0$. In terms of θ the problem is now to find an aligned composite estimator $\hat{\theta}_{AC}$ with elements satisfying the prior restriction $\theta_1 - \theta_2 + \theta_3 = 0$ or, equivalently, $\bar{D} - \bar{Y} + \bar{X} = 0$ or $D - Y + X = 0$. Although there is one restriction in this situation, we treat in this section the somewhat more general case with m restrictions $\left(1 \le m \le 3\right)$. When the prior restrictions are of the linear form $c - R\theta = 0$ where R is a $m \times 3$ matrix of rank m $(m \le 3)$, the optimal unbiased composite estimator for θ is equal to the general restriction (GR) estimator

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right) \tag{5.1}$$

$$K = V_0 R'\left(RV_0 R'\right)^{-1}$$

$$V_{GR} \equiv \text{cov}\left(\hat{\theta}_{GR}\right) = \left(I_3 - KR\right)V_0, \tag{5.2}$$

where I_3 stands for the 3×3 identity matrix. The estimator $\hat{\theta}_{GR}$ is optimal in the sense that when $\hat{\theta}_0$ follows a multivariate normal distribution $N\left(\theta, V_0\right)$, the likelihood of $\hat{\theta}_0$ attains its maximum, under the constraint $c - R\theta = 0$, for $\theta_{\max} = \hat{\theta}_{GR}$. Moreover, given the form $\hat{\theta}_K = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right)$, it can be shown

that minimizing $\mathrm{tr}\left\{\mathrm{cov}\left(\hat{\theta}_K\right)\right\}$ with respect to the $3\times m$ matrix K leads to (5.2). Recall that this means that for any other matrix K the corresponding covariance matrix $\mathrm{cov}\left(\hat{\theta}_K\right)$ exceeds V_{GR} by a positive semidefinite matrix; see Magnus and Neudecker (1988, pages 255-256). For further details on the GR estimator, see Knottnerus (2003, pages 328-332). To illustrate how (5.1) and (5.2) can be used for obtaining an aligned composite (AC) estimator $\hat{\theta}_{AC}$, consider the following example dealing with the estimation of two population means and their difference.

Example 5.1. We use the same data as in Examples 2.1 and 4.2. The initial vector $\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$ is given by $\left(4.89, 97.19, 89.84\right)'$. These estimates do not satisfy the restriction $\theta_1 - \theta_2 + \theta_3 = 0$; note that $R = (1, -1, 1)$ and $c = 0$. Most elements of V_0 have already been discussed. Similar to Example 4.2, for element $\mathrm{cov}\left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}\right)$ we get

$$\begin{aligned}\mathrm{cov}\left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}\right) &= \mathrm{cov}\left(\bar{y}_2 - \bar{x}_2, \bar{y}_{23}\right) \\ &= \mathrm{var}\left(\bar{y}_{23}\right) - \mathrm{cov}\left(\bar{x}_{23}, \bar{y}_{23}\right).\end{aligned} \tag{5.3}$$

Each term in (5.3) can be estimated from s_2 as described before. The other covariances in V_0 have a similar form and can be estimated in the same manner. The variance estimates for $\hat{\bar{D}}_{OLP}$, $\bar{y}_{23}$ and $\bar{x}_{12}$ are 13.12, 38.79 and 22.92, respectively. Next, applying (5.1) and (5.2) with K replaced by $\hat{K} = \hat{V}_0 R'\left(R\hat{V}_0 R'\right)^{-1}$, we obtain the following aligned composite AC estimates

$$\hat{\bar{D}}_{AC} = 5.40\ (12.37), \quad \hat{\bar{Y}}_{AC} = 96.28\ (36.32), \quad \text{and} \quad \hat{\bar{X}}_{AC} = 90.88\ (19.75).$$

Between parentheses the variances are mentioned.

Now three remarks are in order. Firstly, $\hat{\bar{D}}_{COM}$ discussed in the preceding section can also be derived from (5.1) and (5.2) by choosing $\hat{\theta}_0 = \left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)'$ with prior restriction $\theta_1 - \theta_2 = 0$. Secondly, by construction, the estimator $\hat{\bar{D}}_{AC}$ is equal to estimator $\hat{\bar{D}}_{COM}$ and, consequently, they have the same variance. Thirdly, were K known, then the AC estimator would be unbiased. But because K is to be replaced by $\hat{K}$, the AC estimator $\hat{\theta}_{AC}$ is only asymptotically unbiased. The same remark applies to the estimator $\left(I_3 - \hat{K}R\right)\hat{V}_0$ of $\mathrm{cov}(\hat{\theta}_{AC})$. Similar to $\hat{\theta}_{COM}$ described in the preceding section, the bias of $\hat{\theta}_{AC}$ is of order $O(1/n_2)$; for the relationship between $\hat{\theta}_{AC}$ and the regression estimator, see Appendix A.3.

In case of m nonlinear restrictions, say $c - R(\theta) = 0$, a first-order Taylor series approximation around $\theta = \hat{\theta}_0$ yields $c - R\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\left(\theta - \hat{\theta}_0\right) = 0$ or, equivalently,

$$c\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\theta = 0, \quad \text{where} \quad c\left(\hat{\theta}_0\right) = c - R\left(\hat{\theta}_0\right) + D_R\left(\hat{\theta}_0\right)\hat{\theta}_0. \tag{5.4}$$

$D_R(\theta)$ stands for the $m \times 3$ matrix of partial derivatives of $R(\theta)$ $\left(\text{i.e., } D_R(\theta) = \partial R(\theta)/\partial \theta'\right)$. Subsequently, an iterative procedure can be carried out by repeatedly applying (5.1) and (5.2) to the updated linearized versions of the nonlinear restrictions $c - R(\theta) = 0$. This yields

$$\left.\begin{aligned} \hat{\theta}_h &= \hat{\theta}_0 + K_h \hat{e}_h; \\ \hat{e}_h &= c_h - D_h \hat{\theta}_0; \\ K_h &= V_0 D_h' \left(D_h V_0 D_h'\right)^{-1}; \\ \text{cov}(\hat{\theta}_h) &= \left(I_3 - K_h D_h\right) V_0; \\ D_h &= D_R\left(\hat{\theta}_{h-1}\right); \\ c_h &= c - R\left(\hat{\theta}_{h-1}\right) + D_h \hat{\theta}_{h-1} \qquad (h = 1,2,...). \end{aligned}\right\} \tag{5.5}$$

For further details, see Appendix A.2 and Knottnerus (2003, pages 351-354). Note that the first equation can be seen as an update of $\hat{\theta}_0$ rather than of $\hat{\theta}_{h-1}$. This is an important difference with the celebrated Kalman equations; see Kalman (1960). In the present context, the vectors $\hat{\theta}_{h-1}$ are only used in a numerical procedure for finding new (better) Taylor series approximations of the nonlinear restrictions $c - R(\theta) = 0$ around $\theta = \hat{\theta}_{h-1}$ $(h = 1,2,...)$ until convergence is reached. Furthermore, note that $\hat{e}_h$ can be seen as a $m-$vector of restriction errors when substituting $\theta = \hat{\theta}_0$ into the linearized restrictions around $\theta = \hat{\theta}_{h-1}$. To illustrate the use of the Kalman-like equations in (5.5) for deriving aligned composite estimators for growth rates and totals, consider the following example.

Example 5.2. We use the same data as in Example 4.1. The initial vector $\hat{\theta}_0$ is now defined by $\hat{\theta}_0 = \left(\hat{G}_{OLP}, \overline{y}_{23}, \overline{x}_{12}\right)'$ and is given by $(1.050, 97.191, 89.840)'$. These estimates do not satisfy the (nonlinear) prior restriction $\theta_2 - \theta_1 \theta_3 = 0$ $(m = 1)$. All elements of V_0 and their estimation have already been discussed. For the $(h+1)$-th recursion $R\left(\hat{\theta}_h\right)$ and the 1×3 matrix D_{h+1} are given by

$$R\left(\hat{\theta}_h\right) = \left(\hat{\theta}_{h2} - \hat{\theta}_{h1}\hat{\theta}_{h3}\right)$$
$$D_{h+1} = \left(-\hat{\theta}_{h3} \quad 1 \quad -\hat{\theta}_{h1}\right),$$

respectively; $\hat{\theta}_{hk}$ is the k-th element of vector $\hat{\theta}_h$ $(1 \le k \le 3)$. Recall V_0 and $\hat{V}_0$ remain unchanged for all recursions. The first recursion from (5.5) yields

$$\hat{\theta}_1 = (1.0544,\ 95.945,\ 91.000)'.$$

The (nonlinear) restriction is almost satisfied, that is, $R\left(\hat{\theta}_1\right) = -0.005$. The second recursion yields the following aligned composite (AC) estimates

$$\hat{G}_{AC} = 1.0544\ (0.00130), \quad \hat{\overline{Y}}_{AC} = 95.947\ (35.55), \quad \text{and} \quad \hat{\overline{X}}_{AC} = 90.998\ (19.85).$$

Between parentheses the variances are mentioned. The (absolute) error of the second restriction further decreased, that is, $R\left(\hat{\theta}_2\right) = -0.001$ and we stopped the recursions. Due to the nonlinearity of the restriction, the estimates of $\hat{G}_{AC}$ and its variance are slightly different from those of $\hat{G}_{COM}$ and its variance in Example 4.1.

It is noteworthy that in Example 5.2 $\hat{G}_{AC}$ is not much different of $\hat{G}_{OLP}$ (=1.050). A related method for estimating totals is the so-called matched pair (MP) method; see Smith et al. (2003, page 269-271). The original MP method is purely based on $\hat{G}_{OLP}$ (in our notation) between months t and $t-1$ and used by ONS for estimating the monthly retail sales index. In a simulation study the authors found that the MP method gives a good performance for the short-term growth rates but for terms of more than 15 months the performance was worsening with respect to the bias. The bias could be corrected by benchmarking to growth rates on a regular basis. Another drawback of the MP method seems to be that a formula for the variance of the MP estimator is (still) lacking. In the next section we describe an extension of the AC estimator for incorporating auxiliary information into the AC estimation procedure.

6 Extensions

In this section we briefly discuss a number of extensions of the AC estimator described in the preceding section. Firstly, we pay attention to the situation whereby regression estimators, say $\hat{\bar{Y}}_{REG,k}$ and $\hat{\bar{X}}_{REG,k}$, are used instead of SRS estimators $(k = 2,\ 12 \text{ and } 23)$. To avoid a notational burden, we look at the situation with one explanatory variable, say z; a generalization for more auxiliaries is straightforward. Furthermore, for simplicity's sake, we assume that the estimated regression coefficients, denoted by b_{yz2} and b_{xz2}, stem from s_2. In order to derive the aligned composite estimators in this situation, we only need to evaluate (co)variance terms of the form $\text{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{X}}_{REG,l}\right)$ in the different formulas $(k,\ l = 2,\ 12 \text{ and } 23)$. This evaluation can be done as follows. Replace the Y_i and X_i in the formulas by the corresponding (estimated) residuals from a regression on Z_i and a *constant*. That is,

$$\text{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{X}}_{REG,l}\right) = \text{cov}\left(\bar{y}_k^*, \bar{x}_l^*\right), \tag{6.1}$$

where the (estimated) residual variables Y_i^* and X_i^* are defined by

$$Y_i^* = Y_i - \bar{y}_k - b_{yz2}\left(Z_i - \bar{z}_k\right) = Y_i - b_{yz2}Z_i + const.$$
$$X_i^* = X_i - \bar{x}_l - b_{xz2}\left(Z_i - \bar{z}_l\right) = X_i - b_{xz2}Z_i + const.$$

The term $\text{cov}\left(\bar{y}_k^*, \bar{x}_l^*\right)$ on the right-hand side of (6.1) can be calculated in the same manner as $\text{cov}\left(\bar{y}_k, \bar{x}_l\right)$, discussed in preceding sections; see also formula (A.8) in Appendix A.3 and recall $\text{var}\left(\hat{\bar{Y}}_{REG,k}\right) = \text{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{Y}}_{REG,k}\right)$. In addition, the same approach can be applied when use is made of ratio estimators such as $\hat{\bar{Y}}_{R,k} = \bar{y}_k \bar{Z} / \bar{z}_k$ and $\hat{\bar{X}}_{R,l} = \bar{x}_l \bar{Z} / \bar{z}_l$. That is, the residual variables Y_i^* and X_i^* are now to be read as

$$Y_i^* = Y_i - \frac{\overline{y}_2}{\overline{z}_2} Z_i \quad \text{and} \quad X_i^* = X_i - \frac{\overline{x}_2}{\overline{z}_2} Z_i.$$

An alternative option for taking an auxiliary variable into account is to extend both the parameter vector θ and the set of prior restrictions. For instance, in Example 5.1 the parameter θ was implicitly defined by $\theta = \left(\overline{D}, \overline{Y}, \overline{X}\right)'$. When the variable z is observed in samples 12 and 23, the new, extended $\hat{\theta}_0$ is given by

$$\hat{\theta}_0 = \left(\hat{\overline{D}}_{OLP}, \overline{y}_{23}, \overline{x}_{12}, \overline{z}_{23}, \overline{z}_{12}, \overline{z}_2\right)'$$

and the extended set of prior restrictions is

$$\begin{aligned}
&\theta_2 - \theta_1 - \theta_3 = 0; \\
&\theta_4 - \theta_5 = 0; \\
&\theta_4 - \theta_6 = 0; \\
&\theta_4 = \overline{Z}.
\end{aligned}$$

Hence, the new c is $c = \left(0, 0, 0, \overline{Z}\right)'$. In this way the efficiency of $\hat{\theta}_0$ can be further improved.

Secondly, another extension regards births and deaths. With respect to deaths, the population in period $t-12$ can be divided into two (post)strata: one consisting of the deaths in period t and one consisting of the enterprises existing in periods $t-12$ and t. Using such a poststratification still leads to an asymptotically unbiased estimator for the population mean at period t, provided there are no births. In order to take births into account, one should draw an appropriate sample from this substratum of births especially when the number of births is substantial, and when there are no realistic assumptions with respect to the total turnover in this substratum in month t.

Finally, we examine the situation whereby a combination of quarterly and semesterly data is to be analysed. Suppose that in quarters 2, 4 and 6 semesterly samples are drawn which need not be the same as the quarterly samples in those quarters. In order to explain the AC estimator in this situation, consider six consecutive quarterly SRS estimates for the quarterly means of the turnover, say $\overline{y}_1, \overline{y}_2, \overline{y}_3, \overline{y}_4, \overline{y}_5, \overline{y}_6$, and three semesterly SRS estimates for the semesterly means of turnover, say $\overline{x}_2, \overline{x}_4$ and $\overline{x}_6$; note that the subscript refers to the quarter of observation and *not* to a sample set as before. Furthermore, suppose that the following growth ratios are to be estimated: $G_{62} = Y_6/Y_2$, $H_{62} = X_6/X_2$ and $H_{64} = X_6/X_4$ as well as the corresponding quarterly and semesterly totals. In order to obtain a consistent set of estimators for totals (means) and growth rates, define in analogy with the approach in Section 5

$$\hat{\theta}_0 = \left(\hat{G}_{62,OLP}, \hat{H}_{62,OLP}, \hat{H}_{64,OLP}, \overline{y}_1, \overline{y}_2, \overline{y}_3, \overline{y}_4, \overline{y}_5, \overline{y}_6, \overline{x}_2, \overline{x}_4, \overline{x}_6\right)'.$$

The corresponding set of restrictions is

$$\begin{aligned}
\theta_9 - \theta_1\theta_5 &= \overline{Y}_6 - G_{62}\overline{Y}_2 = 0 \\
\theta_{12} - \theta_2\theta_{10} &= \overline{X}_6 - H_{62}\overline{X}_2 = 0 \\
\theta_{12} - \theta_3\theta_{11} &= \overline{X}_6 - H_{64}\overline{X}_4 = 0 \\
\theta_4 + \theta_5 - \theta_{10} &= \overline{Y}_1 + \overline{Y}_2 - \overline{X}_2 = 0 \\
\theta_6 + \theta_7 - \theta_{11} &= \overline{Y}_3 + \overline{Y}_4 - \overline{X}_4 = 0 \\
\theta_8 + \theta_9 - \theta_{12} &= \overline{Y}_5 + \overline{Y}_6 - \overline{X}_6 = 0.
\end{aligned}$$

The matrix V_0 can be estimated in a similar manner as described in Sections 2 and 4.

7 Conclusions and discussion

This section summarizes a number of conclusions and issues for further research.

When totals of turnover are estimated from a panel in months t and $t-12$, two estimators $\hat{g}_{STN}$ and $\hat{g}_{OLP}$ for the growth rate between these months can be distinguished.

When using $\hat{g}_{STN}$, one should be aware that in practice, $\text{var}(\hat{g}_{OLP})$ might be much smaller than $\text{var}(\hat{g}_{STN})$ especially when the turnover in month $t-12$ and the turnover in month t are highly correlated and the overlap ratios λ and μ are not too small.

The efficiency of $\hat{g}_{STN}$ and $\hat{g}_{OLP}$ can be improved by the composite estimator $\hat{g}_{COM}$ described in Section 4.

Using least squares techniques, an aligned composite vector-estimator $\left(\hat{g}_{AC}, \hat{Y}_{AC}, \hat{X}_{AC}\right)'$ can be derived that obeys the nonlinear restriction for totals and growth rates: $\hat{Y}_{AC} = \left(1+\hat{g}_{AC}\right)\hat{X}_{AC}$.

The AC estimator subject to *linear* restrictions can be extended in several ways: (i) for nonlinear restrictions, (ii) for different data sets such as monthly, quarterly and yearly data, (iii) for births and deaths, (iv) for regression and ratio estimators, and (v) for additional auxiliary variables.

Similar to the regression estimator, the AC estimator is asymptotically unbiased. This remark also applies to the covariance-matrix estimator $\left(I_k - \hat{K}R\right)\hat{V}_0$.

There is not yet an unambiguous answer on the question of to what extent data from the past should be included in the vector estimate $\hat{\theta}_0$ each month. The answer depends upon: (i) the NSI's policy and rules with respect to revision of already published figures, (ii) the fact that from a theoretical viewpoint, the sequence of T monthly SRS estimates $\bar{y}_1, \bar{y}_2, ..., \bar{y}_T$ (included as component in $\hat{\theta}_0$) should be so long that the difference between the two AC estimators of $\bar{Y}_1$, say $\hat{\bar{Y}}_{1AC}^{T}$ and $\hat{\bar{Y}}_{1AC}^{T+1}$, is not substantial, and (iii) the size of the samples. That is, in analogy with the regression estimator or, equivalently, the calibration estimator, the sample sizes should be much larger than the number of (calibration) restrictions. For a simulation study on the variance of the regression estimator and the number of regressors, see Silva and Skinner (1997) and for a relationship between the regression estimator and the GR estimator, see Appendix A.3 and Knottnerus (2003).

In the specific case of estimating mutually aligned totals and changes, additional research is needed for finding: (i) the optimal and practical length for the monthly, quarterly, semesterly and yearly series of SRS estimates to be included in the initial vector $\hat{\theta}_0$ and (ii) a rule of thumb with respect to the number of restrictions compared to the sample sizes in order to find an AC estimator $\hat{\theta}_{AC}$ with an improved efficiency.

Acknowledgements

The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands. The author would like to thank Harm Jan Boonstra, Arnout van Delden, Sander Scholtus, the Associate Editor and two anonymous referees for their helpful comments and corrections.

Appendix

A.1 Proofs of (3.4) and (4.5)

The proof of (3.4) is as follows. For $n_{12} = n_{23} = n,$ formula (2.2) can be rewritten as

$$\text{var}\left(\hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}\right\} \tag{A.1}$$

Dividing (2.4) by (A.1) yields

$$\begin{aligned} Q = \frac{\text{var}\left(\hat{g}_{OLP}\right)}{\text{var}\left(\hat{g}_{STN}\right)} &\approx \frac{\left(\frac{1}{\lambda n}-\frac{1}{N}\right)S_{y-Gx}^2}{\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}} \\ &= \frac{\left(\lambda^{-1}-f\right)S_{y-Gx}^2}{(1-f)S_{y-Gx}^2 + 2(1-\lambda)GS_{xy}} \\ &= \left(\lambda^{-1}-f\right)\left(1-f+2(1-\lambda)\frac{GS_{xy}}{S_{y-Gx}^2}\right)^{-1} \\ &\approx \left(\lambda^{-1}-f\right)\left(1-f+2(1-\lambda)\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}. \end{aligned} \tag{A.2}$$

In the last line we used that under the model assumptions mentioned in Section 3, $GS_{xy} \approx \hat{B}^2 S_x^2 = \rho_{xy}^2 S_y^2$ and $S_{y-Gx}^2 \approx \left(1-\rho_{xy}^2\right)S_y^2,$ provided that N is sufficiently large; see also the derivation of (3.2).

Next, under the same assumptions, (4.5) can be derived as follows. Since $n_{12} = n_{23} = n,$ the covariance in (4.3) can be rewritten as

$$\text{cov}\left(\hat{g}_{OLP}, \hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2. \tag{A.3}$$

Combining (2.4), (A.1) and (A.3), we can write k in (4.2) as

$$\begin{aligned} k &\approx \frac{\left(\frac{1}{\lambda n}-\frac{1}{n}\right)S_{y-Gx}^2}{\left(\frac{1}{\lambda n}-\frac{1}{n}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}} \\ &= \left(1+\frac{2\lambda GS_{xy}}{S_{y-Gx}^2}\right)^{-1} \approx \left(1+\frac{2\lambda\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}. \end{aligned}$$

Similar to deriving (A.2), we used in the last line $GS_{xy}/S^2_{y-Gx} \approx \rho^2_{xy}/\left(1-\rho^2_{xy}\right)$.

A.2 Derivation of (5.5)

In case of m linear restrictions $c - R\theta = 0,$ matrix K can be found by minimizing

$$\min_K E\left[\left\{\theta-\hat{\theta}_0-K\left(c-R\hat{\theta}_0\right)\right\}'\left\{\theta-\hat{\theta}_0-K\left(c-R\hat{\theta}_0\right)\right\}\right];$$

see Knottnerus (2003, page 330). The solution of this least squares problem is given by

$$\begin{aligned} K &= E\left\{\left(\theta-\hat{\theta}_0\right)\left(c-R\hat{\theta}_0\right)'\right\}\left[\operatorname{cov}\left(c-R\hat{\theta}_0\right)\right]^{-1} \\ &= V_0R'\left(RV_0R'\right)^{-1}. \end{aligned} \tag{A.4}$$

In case of m nonlinear restrictions, the new minimand is

$$E\left[\left\{\theta-\hat{\theta}_0-K\left[c-R\left(\hat{\theta}_0\right)\right]\right\}'\left\{\theta-\hat{\theta}_0-K\left[c-R\left(\hat{\theta}_0\right)\right]\right\}\right].$$

Similarly to (A.4), it can be shown that this minimand attains its minimum for

$$K = E\left\{\left(\theta-\hat{\theta}_0\right)\left[c-R\left(\hat{\theta}_0\right)\right]'\right\}\left[\operatorname{cov}\left\{c-R\left(\hat{\theta}_0\right)\right\}\right]^{-1}. \tag{A.5}$$

Substituting Taylor's linearization $R\left(\hat{\theta}_0\right) \approx R(\theta) + D_R(\theta)\left(\hat{\theta}_0-\theta\right)$ into (A.5), we get the following approximation, say K_1, for K

$$\begin{aligned} K_1 &\approx V_0D_R'(\theta)\left[D_R(\theta)V_0D_R'(\theta)\right]^{-1} \\ &\approx V_0D_R'\left(\hat{\theta}_0\right)\left[D_R\left(\hat{\theta}_0\right)V_0D_R'\left(\hat{\theta}_0\right)\right]^{-1}. \end{aligned} \tag{A.6}$$

Assuming that $\hat{\theta}_0 \sim N\left(\theta, V_0\right),$ the first approximation for the constrained maximum likelihood (ML) solution, say $\hat{\theta}^{(1)}_{ML},$ can be calculated in the standard manner by using the linearized restrictions

$$\hat{\theta}^{(1)}_{ML} = \hat{\theta}_0 + K_1\left\{c\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\hat{\theta}_0\right\}, \tag{A.7}$$

where $c\left(\hat{\theta}_0\right)$ is defined by (5.4). If $\hat{\theta}^{(1)}_{ML}$ does not satisfy the nonlinear restrictions $c - R(\theta) = 0,$ a better approximation of K might be obtained by replacing $\hat{\theta}_0$ in (A.6) by update $\hat{\theta}^{(1)}_{ML}$ resulting in a new matrix K_2. In turn, in analogy with (A.7) K_2 leads to a better approximation or update of $\hat{\theta}_0,$ say $\hat{\theta}^{(2)}_{ML},$

$$\hat{\theta}^{(2)}_{ML} = \hat{\theta}_0 + K_2\left\{c\left(\hat{\theta}^{(1)}_{ML}\right) - D_R\left(\hat{\theta}^{(1)}_{ML}\right)\hat{\theta}_0\right\},$$

where we used Taylor's linearization of the nonlinear restrictions around $\theta = \hat{\theta}^{(1)}_{ML}$. Repeating this procedure, we get the following recursions for $\hat{\theta}^{(h)}_{ML}$ or, for short, $\hat{\theta}_h$

$$\hat{\theta}_h = \hat{\theta}_0 + K_h\left\{c_h - D_h\hat{\theta}_0\right\}$$
$$K_h = V_0 D_h'\left[D_h V_0 D_h'\right]^{-1} \qquad (h = 1, 2, ...).$$

For definitions of c_h and D_h, see Section 5; in practice, V_0 should be replaced by its estimate $\hat{V}_0$. By construction, for each h we have

$$\begin{aligned} 0 &= c\left(\hat{\theta}_{ML}^{(h-1)}\right) - D_R\left(\hat{\theta}_{ML}^{(h-1)}\right)\hat{\theta}_{ML}^{(h)} \\ &= c - R\left(\hat{\theta}_{ML}^{(h-1)}\right) + D_R\left(\hat{\theta}_{ML}^{(h-1)}\right)\hat{\theta}_{ML}^{(h-1)} - D_R\left(\hat{\theta}_{ML}^{(h-1)}\right)\hat{\theta}_{ML}^{(h)}; \end{aligned}$$

see (5.4). Hence, when $\hat{\theta}_{ML}^{(h)}$ converges to the (constrained) maximum likelihood solution $\hat{\theta}_{ML}$, $c - R\left(\hat{\theta}_{ML}^{(h-1)}\right)$ converges to zero. Also, assuming K_h converges to say $\hat{K}_{ML}$, the corresponding covariance matrix of $\hat{\theta}_{ML}$, say V_{ML}, can be approximated by

$$V_{ML} \approx \left\{I_k - KD_R(\theta)\right\}V_0,$$

which for sufficiently large h can be estimated by $\hat{V}_{ML} = \left(I_k - K_h D_h\right)\hat{V}_0$; see also Cramer (1986, page 38).

A.3 Regression estimator as GR estimator

Suppose that Y_i and the auxiliary variable Z_i, with known population mean $\overline{Z}$, are observed in s_2. In order to apply the GR estimator to this situation, define

$$\hat{\theta}_0 = \begin{pmatrix} \overline{y}_2 \\ \overline{z}_2 \end{pmatrix}, \quad V_0 = \operatorname{cov}\left(\hat{\theta}_0\right) = \left(\frac{1}{n_2} - \frac{1}{N}\right)\begin{pmatrix} S_y^2 & S_{yz} \\ S_{yz} & S_z^2 \end{pmatrix}.$$

The prior restriction is

$$0 = c - R\theta = \overline{Z} - (0,\ 1)\begin{pmatrix} \theta_y \\ \theta_z \end{pmatrix}.$$

Applying (5.1) and (5.2) to this case yields the following GR estimator

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right) = \begin{pmatrix} \overline{y}_2 \\ \overline{z}_2 \end{pmatrix} + K\left(\overline{Z} - \overline{z}_2\right)$$
$$K = V_0 R'\left(RV_0R'\right)^{-1} = \begin{pmatrix} S_{yz} \\ S_z^2 \end{pmatrix}\frac{1}{S_z^2} = \begin{pmatrix} b_{yz} \\ 1 \end{pmatrix} \qquad \left(b_{yz} = S_{yz}/S_z^2\right)$$
$$V_{GR} = \left(I_2 - KR\right)V_0 = \begin{pmatrix} 1 & -b_{yz} \\ 0 & 0 \end{pmatrix}V_0.$$

Hence, replacing b_{yz} by its estimate $b_{yz2} = s_{yz2}/s_{z2}^2$, we can approximate the first element in $\hat{\theta}_{GR}$ by $\hat{\theta}_{GRy} \approx \overline{y}_2 + b_{yz2}\left(\overline{Z} - \overline{z}_2\right)$ which corresponds to the familiar regression estimator, often denoted by $\hat{\overline{Y}}_{REG}$. For sufficiently large n_2, the variance of $\hat{\overline{Y}}_{REG}$ can be approximated by

$$\begin{aligned}\operatorname{var}\left(\hat{\bar{Y}}_{REG}\right) &\approx \operatorname{var}\left(\hat{\theta}_{GRy}\right) = \left[V_{GR}\right]_{11} = \left(\frac{1}{n_2}-\frac{1}{N}\right)\left(S_y^2 - b_{yz}S_{yz}\right) \\ &= \left(\frac{1}{n_2}-\frac{1}{N}\right)S_e^2; \qquad \text{(A.8)} \\ S_e^2 &= \frac{1}{N-1}\sum_{i\in U}\left\{Y_i - \bar{Y} - b_{yz}\left(Z_i - \bar{Z}\right)\right\}^2;\end{aligned}$$

recall from regression theory that $b_{yz}S_{yz} = b_{yz}^2 S_z^2$ and $S_y^2 = b_{yz}^2 S_z^2 + S_e^2$. The variance in (A.8) can be estimated by the well-known variance estimator

$$\hat{\operatorname{var}}\left(\hat{\bar{Y}}_{REG}\right) = \left(\frac{1}{n_2}-\frac{1}{N}\right)s_{\hat{e}2}^2, \quad \text{where} \quad s_{\hat{e}2}^2 = \frac{1}{n_2-1}\sum_{i\in s_2}\left\{Y_i - \bar{y}_2 - b_{yz2}\left(Z_i - \bar{z}_2\right)\right\}^2.$$

Similar results can be derived for more than one auxiliary variable. This illustrates once more that with respect to the bias and the variance approximation the AC estimator strongly resembles the regression estimator or, equivalently, the calibration estimator.

References

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

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

Cramer, J.S. (1986). *Econometric Applications of Maximum Likelihood Methods*. Cambridge: Cambridge University Press.

Hidiroglou, M.A., Särndal, C.E. and Binder, D.A. (1995). Weighting and estimation in business surveys. In *Business Survey Methods*, (Eds., B.G. Cox et al.). New York: John Wiley and Sons, Inc.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. Transactions ASME, *Journal of Basic Engineering*, 82, 35-45.

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

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.

Knottnerus, P. and Van Delden, A. (2012). On variances of changes estimated from rotating panels and dynamic strata. *Survey Methodology*, 38(1), 43-52.

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 496-500.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley and Sons, Inc.

Nordberg, L. (2000). On variance estimation for measures 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 changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34(2), 173-181.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Silva, P.L.D.N. and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1), 23-32.

Smith, P., Pont, M. and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994–2000. *Journal of the Royal Statistical Society D*, 52, 257-295.

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

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.