9th Workshop on Labour Force Survey Methodology
Rome, 15th and 16th of May 2014

REGRESSION COMPOSITE ESTIMATION FOR THE FINNISH LFS FROM A PRACTICAL PERSPECTIVE

Riku Salonen[1]

[1] Statistics Finland
e-mail: riku.salonen@stat.fi

Abstract

This paper examines the regression composite (RC) estimator in a complex rotating panel design from a practical perspective. Empirical results are based on real data from the Finnish Labour Force Survey (LFS). Currently, the Finnish LFS estimation system uses generalised regression (GREG) estimation and calibration techniques. Estimation for employment and unemployment are based on cross-sectional data. It is expected that estimation can be improved by using the rotating panel property, because employment and unemployment tend to be correlated over time. The RC estimator extends the standard GREG estimator by taking advantage of the temporal correlations. The RC estimator can be implemented within the current LFS estimation system by adding control totals and auxiliary variables to the estimation program. The empirical results shown that the RC estimator outperforms the standard GREG estimator for estimates of both level and change in employment and unemployment.

Keywords: Complex rotating panel; Regression composite estimator.

1 Introduction

The regression-based composite estimation method was developed in Canada. The first variant of regression composite estimation was called level-driven modified regression (MR1) estimator (Singh, Merkouris and Wu 1995, Singh 1996). MR1 produced reduced variance estimates of level relative to standard generalised regression (GREG) estimation based on cross-sectional data. The second variant of regression composite estimation was described by Singh, Kennedy, Wu and Brisebois (1997). This method was called change-driven modified regression (MR2) estimator. It produced reduced variance estimates of change compared to the standard GREG estimator.

The regression composite (RC) estimator was suggested by Fuller and Rao (2001). It is a compromise between MR1 and MR2. RC estimation has also been studied by Singh, Kennedy and Wu (2001), Gambino, Kennedy and Singh (2001), Bell (2001), Beaumont (2005) and Beaumont and Bocci (2005). Exploitation of sample overlap over time to improve the efficiency of estimates can be done via calibration by using the RC estimator. This method extends the GREG estimator by using information from the previous wave in a similar manner as the standard GREG estimator uses auxiliary variables. The RC method uses correlation between labour force characteristics of two consecutive waves. Level estimates based on cross-sectional data can be improved by using past data because of the correlation due to the common samples. The resulting estimates of change and average over time can also be improved. A further advantage of the new approach is that it yields a single set of estimation weights, leading to internal consistency of estimates. Since 2000, the RC estimator has been successfully used in the Canadian LFS. (See Gambino, Kennedy and Singh 2001.)

2 Design of the LFS

The target population of the Finnish LFS is persons aged 15 to 74, including foreign workers, citizens temporarily abroad, members of the armed forces, non-resident citizens, and unsettled and institutional population. The LFS is a monthly survey of individuals selected by systematic sampling. For estimation purposes, the sampling design is approximated with a without-replacement simple random sampling design (SRSWOR). The sampling frame is based on the database of the total population maintained by Statistics Finland. The sample size is approximately 12,500 individuals each month divided into five waves and four or five reference weeks. The monthly sample is allocated so that the weekly sample sizes are equal in each wave. The reference quarters and years are groups of 13 or 52 consecutive weeks.

The survey is repeated over time with partially overlapping samples. Each person will be included five times during 15 months. The rotation pattern in the LFS can be described as follows 1-2-1-2-1-5-1-2-1 (see Djerf 2004). In the first month, an individual is in the panel in wave one and after a two-month break, he/she will be included in the interview in the second wave, and so on. The lag between the interviews is three months except for one occasion, when it is six months.

The design of the LFS ensures the independence of the monthly samples in each three-month period, i.e. a sample for a quarter consists of separate monthly samples. Each sampled person is included once per quarter. This simplifies the estimation of quarterly figures. In the LFS the sample size is 37,500 persons per quarter. There is dependence between successive quarters; the overlap from one quarter to the next is 3/5. There is also a 2/5 overlap between two consecutive years. The disadvantage of this rotating panel structure is that the annual average (of four quarters) will be estimated with a larger variance compared to independent samples.

3 LFS estimation system

3.1 Current GREG estimation

The current GREG estimation system on monthly level was introduced in 2000. For this purpose i) the monthly weights need to be divided by three to create quarterly weights and ii) the monthly weights need to be divided by twelve to create annual weights. This automatically means that monthly, quarterly and annual estimates are consistent. Register data on unemployment were used as auxiliary information at the estimation stage. The use of such auxiliary data significantly improved estimates on unemployment by reducing sampling errors and non-response bias (Djerf 1997).

Denote the finite population by $U = \{1,\ldots, k,\ldots, N\}$. A sample $s \subset U$ of size n is drawn by a sampling design $p(s)$ with inclusion probabilities π_k, $k \in U$. Under SRSWOR, the inclusion probabilities are $\pi_k = n/N$. The design weight of unit k is $a_k = 1/\pi_k = N/n$. Denote by y the variable of interest and by y_k its value for unit k.

In the Finnish LFS, post-stratification is used to improve the precision of estimation. The H = 240 post-strata are constructed by sex (2 classes), age group (6 groups) and region (20 regions). Let n_h be the number of sampled units in post-stratum h, so $\sum_{h=1}^{H} n_h = n$. At the population level, $\sum_{h=1}^{H} N_h = N$.

There is also missingness due to unit non-response. The weight adjusted for non-response is $d_k = 1/(\pi_k \hat{\theta}_k) = (N_h / n_h) \times (n_h / m_h) = N_h / m_h$ for element k in post-stratum h, where m_h is the number of responding units in post-stratum h and $\hat{\theta}_k = m_h / n_h$ is the estimated response probability for element k in post-stratum h. The weights d_k adjusted for non-response are calibrated using the available auxiliary information. The GREG estimator with linear fixed-effects assisting model is a special case of the calibration estimator (e.g. Särndal, Swensson and Wretman 1992).

As Deville and Särndal (1992 and 1993) show, the GREG estimator of a population total $t_y = \sum_U y_k$ can be given as $\hat{t}_{ygr} = \sum_r w_k^{gr} y_k$ where r refers to the respondent group and the calibrated weights are $w_k^{gr} = d_k g_k^{gr}$ with

$$g_k^{gr} = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})' \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k ' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \mathbf{x}_k q_k \,. \qquad (1)$$

The known auxiliary totals, called control totals, are $\mathbf{t_x} = \left(t_{x1}, \ldots, t_{xj}, \ldots\ t_{xJ}\right)'$ and $\hat{\mathbf{t}}_\mathbf{x} = \left(\hat{t}_{x1}, \ldots, \hat{t}_{xj}, \ldots\ \hat{t}_{xJ}\right)'$ is a vector of estimates of the elements in $\mathbf{t_x}$. The auxiliary information vector is defined as $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots x_{Jk})'$ and q_k is a known constant (usually set equal to one). The calibration property assures that $\hat{\mathbf{t}}_\mathbf{x} = \sum_r w_k^{gr} \mathbf{x}_k = \sum_U \mathbf{x}_k = \mathbf{t_x}$. In the Finnish LFS the auxiliary information vector is defined by four auxiliary variables taken from administrative registers: x_1 = sex (2 classes), x_2 = age (12 groups), x_3 = region (20 regions), x_4 = employment status in Ministry of Labour's job-seeker register (8 classes). Weekly balancing of weights on monthly level was also included in the calibration (4 or 5 reference weeks).

We used a linear distance function in the calibration procedure, available in CLAN (until 2013) and ETOS (since 2014), programs developed by Statistics Sweden for calibration and GREG estimation. Variance estimation is based on GREG estimation. For variance estimation we need the residuals $e_k = y_k - \mathbf{x}_k' \hat{B}$, where

$$\hat{B} = \left(\sum_r \frac{x_k x_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \sum_r \frac{x_k y_k q_k}{\pi_k \hat{\theta}_k} \,.$$

The variance estimator of $\hat{t}_{ygr}$ under SRSWOR is given by

$$\hat{V}\left(\hat{t}_{ygr}\right) = \sum_{h=1}^{H} \frac{N_h^2}{m_h} \left(1 - \frac{m_h}{N_h}\right) \frac{1}{m_h - 1} \left[\sum_{r_h} \left(g_k^{gr} \times e_k\right)^2 - \frac{\left(\sum_{r_h} g_k^{gr} \times e_k\right)^2}{m_h} \right], \qquad (2)$$

where r_h denotes the respondent set in post-stratum h.

3.2 RC estimation

The GREG estimator is based on the current period's data and does not use the rotation panel pattern. In RC estimation auxiliary data, known as composite auxiliary variables $\mathbf{z}_k$, are taken from the previous time period t−1. The composite auxiliary variables have random benchmarks determined by setting the weighted sum of variables $\mathbf{z}_k$ equal to the previous period's estimates. These estimated control totals are called composite control totals. Under a rotating panel design, however, values for the composite auxiliary variables are known for the overlapping part of the sample. For the non-overlapping part the values are imputed.

There is dependence between successive quarters; the overlap from a quarter to the next is 3/5. The part of the sample which is common for the current and previous the previous wave of interview is referred to as the *matched*, i.e. overlap, sample. The remaining 2/5 part of the sample is known as the *unmatched*, i.e. non-overlap, sample.

In the RC estimation system for the LFS we used the following composite auxiliary data: labour force status (employed, unemployed) by sex/age group, labour force status by NUTS2 region and labour force status by industry. The corresponding composite auxiliary variables were defined as a linear combination of *MR*1 and *MR*2 as suggested by Fuller and Rao (2001). For the level-driven predictor *MR*1, data from the previous wave of interview were used for the matched sample, and mean imputation was used for the unmatched part. For the change-driven predictor *MR*2, carry backward imputation was used for the unmatched sample, and transformed values of the previous wave of interview were used for the matched sample.

The composite auxiliary variables $\mathbf{z}_k$ were formulated as $\mathbf{z}_k = (1-\alpha)MR1_k + \alpha\, MR2_k$, $\alpha \in [0,1]$, where the choice of the coefficient α depends on the variable of interest and on the relative importance of level versus change (Gambino, Kennedy and Singh 2001). The level-driven and change-driven predictors are special cases corresponding to $\alpha = 0$ and $\alpha = 1$, respectively. The level-driven predictor is given by

$$MR1_k = \begin{cases} \hat{t}_{t-1} / N_{t-1} & \text{if element } k \text{ belongs to the unmatched part of sample} \\ y_{t-1,k} & \text{if element } k \text{ belongs to the matched part of sample} \end{cases}$$

with $\hat{t}_{t-1}$ as an imputed value defined as the previous wave of interview estimate of the total of the study variable and $N_{t\text{-}1}$ is the corresponding population size, and $y_{t\text{-}1,k}$ refers to the observed value of the study variable for unit k at time point t−1. The change-driven predictor is given by

$$MR2_k = \begin{cases} y_{tk} & \text{if element } k \text{ belongs to the unmatched part of sample} \\ y_{tk} + R^{-1}(y_{t-1,k} - y_{tk}) & \text{if element } k \text{ belongs to the matched part of sample} \end{cases}$$

where R is a ratio that adjusts the sample overlap from one quarter to the next (R = 3/5), and y_{tk} refers to the observed value of the study variable for unit k at time point t.

As Singh, Kennedy and Wu (2001) show, the RC estimator of $t_y = \sum_U y_k$ can be expressed in the form of $\hat{t}_{yrc} = \sum_r w_k^{rc} y_k$ where the calibrated RC weights w_k^{rc} are obtained in a similar manner as the GREG weights w_k^{gr}, except that the constraint $\sum_r w_k^{gr}\mathbf{x}_k = \mathbf{t}_\mathbf{x}$ is replaced by the constraints $\sum_r w_k^{rc}\mathbf{x}_k = \mathbf{t}_\mathbf{x}$ and $\sum_r w_k^{rc}\mathbf{z}_k = \hat{\mathbf{t}}_\mathbf{z}$.

The vector of estimated composite control total $\hat{\mathbf{t}}_{\mathbf{z}}$ must be computed using from the previous wave of interview data. The RC weights w_k^{rc} are calibrated on the usual control totals $\mathbf{t}_{\mathbf{x}}$ given by $w_k^{rc} = d_k g_k^{rc}$, where g_k^{rc} has the same form as (1), with the exception that $\mathbf{x}_k$ and $\mathbf{t}_{\mathbf{x}}$ are replaced by $\left(\mathbf{x}_k^{'}, \mathbf{z}_k^{'}\right)^{'}$ and $\left(\mathbf{t}_{\mathbf{x}}^{'}, \hat{\mathbf{t}}_{\mathbf{z}}^{'}\right)^{'}$, respectively.

The approximate variance of $\hat{t}_{yrc}$ is calculated by using g_k^{rc} instead of g_k^{gr} in the GREG variance formula (2). Thus the variance of $\hat{t}_{yrc}$ is estimated by

$$\hat{V}\left(\hat{t}_{yrc}\right) = \sum_{h=1}^{H} \frac{N_h^2}{m_h}\left(1 - \frac{m_h}{N_h}\right)\frac{1}{m_h - 1}\left[\sum\nolimits_{r_h}\left(g_k^{rc} \times e_k\right)^2 - \frac{\left(\sum\nolimits_{r_h} g_k^{rc} \times e_k\right)^2}{m_h}\right]. \qquad (3)$$

Here we have used the CLAN and ETOS programs for point and variance estimation.

4 Some practical viewpoints

We present a summary of some of the features and properties of the RC estimator from a practical perspective.

System implementation. The RC estimator can be implemented within the current LFS estimation system by adding control totals and auxiliary variables to the estimation program. It can be performed by using, with minor modification, standard software for GREG estimation, such as ETOS. It yields a single set of estimation weights. Leading to internal consistency of estimates (e.g. Employment + Unemployment = Labour Force).

Empirical results. We have compared the RC estimator to the GREG estimator in the Finnish LFS real data. Here we have used the ETOS program for point and variance estimation (Taylor linearisation method). Sampling variance is a measure of the design's efficiency. For the variables that were included as composite control totals, there are substantial gains in efficiency for both estimates of level and of change. In particular, this holds for employment by Standard Industrial Classification. A reason for large efficiency improvement is the high correlation of employment over time. For unemployment estimates, the efficiency gains were modest. An explanation for this is that unemployment is only moderately correlated over time and the register data on unemployment (labour force status in Ministry of Labour's job-seeker register) are used as auxiliary information already at the GREG estimation stage. For variables that were not controlled, there were little or no efficiency gains from RC estimation, unless the variable in question was highly correlated with a composite auxiliary variable. The results are well comparable with results reported from other countries.

The quality of the estimates. The RC estimator produced level and change estimates that were usually more efficient than the estimates produced by the current GREG estimator. When designing the sample, we try to reduce the sampling variance. From another viewpoint, a more efficient sample design, or one that results in a smaller sampling variance, helps to control the impacts of growing nonresponse compared to another less efficient design, while maintaining the quality of the estimates (see Statistics Canada 2008).

Some problems with RC method? Dever and Valliant (2010) consider the problems of estimated control totals. They compared several estimated-control (EC) variance estimators and they supposed that traditional variance estimators can underestimate the population sampling variance resulting.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach, *Journal of the Royal Statistical Society*. Series B, 67, 445-458.

Beaumont, J.-F. and Bocci, C. (2005). A Refinement of the Regression Composite Estimator in the Labour Force Survey for Change Estimates. SSC Annual Meeting, *Proceedings of the Survey Methods Section*, June 2005.

Bell, P. (2001). Comparison of Alternative Labour Force Survey Estimators. Survey Methodology, 27, 53—63.

Chen, E.J. and Liu, T.P. (2002). Choices of Alpha Value in Regression Composite Estimation for the Canadian Labour Force Survey: Impacts and Evaluation. *Methodology Branch Working Paper*, HSMD-2002-005E, Statistics Canada.

Dever, A.D., and Valliant, R. (2010). A Comparison of Variance Estimators for Poststratification to Estimated Control Totals. Survey Methodology, 36, 45-56

Deville J.-C. and Särndal C.E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376—382.

Deville J.-C., Särndal C.E. and Sautory O. (1993): Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013—1020.

Djerf, K. (1997): Effects of Post-Stratification on the Estimates of the Finnish LFS. *Journal of Official Statistics*, 13, 29—39.

Djerf, K. (2004): Non-response in Time: A Time Series Analysis of the Finnish Labour Force Survey. *Journal of Official Statistics*, 20, 39—54.

Fuller, W.A., and Rao, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

Gambino, J., Kennedy, B., and Singh, M.P. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation ja Implementation. *Survey Methodology*, 27, 65-74.

Salonen, R. (2007). Regression Composite Estimation with Application to the Finnish Labour Force Survey. *Statistics in Transition*, 8, 503-517.

Singh, A.C., Merkouris, P. and Wu, S. (1995). Composite Estimation by modified regression for repeated survey. *ASA Proc., Surv. Res. Meth*. Sec., 420—425.

Singh, A.C. (1996). Combining information in survey sampling by modified regression. *ASA Proc., Surv. Res. Meth*. Sec., Vol. 1, 120—129.

Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F. (1997). Composite Estimation for the Canadian Labour Force Survey . *ASA Proc., Surv. Res. Meth*. Sec., 300—305.

Singh, A.C., Kennedy, B., and Wu, S. (2001). Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design. *Survey Methodology*, 27, 33-44.

Statistics Canada (2008). *Methodology of the Canadian Labour Force Survey*, Cat 71-526-X.

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