Estimation in the Swedish LFS – an example of combining survey data from independent samples

Martin Axelson[1] and Frida Videll[2]
[1]Statistics Sweden, Research & Development, SE701 89 Örebro, Sweden
[2] Statistics Sweden, Process Department, Box 24300, SE104 51 Stockholm, Sweden

1. Introduction and outline

In 2009 the Swedish parliament decided to increase the budget for the LFS. In this paper, we provide some information on the background to the decision and the impact it eventually had on how the Swedish LFS is performed. In particular, the paper presents the background to and the basic features of the new regression estimator that is today implemented in the LFS. The new estimator, which is easy to justify theoretically, results in a single weight system to be used for all study variables and is, in general, more efficient than the method used initially. In addition, the new estimator was very easy to implement using the available IT-infrastructure.

In section 2, some more information on the sampling design currently in use is provided. Section 3 provides some theoretical results regarding different ways of combining survey data from independent samples from the same population. In section 4, we discuss how the results in section 3 relate to the estimator today implemented in the LFS and exemplify numerically the precision gains realized by the new estimator, before the paper is concluded with some final remarks.

2. The current sampling design for the Swedish LFS

In 2008, Statistics Sweden conducted a project in response to increased political and public interest in groups outside or with a weak attachment to the labour market. The goal of the project was to suggest a cost-efficient way to secure better statistics for both stocks and flows for certain small domains of particular interest. The project did not simply suggest an increase in sample size for the sampling design already in use in the LFS, nor did it suggest that the sampling design in use should be replaced by a completely new design. Instead, it suggested an approach that would combine data from two different samples drawn from the same sampling frame. In addition to a sample drawn according to the ordinary sampling design already in use, the project suggested a second sample to be drawn according to an additional sampling design constructed with high precision for specific parameters and domains of study in mind.

The LFS suggested by the project would thus comprise of two monthly samples, drawn as stratified samples but based on very different stratifications. Whereas the existing design in effect results in a self-weighting sample, the suggested additional sampling design would use a stratification and sample size allocation aimed at resulting in a clear over-representation of respondents either unemployed or not in the labour force. Along with fairly detailed discussions on practical issues regarding the construction and implementation of the suggested "new" sampling design, the project report also presented results on possible precision gains for a number of parameters.

In 2009 the Swedish government decided to increase the budget for the LFS. The decision, in part based on the results from the project in 2008, allowed for an increase of the total monthly sample size by almost 40 %. Following the political decision, Statistics Sweden initiated a new project to finalize and implement the suggestions from the previous project. The project delivered detailed instructions for the construction and

implementation of an additional sampling design, p_A, to be used in combination with p_O, the ordinary LFS-design already in use. The new approach was implemented from 2010. Thus, from January 2010, the monthly LFS is based on two samples, one drawn according to the ordinary design p_O, with sample size $n_O \approx 21500$, i.e. the sample size used prior to the budget increase, and one drawn according to the additional design p_A, with sample size $n_A \approx 8000$.

3. Combining data from independent samples – some theory

Below some results presented in Statistics Sweden (2014) are reiterated. The results on explicit weighting presented in section 3.1 follow from well-known results from sampling theory and statistical theory. However, the results in section 3.2 on implicit weighting, which may be seen as an extension of some of the results presented by Sing and Mecatti (2011), are to our knowledge novel.

Let $U = \{1,...,k,...,N\}$ denote the population of interest, let y_k denote the (fixed) value of the study variable y associated with element k, and let $t = \sum_{k \in U} y_k$ denote the parameter of interest. Appropriate definition of the variable y allows the parameter t to be defined at the domain level. Suppose that in order to estimate t, not one but J different samples are drawn from the population of interest. More specifically, let s_j denote the j:th sample, drawn from according to the sampling design p_j ,and let $\pi_{j,k}$ and $\pi_{j,kl}$ denote the first- and second-order inclusion probabilities under the design p_j, $j = 1,...,J$. Below it is assumed that p_j is such that $\pi_{j,k} > 0$ for all $k \in U$ and $\pi_{j,kl} > 0$ for all pairs $\{k,l\} \in U$, , $j = 1,...,J$. Moreover, it is assumed that the designs are such that the samples s_j, $j = 1,...,J$, are selected independently of each other.

3.1 Explicit weighting – weighting at the estimator level

Let a_j, $j = 1,...,J$, be constants such that $\sum_{j=1}^{J} a_j = 1$, and let $\mathbf{x}$ denote a vector valued variable for which the total $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ is known. If y_k och $\mathbf{x}_k$ are observed for $k \in s_j$, a GREG-estimator for for t is given by

$$\hat{t}_j = \sum_{k \in s_j} \frac{g_{j,k} y_k}{\pi_{j,k}} = \sum_{k \in s_j} \frac{y_k}{\pi_{j,k}} + (\mathbf{t}_x - \sum_{k \in s_j} \frac{\mathbf{x}_k}{\pi_{j,k}})' \hat{\mathbf{B}}_j \approx \sum_{k \in s_j} \frac{y_k}{\pi_{j,k}} + (\mathbf{t}_x - \sum_{k \in s_j} \frac{\mathbf{x}_k}{\pi_{j,k}})' \mathbf{B}$$

where $g_{j,k} = 1 + (\mathbf{t}_x - \sum_{k \in s_j} \mathbf{x}_k / \pi_{j,k})' (\sum_{k \in s_j} \mathbf{x}_k \mathbf{x}_k' / \pi_{j,k})^{-1} \mathbf{x}_k$, $\hat{\mathbf{B}}_j = (\sum_{k \in s_j} \mathbf{x}_k \mathbf{x}_k' / \pi_{j,k})^{-1} \sum_{k \in s_j} \mathbf{x}_k y_k / \pi_{j,k}$ and $\mathbf{B} = (\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{k \in U} \mathbf{x}_k y_k$. Weighting the estimator $\hat{t}_j$ with a_j, $j = 1,...,J$, and summing, we get the explicitly weighted GREG-estimator

$$\hat{t}_{Exp} = \sum_{j=1}^{J} a_j \hat{t}_j \qquad (1)$$

Given that the sample sizes are large enough for $E_{p_j}(\hat{t}_j) \approx t$ to hold for $j = 1,...,J$, it follows that $E(\hat{t}_{Exp}) \approx t$, i.e. $\hat{t}_{Exp}$ is approximately unbiased for t. An approximate variance expression and a variance estimator are given by

$$V(\hat{t}_{Exp}) = V(\sum_{j=1}^{J} a_j \hat{t}_j) = \sum_{j=1}^{J} a_j^2 V_{p_j}(\hat{t}_j) \approx \sum_{j=1}^{J} a_j^2 \sum_{k \in U} \sum_{l \in U} (\pi_{j,kl} - \pi_{j,k}\pi_{j,l}) \frac{y_k - \mathbf{x}_k'\mathbf{B}}{\pi_{j,k}} \frac{y_l - \mathbf{x}_l'\mathbf{B}}{\pi_{j,l}}$$

and

$$\hat{V}(\hat{t}_{Exp}) = \sum_{j=1}^{J} a_j^2 \sum_{k \in s_j} \sum_{l \in s_j} (\pi_{j,kl} - \pi_{j,k}\pi_{j,l}) \frac{g_{j,k}(y_k - \mathbf{x}_k'\hat{\mathbf{B}}_j)}{\pi_{j,k}} \frac{g_{j,l}(y_l - \mathbf{x}_l'\hat{\mathbf{B}}_j)}{\pi_{j,l}}$$

Optimal explicit weights are given by $a_j = V_{p_j}(\hat{t}_{G,j})^{-1} / \sum_{j'=1}^{J} V_{p_{j'}}(\hat{t}_{G,j'})^{-1}$, $j = 1,...,J$.

3.2 Implicit weighting – weighting at the element level

Let $b_{j,k}$, $j = 1,...,J$, be non-stochastic variables such that $\sum_{j=1}^{J} b_{j,k} = 1$ for every $k \in U$ and let

$$h_k = 1 + (\mathbf{t}_x - \sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k}\mathbf{x}_k}{\pi_{j,k}})' (\sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k}\mathbf{x}_k\mathbf{x}_k'}{\pi_{j,k}})^{-1} \mathbf{x}_k$$

An implicitly weighted GREG-estimator for t is given by

$$\begin{aligned} \hat{t}_{Imp} &= \sum_{j=1}^{J} \sum_{k \in s_j} \frac{h_k b_{j,k} y_k}{\pi_{j,k}} = \sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k} y_k}{\pi_{j,k}} + (\mathbf{t}_x - \sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k}\mathbf{x}_k}{\pi_{j,k}})' \hat{\mathbf{B}} \\ &\approx \sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k} y_k}{\pi_{j,k}} + (\mathbf{t}_x - \sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k}\mathbf{x}_k}{\pi_{j,k}})' \mathbf{B} \end{aligned} \qquad (2)$$

where $\hat{\mathbf{B}} = (\sum_{j=1}^{J} \sum_{k \in s_j} b_{j,k}\mathbf{x}_k\mathbf{x}_k' / \pi_{j,k})^{-1} \sum_{j=1}^{J} \sum_{k \in s_j} b_{j,k}\mathbf{x}_k y_k / \pi_{j,k}$ and $\mathbf{B}$ as previously defined. Since

$$E(\sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k} y_k}{\pi_{j,k}}) = \sum_{j=1}^{J} \sum_{k \in U} b_{j,k} y_k = \sum_{k \in U} y_k \sum_{j=1}^{J} b_{j,k} = \sum_{k \in U} y_k = t$$

and thus $E(\sum_{j=1}^{J} \sum_{k \in s_j} b_{j,k}\mathbf{x}_k / \pi_{j,k}) = \mathbf{t}_x$ by analogy, it follows that $E(\hat{t}_{Imp}) \approx t$, i.e., $\hat{t}_{Imp}$ is approximately unbiased for t. An approximate variance expression and a variance estimator are given by

$$V(\hat{t}_{Imp}) \approx V(\sum_{j=1}^{J} \sum_{k \in s_j} \frac{b_{j,k}(y_k - \mathbf{x}_k'\mathbf{B})}{\pi_{j,k}}) = \sum_{j=1}^{J} \sum_{k \in U} \sum_{l \in U} (\pi_{j,kl} - \pi_{j,k}\pi_{j,l}) \frac{b_{j,k}(y_k - \mathbf{x}_k'\mathbf{B})}{\pi_{j,k}} \frac{b_{j,l}(y_l - \mathbf{x}_l'\mathbf{B})}{\pi_{j,l}}$$

and

$$\hat{V}(\hat{t}_{Imp}) = \sum_{j=1}^{J}\sum_{k\in s_j}\sum_{l\in s_j}(\pi_{j,kl} - \pi_{j,k}\pi_{j,l})\frac{h_k b_{j,k}(y_k - \mathbf{x}_k'\hat{\mathbf{B}})}{\pi_{j,k}}\frac{h_l b_{j,l}(y_l - \mathbf{x}_l'\hat{\mathbf{B}})}{\pi_{j,l}}$$

Suppose the J sampling designs are such that the approximation

$$V(\hat{t}_{Imp}) \approx \sum_{j=1}^{J}\sum_{k\in U}(\pi_{j,k} - \pi_{j,k}^2)\frac{b_{j,k}^2(y_k - \mathbf{x}_k'\mathbf{B})^2}{\pi_{j,k}^2} = \sum_{k\in U}(y_k - \mathbf{x}_k'\mathbf{B})^2\sum_{j=1}^{J}\frac{b_{j,k}^2(1-\pi_{j,k})}{\pi_{j,k}} \quad (3)$$

is valid. The right hand side of (3) is minimized by

$$b_{j,k} = \frac{\pi_{j,k}/(1-\pi_{j,k})}{\sum_{j'=1}^{J}\pi_{j',k}/(1-\pi_{j',k})},\ j = 1,...,J \text{ and } k \in U \quad (4a)$$

If $\pi_{j,k}/(1-\pi_{j,k}) \approx \pi_{j,k}$, $j = 1,...,J$ and $k \in U$, (4a) implies that

$$b_{j,k} \approx b_{j,k}^{*} = \pi_{j,k}/\sum_{j'=1}^{J}\pi_{j',k},\ j = 1,...,J \text{ and } k \in U \quad (4b)$$

4. Estimation in the Swedish LFS

It can be argued that the possible gains in precision presented by the project in 2008 were unrealistic. The reason for this is quite simple; for each studied parameter, the numerical result on expected gain in precision was based on (almost) optimal explicit weighting of GREG-estimators based on the samples selected according to the ordinary LFS-design and the suggested additional sampling design. Clearly, such weighting amounts to using parameter specific weight systems and is thus of limited interest to NSIs, who typically try very hard to avoid multiple weight systems. Consequently, the question of estimation was further addressed in the implementation project carried out in 2009. The project proposed an estimator based on a common weight system for all parameters of interest. The estimator can be seen as an extended version of (1), allowing the estimators to be combined to be statistically dependent. In principle, the explicit weights were chosen to reflect the relation between the sample sizes under p_O and p_A for certain pre-specified subgroups. However, once the estimator was implemented and estimates for 2010 were actually produced, the achieved precision gains turned out to be smaller, or even much smaller, than anticipated for many of the parameters of interest. For some parameters, the weight system even resulted in a precision loss when compared to estimates solely based on the ordinary sample.

Against this background an overview of the estimation was initiated in 2011, aimed at finding a method for constructing a single weight system with more appealing precision properties. Initially, a GREG-type estimator, based on what O'Muircheartaigh and Pedlow (2002) denote cumulative inclusion probabilities, was considered. However, once the results presented in section 3.2 had been derived, an estimator based on implicit weighting was chosen. The estimator, below denoted $\hat{t}_{LFS}$, is best described as a non-response adjusted version of (2), with weights inspired by (4b). More information on the result of the overview is found in Statistics Sweden (2014). When implemented, $\hat{t}_{LFS}$ was used to produce revised statistics from 2010.

Let $\hat{t}_{O,LFS}$ denote the "old" LFS-estimator that would be used if only the sample drawn according to p_O was available .Table 1 and 2 illustrate the estimated gain in precision that follows from using $\hat{t}_{LFS}$ instead of $\hat{t}_{O,LFS}$ for estimation of domain parameters of the type for which the new estimation approach should be efficient, i.e. parameters defined for groups outside or with a weak attachment to the labour market.

Table 1: Population 20-64 by age and labor status, January 2013

Age	$[\hat{V}(\hat{t}_{LFS})/\hat{V}(\hat{t}_{O,LFS})]^{0.5}$	
	Unemployed	*Not in the labor force*
20-24	0.86	0.85
25-34	0.83	0.82
35-44	0.79	0.81
45-54	0.82	0.77
55-64	0.76	0.88

Table 2: Population 20-64 not in the labor force by gender, January 2013

Sex	$[\hat{V}(\hat{t}_{LFS})/\hat{V}(\hat{t}_{O,LFS})]^{0.5}$					
	Full-time students	*Working at home*	*Jobseekers, not available*	*Retired*	*Long-term ill*	*Others*
Male	0.83	0.72	0.79	0.93	0.80	0.86
Female	0.84	0.85	0.78	0.91	0.82	0.90

The precision gains presented should be compared to $1.4^{-0.5} \approx 0.85$, the expected precision gain, had the budget increase in 2009 been used for a of 40 % increase of the sample size at the stratum level under p_O. For the 22 parameters in table 1 and 2, the ratio $[\hat{V}(\hat{t}_{LFS})/\hat{V}(\hat{t}_{O,LFS})]^{0.5}$ is less than or equal to 0.85 in 19 cases.

Thus, the tables reflect the main conclusion of in Statistics Sweden (2014): using implicit weighting to combine data from the samples drawn according to p_O and p_A contributes to the realization of the main goal behind the budget increase – to secure better statistics for groups outside or with a weak attachment to the labour market. The resulting estimator, $\hat{t}_{LFS}$, uses a single weight system for all study variables and is easy to justify theoretically. In addition, the construction of $\hat{t}_{LFS}$ is such that it was easy to implement, using the already existing IT-infrastructure for the LFS.

References

O'Muircheartaigh, C. and Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, pp. 2557-2562.

Sing, A.C. and Mecatti, F. (2011). Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics* 27, 4, 633-650.

Statistics Sweden (2014). Method for estimation when combining samples with different designs in the Swedish Labour Force Surveys. Background Facts, Labour and Education Statistics 2014:1.

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