Weighting and Estimation of Sampling Errors in the UK Annual Population Survey

Mārtiņš Liberts[1]

[1] Office for National Statistics, UK
e-mail: martins.liberts@ons.gov.uk

Abstract

The Annual Population Survey is a survey of households in the United Kingdom. The sampling design of the survey can be approximated by a stratified simple random sample for households and a stratified cluster sample for individuals where the households are clusters of individuals. The topics discussed in the paper are two-phase calibration, calibration with many calibration constraints and estimation of standard errors using a jackknife linearization variance estimator.

1 Annual Population Survey

The Annual Population Survey (APS) is a survey of households in the United Kingdom (UK). Its purpose is to provide information on key social and socioeconomic variables between the 10-yearly censuses, with particular emphasis on providing information relating to UK unitary authorities and local authority districts (UALAD). The first publication of APS data covered the survey period January to December 2004. Subsequently, APS data has been published on a quarterly basis, but with each publication covering a year's data.

1.1 Sample Design

The APS sample consists of the Labour Force Survey (LFS) sample, the Local Labour Force Survey (LLFS) sample and an additional sample (called APS boost sample) during the time period 2004-2005. The LFS sample covers the whole UK. It is a simple random sample of addresses. The LLFS covers Great Britain (England, Wales and Scotland, excluding Northern Ireland). APS boost sample covers only England. The LLFS sample and the APS boost sample use stratified random sampling of addresses. The APS boost sample is designed to ensure that the APS achieves a sample of at least 500 economically active persons in each English Local Authority District. The APS boost sample is sampled from English Local Authority Districts where the combination of LFS sample and LLFS samples is too small. We can assume that the APS sample is a stratified random sample of addresses.

All households at the selected address are sampled and all individuals belonging to the selected households are sampled.

There are two types of estimates provided from the APS – those from individual records and those from household records. This is stratified cluster sampling of households or individuals where addresses form clusters. The postcode address file is used as the sampling frame.

2 Weighting

2.1 Design Weights

Design weights are derived from the sample and sampling frame. Design weights are computed for each address as the inverse of the inclusion probability. The same design weight is assigned to each household belonging to a sampled address. The same design weight is also assigned to each individual belonging to a sampled household.

Design weights are scaled to known population totals to derive the initial weights for calibration. The scaling is done with one coefficient for broad regional subgroups. Here the regions are defined as Government Office Regions (GOR). There are 10 GOR in England; Wales, Scotland and Northern Ireland each form a separate GOR; thus there are 13 GOR in total. The coefficient is derived as the population total divided by the sum of the design weights over all responding units.

2.2 Weighting for Household File

There are 463 calibration variables created for weighting the APS household file. There are 32 variables for sex and age groups and 431 variables for regional breakdown. The variables contain the number of household members belonging to each of these groups. The totals are delivered from an auxiliary source of population statistics.

The calibration is done using the initial weights described in Section 2.1 and c_i weights which define the variance in the GREG model (Lundström & Särndal 2005, p. 34), where c_i is equal to the inverse of the number of household members in each household. Households with a higher number of members are less affected by calibration. The usage of c_i weights results in less variable calibration factors or *g*-weights.

2.3 Weighting for Individual File

The calibration of the APS individual file is done in two phases. The APS individual sample can be thought of as conforming to a two phase sampling design. The whole APS sample (the LFS, LLFS and APS boost samples) shall be referred to as the core sample. There are core questions asked of the core sample. The core questions consist of demographic questions and

key LFS questions for example employment status. Extra questions are asked to a sub-sample (the LFS and LLFS samples) of the core sample.

Calibration of the weights for the core sample is done in the first phase. Altogether there are 3,988 calibration variables used in the first phase. The huge number of calibration variables comes from the detailed breakdown by local authority areas required by APS. The calibration variables can be split into three sets. The first set consists of variables describing sex, age groups and geographic areas. The geographic areas used are artificially created by grouping UALAD in this case. The grouping is done only for weighting purposes. The second set of auxiliary variables consists of variables describing the working age population (16-64 for males and 16-59 for females) and UALAD areas. The third set of calibration variables contains variables describing sex and age groups. The age groups are defined so as to improve estimates for the young working age population in this set of variables. The variables can be considered as dummy variables defining which of the groups the respondent belongs to. Each respondent belongs to one group from each set of auxiliary variables. The totals for the calibration variables are delivered from an auxiliary source. Note that the creation of calibration variables and totals is not described in full detail in this paper.

The calibration is carried out using the initial weights described in Section 2.1. Unlike in the household weighting the c_i weights are not used for individual weighting and as such the c_i weights are set equal to 1 for all units.

The second phase of weighting is performed only for the population of England. There are 1,664 calibration variables used for the second phase calibration. The calibration variables can be split into two sets of variables. The first set contains some of the variables used in the first phase of weighting i.e. those describing the working age population and UALAD areas. The second set of calibration variables contains variables describing geographic areas and the status of the economic activity of the population. The totals of the second phase calibration are estimated from the core sample.

The calibration is done using the calibrated weights from the first phase of weighting. The c_i weights are not used in this case either.

3 Standard Error Estimation

A procedure capable of estimating standard errors is required for the production of official statistics. The procedure should be flexible to different sample designs used in the institution.

The estimation should be done in an efficient way – providing balance between precision and estimation time required.

3.1 GREG Estimator

The generalised regression (GREG) estimator of totals can be described as

$$\hat{\Theta}^T = 1^T \Omega Y + \left(c^T - 1^T \Omega X\right)\left(X^T \Delta \Omega X\right)^{-1} X^T \Delta \Omega Y\,, \qquad (1)$$

where $\hat{\Theta}$ is the $(l \times 1)$ vector of GREG estimates of totals;

1 is a $(n \times 1)$ vector with all elements equal to 1;

Ω is a square $(n \times n)$ diagonal matrix with design or initial weights on the main diagonal;

Y is a $(n \times l)$ matrix where the columns are study variables;

c is a $(m \times 1)$ vector of auxiliary totals;

X is a $(n \times m)$ matrix where the columns are the auxiliary variables;

Δ is a square $(n \times n)$ diagonal matrix with user specified weights (defining the variance in the GREG model) on the main diagonal, often Δ is equal to the identity matrix, otherwise elements on the diagonal, δ_{ii}, normally take on values in the range $0 \le \delta_{ii} \le 1$;

n is the number of respondents;

m is the number of auxiliary variables;

l is the number of study variables.

We can rewrite (1) as

$$\hat{\Theta}^T = \left(1^T + \left(c^T - 1^T \Omega X\right)\left(X^T \Delta \Omega X\right)^{-1} X^T \Delta\right) \Omega Y\,. \qquad (2)$$

From (2) calibration factors or *g*-weights can be derived

$$g^T = 1^T + \left(c^T - 1^T \Omega X\right)\left(X^T \Delta \Omega X\right)^{-1} X^T \Delta\,, \qquad (3)$$

where g is a $(n \times 1)$ vector of g-weights.

To estimate either the variance or the standard error (SE) of $\hat{\Theta}$ it is first necessary to estimate the regression residuals arising from the regression of Y on X. The $(n \times l)$ matrix of estimated residuals can be defined as $\hat{E} = Y - X\hat{B}$, where

$$\hat{B} = \left(X^T \Delta \Omega X \right)^{-1} X^T \Delta \Omega Y \qquad (4)$$

is a $(n \times l)$ matrix containing the estimates of the regression coefficients.

The variance of $\hat{\Theta}$ can be estimated as function of $\hat{E}$, where the function depends on sampling design used.

M is a $(m \times m)$ matrix defined as

$$M = X^T \Delta \Omega X .$$

It can be seen from (3) and (4) that the matrix M^{-1} is required both for calculating the *g*-weights and also for estimating the variance of the GREG estimates. The inversion of the matrix M is one of the most computer intensive tasks in GREG estimation.

3.2 Practical Implementation of GREG Estimator

If the parameters n, m and l used in the weighting process are large numbers, it can cause practical problems in the implementation of the weighting. The problems can arise from the long computation time required and the size of the matrices used in weighting. All the parameters n, m and l are quite large numbers in APS. n, the number of respondents can be approximately 150,000 for APS household files and 300,000 for the APS individual file. m, the number of auxiliary variables can be approximately 500 for APS household files and 4,000 for APS individual file. It is obvious that l, the number of study variables can be very large.

To gain the efficiency of the SE estimation procedure the following scheme is proposed in Table 1.

Stage	Input	Output
I Weighting	X, Ω, Δ, c	g, M^{-1}
II Estimation	X, Ω, Δ, Y, g, M^{-1}	$\hat{\Theta}$, $\hat{E}$

Table 1, Scheme for SE Estimation Procedure

The idea is to compute the matrix M^{-1} once during the weighting procedure and then save it so that it can also be used in the estimation of standard errors. It should theoretically increase the speed of the procedure when compared to the procedure where M^{-1} is computed each time a SE is estimated.

3.3 Jackknife Linearization

Estimates of SEs are required for level (total), ratio (of two levels) and change of level and ratio estimators. The Jackknife linearization method has been proposed as suitable method to deal with this task.

The Jackknife linearization method is an approximation method but it is very efficient with respect to the computation time required when compared to the classical Jackknife and other re-sampling based methods. The weak point of the method is that it does not possess the flexibility to cope with different types of estimators. For each type of estimator an empirical influence function has to be derived. This is not, however, a big issue in this case because in practical applications the set of estimators used for the survey data is fixed. Empirical influence functions derived for level, ratio and change estimators can be found in the literature e.g. by Canty and Davison (1999, p. 389).

The Jackknife linearization method for a stratified cluster design is now briefly described. For a more details about Jackknife linearization see e.g. Canty and Davison (1999, p. 382). Denote h as the index for strata, j as the index for the cluster inside stratum h, i as the index for the unit inside cluster j. The Jackknife linearization variance estimator is

$$\upsilon_L = \sum_h (1-f_h)\frac{1}{n_h(n_h-1)}\sum_{j=1}^{n_h} l_{hj}^2 ,$$

where $f_h = \frac{n_h}{N_h}$ is the sampling fraction of clusters in stratum h. n_h is the number of clusters sampled in stratum h. N_h is the number of clusters in the population of stratum h. l_{hj} is an empirical influence value for the cluster hj.

The key to the method is the calculation of l_{hj} for each type of estimator. In the case of level estimation (without using a GREG estimator) l_{hj} is computed as

$$l_{hj} = n_h y'_{hj} - \sum_i y'_{hj} ,$$

where $y'_{hj} = \sum_i \omega_{hji} y_{hji}$ and ω_{hji} is the design or initial weight (before GREG estimation) for the unit hji. The value of the study variable for the unit hji is y_{hji} and y'_{hj} is the weighted sum of y_{hji} over cluster hj.

If the GREG estimator is used for level, then l_{hj} is computed as

$$l_{hj} = n_h e'_{hj} - \sum_i e'_{hj} ,$$

where $e'_{hj} = \sum_i w_{hji} e_{hji}$, w_{hji} is a calibrated weight, $w_{hji} = \omega_{hji} g_{hji}$, g_{hji} is the calibration factor or g-weight for the unit hji, e_{hji} is a regression residual.

In the case of ratio estimation l_{hj} is computed as

$$l_{hj} = \frac{l^y_{hj} - \hat{R} l^z_{hj}}{\hat{Z}},$$

where l^y_{hj} is the empirical influence value for the Y variable and l^z_{hj} is the empirical influence value for the Z variable, $\hat{R}$ is the estimated value for the ratio, $\hat{R} = \frac{\hat{Y}}{\hat{Z}}$. $\hat{Z}$ is the estimated value for the Z level.

4 Conclusions

Prototype code has been written in SPSS to deliver SE estimates using the methodology described in this paper. The main benefits of the methodology applied are computing the

matrix M^{-1} only once during the weighting and the possibility to define several study variables as a matrix Y. The first results from testing the code show that it works faster than current procedures available for SE estimation giving the same SE estimates.

There are still several open issues to study in the near future such as checking for co-linearity between X variables and excluding linearly dependent variables from X, estimation of SE for the change estimator, estimation of SE in the case where bounding of *g*-weights is used, estimation of SE in the case of two-phase calibration, the possibility to define domains of interest as parameters of the procedure, which is required by users (currently a separate variable has to be created for each domain), and transferring the code from SPSS to SAS environments.

References

Canty A. J., Davison A. C. (1999) Resampling-based variance estimation for labour force surveys. *The Statistician*, No. **48**, 379-391.

Holmes D. J., Skinner C. J. (2000) Variance Estimation for Labour Force Survey Estimates of Level and Change. *Governmental Statistical Service Methodology Series*, No. **21**, 1-40.

Lundström S., Särndal C.-E. (2001) *Estimation in the presence of Nonresponse and Frame Imperfections*. Statistics Sweden, Örebro.

Lundström S., Särndal C.-E. (2005) *Estimation in Surveys with Nonresponse*. Wiley, Chichester.

Stukel D. M., Hidiroglou M. A., Särndal C.-E. (1996) Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization. *Survey Methodology*, Vol. **22**, No. 2, 117-125.

Yung W., Rao J. N. K. (1996) Jackknife Linearization Variance Estimators Under Stratified Multi – Stage Sampling. *Survey Methodology*, Vol. **22**, No. 1, 23-31.