Technical Note: the calculation of Sampling Variability for the Labour Force Survey (LFS)

Introduction

This note provides the formulas used for the calculation of standard errors (SEs) on the quarterly LFS and describes how the impact of the sample design and weighting are captured in these calculations.

Sample Design of the LFS

Addresses are sampled using systematic random sampling from the Postcode Address File (PAF) for Great Britain. The PAF is ordered by postcode sector. This means that addresses selected in close proximity on the PAF are also in close geographical proximity. This systematic random sampling from an ordered list is sometimes described as *implicit stratification*, because its impact on the sampling error is similar to stratified sampling.

To allocate addresses to interviews, Great Britain has been partitioned into a fixed number of interviewer areas (IAs) and within that into stints. There are 13 stints which represent each week of a quarter and also the interviewer's weekly workload. The 13 stinted areas within an IA have been randomly assigned. The addresses sampled from the PAF every quarter can then be matched up to the IA area and stints within that. The addresses are numbered within the stint in the original order they were sampled from the PAF.

This method of sampling and allocation means that addresses that are within close proximity on the PAF, and as a result geographically too, are accordingly assigned to the same IA and stint when interviewing takes place. It then follows that consecutive addresses in the same IA and stint will have greater similarities to each other than compared to two addresses sampled from completely different areas. We can capture this implicit stratification, by pairing consecutive addresses within the same stint and IA[1] and treating these pairs as the strata in standard error calculations.

The sampling for Northern Ireland is slightly different, as it is stratified on Local Authority (LA) and a random sample is drawn every week. However, for the purpose of SE calculations and the pairing of addresses, pairs are created within the weeks of a quarter. The actual LA stratification is ignored in the calculations, as we feel this would make only a small difference to the SEs.

However, when calculating the standard errors, the stratification at LA and ward level is ignored and instead addresses are paired within the weeks of a quarter. So two addresses in a pair will belong to the same week but may come from different LAs in Northern Ireland. We cannot observe LA and ward boundaries in our representation of the stratification in the standard error calculations as LA information within Northern Ireland is not made available for this purpose[2]. We expect that the resulting standard errors will be very slightly overstated for Northern Ireland estimates and that the impact on standard errors for the UK estimates will be negligible.

Weighting in the LFS

[1] If there is an odd number within an IA, the last address is added to the previous pair. However, addresses may be paired across the stints within the same IA.

[2] LA indicators have been provided to ONS but only with the specific purpose to weight the data to population totals.

This is a brief reminder of the weighting used on the main LFS. For more detail and background see the LFS User Guide[3].

The LFS weighting applies an iterative post-stratification procedure in the SIR raking algorithm using three sets of population controls:

i. Local Authority Districts (LAD) – 454 cells
ii. Age1 by sex (where Age1 is 0-15, single years between 16 and 24 and 25+) – 22 cells[4]
iii. Region by Age2 by sex (where Age2 is 5-year bands up to 79, and 80+) – 612 cells

This reason for this weighting is to reduce the bias (from nonresponse and undercoverage), with the added advantage of reducing the variance of survey estimates, particularly totals. However, in order to realise this advantage we must incorporate the post-stratification into the calculation of the variances.

The Jacknife Linearization Estimator

The Jacknife linearization estimator is used to calculate the variance of LFS estimates (ratios and totals). This was recommended by Holmes and Skinner (2000)[5], their conclusions were that the Jacknife linearization estimator was preferred to others (namely standard linearization and linearization (post-stratified)), as it allows for the full variation in the weights as well as being computationally undemanding.

It was computationally impossible to calculate post-stratified variances, in the chosen software STATA, using all three levels of stratification variables, so Holmes and Skinner looked for a close approximation. They found that a close approximation could be achieved using a single variable (Age2 by Sex by Region – 17×2×18 or 612 cells) in the calculations.

Notation

Individuals are labelled by a three-level index (hik), where

h (=1,…,H=110) denotes the sample stratum, here the address pair
i (=1,…,n_h) denotes the address within the stratum (or IA)
k indexes the individual within the address

So, y_{hik} is the estimate we are interested in for the kth individual within the ith address of hth stratum.

d_{hik} – the sample design weight, equal to the reciprocal of the probability of selection; this is the same for all individuals on the LFS, as there is an equal selection probability
w_{hik} – the grossing weight (produced after calibration to i, ii, iii above) in the SIR algorithm
M^c (=1,…,C) – these are known population counts for our post-stratification variable (Age2 by Sex by Region)
s^c – the sample of individuals within the cth cell.

Formulas for Totals

We can estimate the following cell-specific quantities using the following estimators,

[3] Labour Force Survey User Guide. Volume 1: Background and Methodology. Section 10. http://www.statistics.gov.uk/statbase/Product.asp?vlnk=1537
[4] These are applied separately for Great Britain and Northern Ireland.
[5] Holmes, D.J and Skinner, C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change. GSS Methodology Series No. 21. http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9220

$\hat{M}^c = \sum_{(hik)\in s^c} d_{hik}$ the design-weighted estimated population count in the cth cell

$\hat{Y}^c = \sum_{(hik)\in s^c} d_{hik} \times y_{hik}$ the design-weighted estimated total of our variable of interest in the cth cell

So then our estimator for the total, Y, is,

$$\hat{Y} = \sum_{(hik)} w_{hik} \times y_{hik} \qquad (1)$$

Now we want to calculate the variance for this estimated total.

$$Var(\hat{Y}_{ps}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \overline{z_h}) \qquad (2)$$

where $$z_{hi} = \sum_{\text{c}} \sum_{\text{k}\in\text{s}^\text{c}} \text{w}_{\text{hik}} \times e_{hik}^{c} \qquad (3)$$

and $$e_{hik}^{c} = y_{hik} - \frac{\hat{Y}^c}{\hat{M}^c} = y_{hik} - \frac{\sum_{(hik)\in s^c} d_{hik} \times y_{hik}}{\sum_{(hik)\in s^c} d_{hik}} \qquad (4)$$

Here we are calculating the variance in each cell and then summing over all the cells. The calculation of the residuals in (4), which are the difference between each individual's reported estimate and their post-stratum average, is carried out explicitly in STATA. The calculation of (2) uses the SVYTOTAL command on the post-stratified residuals. Note that the calculation of the overall variance requires both the calibration and design weights.

Formulas for Rates and Proportions

Ratios can be represented as a ratio of two totals. Means and proportions are special types of ratio where, the denominator (for means) and both the numerator and denominator (for proportions) are estimated counts.

$\hat{R}_{ps} = \frac{\hat{Y}_{ps}}{\hat{V}_{ps}}$ we estimate the ratio using the post-stratified totals for Y and V

We calculate the variance for this ratio using the estimator in (2) and redefine z_{hi}.

$$Var(\hat{Y}_{ps}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \overline{z_h}) \qquad (2)$$

where $$z_{hi} = \sum_{\text{c}} \sum_{k \in s^c} w_{hik} \times r_{hik}^c \qquad (5)$$

and $$r_{hik}^c = \frac{e_{yhik}^c - \hat{R} e_{vhik}^c}{\hat{V}_{ps}} \qquad (6)$$

$$e_{yhik}^c = y_{hik} - \frac{\hat{Y}^c}{\hat{M}^c} = y_{hik} \text{ - } \frac{\sum_{(hik) \in s^c} d_{hik} \times y_{hik}}{\sum_{(hik) \in s^c} d_{hik}} \qquad (7)$$

$$e_{vhik}^c = v_{hik} - \frac{\hat{V}^c}{\hat{M}^c} = v_{hik} \text{ - } \frac{\sum_{(hik) \in s^c} d_{hik} \times v_{hik}}{\sum_{(hik) \in s^c} d_{hik}} \qquad (8)$$

As before the explicit calculation of residuals for the estimated post-stratified totals Y and V are calculated in STATA and then fed into the SVYTOTAL command to give the variance estimate.

Sampling Variability for LFS Estimates of the Redundancy Rate

The Theory (from Dave Holmes, Southampton University)

The redundancy rate is defined as the ratio of the total number of redundant adults in the current quarter to the total number of employees in the previous quarter, $R = Y_2 / V_1$, and this is estimated by

$$\hat{R} \;=\; \hat{Y}_2 / \hat{V}_1 \quad , \qquad (1)$$

where $\hat{Y}_2$ and $\hat{V}_1$ are defined as

$$\hat{Y}_2 \;=\; \sum_{(hik) \in s_2} w_{2hik} y_{2hik} \qquad \hat{V}_1 \;=\; \sum_{(hik) \in s_1} w_{1hik} v_{1hik} \quad ,$$

(2)

where y_{2hik} is the value of the numerator variable for (*hik*)*s2* in the current quarter, and v_{1hik} is the value of the denominator variable for (*hik*)*s1* in the previous quarter.

The approximate variance of the estimated ratio in (1) is given by

$$\operatorname{var}(\hat{R}) \;=\; \frac{1}{V_1^2} \operatorname{var}(\hat{Y}_2 - R\hat{V}_1) \quad ,$$

$$= \frac{1}{V_1^2} \left\{ \operatorname{var}(\hat{Y}_2) + R^2 \operatorname{var}(\hat{V}_1) - 2R \operatorname{cov}(\hat{Y}_2, \hat{V}_1) \right\} \quad ,$$

$$= \frac{1}{V_1^2}\left\{ \operatorname{var}(\hat{Y}_2) + R^2 \operatorname{var}(\hat{V}_1) - 2R \operatorname{cov}(\sum_h \sum_{i \in s_h^*} z_{2hi}, \sum_h \sum_{i \in s_h^*} z_{1hi}) \right\},$$

(3)

where $z_{2hi} = \sum_k w_{2hik} y_{2hik}$, $z_{1hi} = \sum_k w_{1hik} v_{1hik}$, and s_h^* is the subsample of delivery points in stratum *h* for which data are available in both quarters.

The variance in (3) can be estimated using the same ideas as in Section 8 of Holmes & Skinner (2000). *R* and *V*1 are estimated using (1) and (2). The two variances (inside the bracket) come simply from the variance estimates of level at each quarter. For example, $\operatorname{var}(\hat{V}_1)$ is estimated using equation (2) of our original report with *zhi* defined as

$$z_{hi} = \sum_k w_{1hik} e_{hik},$$

(4)

where $e_{hik} = v_{1hik} - \mathbf{x}_{1hik}^T \hat{\mathbf{B}}_{1v}$ are the estimated residuals.

The covariance term inside the bracket in (3) can be expressed as equation (28) of Holmes & Skinner (2000) and estimated in the same way.

Standard Errors of change estimates

The formula for the standard error of change between quarter *t* and quarter *k* is given in section 3.3.3 of Butcher & Elliot (1987):

$$\operatorname{var}(\hat{R}_t - \hat{R}_k) = \operatorname{var}(\hat{R}_t) + \operatorname{var}(\hat{R}_k) - 2\operatorname{cov}(\hat{R}_t, \hat{R}_k),$$

(5)

where

$\hat{R}_t = \frac{\hat{Y}_t}{\hat{V}_{t\text{-}1}}$ and $\hat{R}_k = \frac{\hat{Y}_k}{\hat{V}_{k\text{-}1}}$

The covariance term expands to:

$$\operatorname{cov}(\hat{R}_t, \hat{R}_k) \square (\hat{V}_{t\text{-}1}\hat{V}_{k\text{-}1})^{-1}\left[\operatorname{cov}(\hat{Y}_t, \hat{Y}_k) + \hat{R}_t\hat{R}_k \operatorname{cov}(\hat{V}_{t\text{-}1}, \hat{V}_{k\text{-}1}) - \hat{R}_t \operatorname{cov}(\hat{V}_{t\text{-}1}, \hat{Y}_k) - \hat{R}_k \operatorname{cov}(\hat{Y}_t, \hat{V}_{k\text{-}1})\right],$$

(6)

Each of the covariance terms in equation (6) are computed using the linearised variable e_{hik} and equation 28 in Holmes and Skinner. Note that for estimates of annual change the fourth covariance term in (6) is zero as there is no overlap. Also evidence has shown that the second and third covariance terms are small and add/subtract little to the final estimates of standard error since the redundancy rates $\hat{R}_t$ and $\hat{R}_k$ are small. They have been included in my estimates but could be dropped in future if many rates were to be produced in order to save computing time. The standard error computed without the second third and fourth term tends to be conservative.

References

Holmes, D. & Skinner, C (2000) *Variance Estimation for Labour Force Survey Estimates of Level and Change*. GSS Methodology Series no. 21
Butcher, B & Elliot D (1987) *A Sampling Errors Manual*. OPCS