Combining Unemployment Benefits Data and LFS Data to Estimate ILO Unemployment for Small Areas: an Application of a Modified Fay-Herriot Method

R. Ambler (UK)
E-mail: Rebecca.Ambler@ons.gov.uk
D. Caplan (UK)
E-mail: David.Caplan@ons.gov.uk
R. Chambers (UK)
E-mail: rc6@soton.ac.uk
M. Kovacevic (Canada)
E-mail: milorad.kovacevic@statcan.ca
S. Wang (USA)
E-mail: sjwang@stat.tamu.edu

1. Background

Users of official statistics are often interested in information about local areas for economic planning, resource allocation and policy making. However, the traditional survey sources of data are not well equipped to meet this need. Within the UK, there has recently been a growth in demand for small area statistics, particularly in the context of the measurement of social exclusion and participation in the labour market. The Labour Force Survey (LFS) is the key source of national information on the labour market, but direct LFS estimates are of limited use for local data. This paper reports on the use of statistical modelling techniques to enhance the quality of small area statistics derived from the LFS.

The LFS is a continuous, large-scale survey, with a sample of around 60,000 households in each three-month period. These include around 150,000 people, of whom over 110,000 are aged 16 or over, in each three-month period. This is used to measure unemployment using the International Labour Organisation (ILO) definition. Throughout this paper, when we refer to unemployed we mean unemployed under this definition. The LFS sample selection procedure can be considered as equivalent to simple random sampling and is primarily designed to produce national estimates. In addition, the ONS produces an annual Local Area Database that contains estimates at the Unitary Authority/Local Authority (UA/LAD) level. Small sample sizes for many of these areas means that estimates can only be published for around 100 of the 407 UA/LADs in Great Britain. The ONS recognises that there is a need for reliable information for UA/LADs and has therefore undertaken a research project designed to improve the accuracy and widen the availability of statistics of unemployment for UA/LADs. This work has a close relationship with more general work in the field of small area estimation being carried out by the ONS (see Heady et al, 2000).

Small area estimation is the title given to a range of statistical techniques used to produce estimates for small areas when the "standard" survey estimates for these areas are unreliable or cannot be calculated. For a survey of common approaches see Ghosh and Rao (1994). The techniques involve the use of models to "borrow strength" over space, over time or from correlation with auxiliary information. The main source of auxiliary information available for this work is the claimant count. This is the administrative count of the number of people claiming unemployment related benefits. Because it is derived from an administrative system, the data are available without sampling error and can reliably be broken down, for example, into different age and sex categories as well as for any geographical unit down to very low level. There is a strong relationship between claimant count and ILO unemployment, although this relationship varies over time, between different areas and between men and women. The changes in the relationship over time are due to both changing administrative rules and factors changing through the economic cycle.

In association with the University of Southampton, the ONS has looked at three approaches to estimating unemployment at UA/LAD level by combining LFS and claimant count data; the

SPREE method (Purcell and Kish, 1980), a modified Fay-Herriot (1979) approach based on a logistic model for proportion of people unemployed in an age-sex class and a multilevel modelling approach (Goldstein, 1995), also based on a logistic model for age-sex counts derived from LFS data. This has allowed comparison of the results from each and consequent insights into their strengths and weaknesses. Since the two approaches based on logistic modelling performed very similarly, and were substantially better than the SPREE approach, we focus on the modified Fay-Herriot approach below.

2. Methodology

In what follows, i,j index age-sex class and g,h index area (i.e. UA/LAD). The sample data then consist of LFS estimates plus associated claimant counts for "cells" defined by age-sex classes within each area. Put N_{ig} equal to the total number of people in cell (i,g), with U_{ig} equal to the total number of unemployed in the same cell. Then $z_{ig} = U_{ig}/N_{ig}$ is the proportion of unemployed in the cell. In general, the distribution of the z_{ig} will be determined by the characteristics of area g (including its claimant count distribution). Put $E(z_{ig}) =$ $_{ig}$ and $var(z_{ig}) =$ $_{ig}(1$ $_{ig})/N_{ig}$. A linear logistic model is used to specify how the characteristics of area g influence the value $_{ig}$. That is, $\text{logit}(\pi_{ig}) = \mathbf{x}'_{ig}\beta$ where $\mathbf{x}_{ig}$ is a vector of known attributes for age-sex class i in area g. Let N_{ig} denote the LFS estimate of the number of people in cell (i,g), with U_{ig} denoting the LFS estimate of number of unemployed in cell (i,g). Then $z_{ig} = U_{ig}/N_{ig}$ is the LFS estimate of z_{ig}. We further assume that the age-sex classes are "finely enough" defined so that there is very little variation in sample weights for selected individuals within a class. Consequently, the LFS estimate for the proportion of unemployed individuals in cell (i,g) can be approximated by the sample proportion of unemployed in this cell. Since the LFS sample is essentially a simple random sample, we have

(1a) $E(z_{ig}|z_{ig}) = z_{ig}$

(1b) $var(z_{ig}|z_{ig}) = [(N_{ig} \quad n_{ig})/(N_{ig}-1)][z_{ig}(1 \quad z_{ig})/n_{ig}] = z_{ig}(1 \quad z_{ig})/n^{*}_{ig}$

where n_{ig} is the LFS sample size in cell (i,g) and $n^{*}_{ig} = n_{ig}(N_{ig}-1)/(N_{ig}-n_{ig})$. Provided z_{ig} and $\mathbf{x}_{ig}$ are independent given z_{ig}, it follows

(2a) $E(z_{ig}|\mathbf{x}_{ig}) = E[E(z_{ig}|z_{ig},\mathbf{x}_{ig})|\mathbf{x}_{ig}] = E(z_{ig}|\mathbf{x}_{ig}) =$ $_{ig}$

(2b) $var(z_{ig}|\mathbf{x}_{ig}) = E[z_{ig}(1 \quad z_{ig})/n^{*}_{ig}|\mathbf{x}_{ig}] + var(z_{ig}|\mathbf{x}_{ig}) =$ $_{ig}(1$ $_{ig})/n^{**}_{ig}$

where $n^{**}_{ig} = n^{*}_{ig}\left[1+(n^{*}_{ig}-1)/N_{ig}\right]^{-1}$. Typically n_{ig} is small relative to N_{ig}, so (2) can be combined with the logistic specification for $_{ig}$ to define an "approximate" binomial logistic model for the estimated sample proportions z_{ig}. This model can be fitted to the sample data via standard logistic regression software, using as inputs the "effective sample size" $n_{ig} = rnd(n^{**}_{ig})$ and "effective sample unemployed" counts $m_{ig} = rnd(n_{ig} \times z_{ig})$. Here rnd(.) is the "round to nearest integer" function. This leads to an estimate β of and an associated estimate $V(\beta)$ of $var(\beta)$. A naive estimator of z_{ig} is then $\pi_{ig} =$ antilogit $(\mathbf{x}'_{ig}\beta)$. However, this is not unbiased , even though β may be asymptotically unbiased. A first order bias corrected version is

(3) $$\pi_{ig} = \pi_{ig}\left(1 - \frac{1}{2}(1-\pi_{ig})(1-2\pi_{ig})\left[\mathbf{x}'_{ig}V(\beta)\mathbf{x}_{ig}\right]\right).$$

The final estimator of the total number of unemployed in area g is then

(4) $$\theta_g = \sum_{i \in g} \alpha_{ig} N_{ig} \pi_{ig}$$

where $i \in g$ denotes those age-sex classes "represented" in area g, and where the calibration coefficients $\{$ $_{ig}\}$ are calculated so that the estimators defined by (4) sum up to "standard" LFS

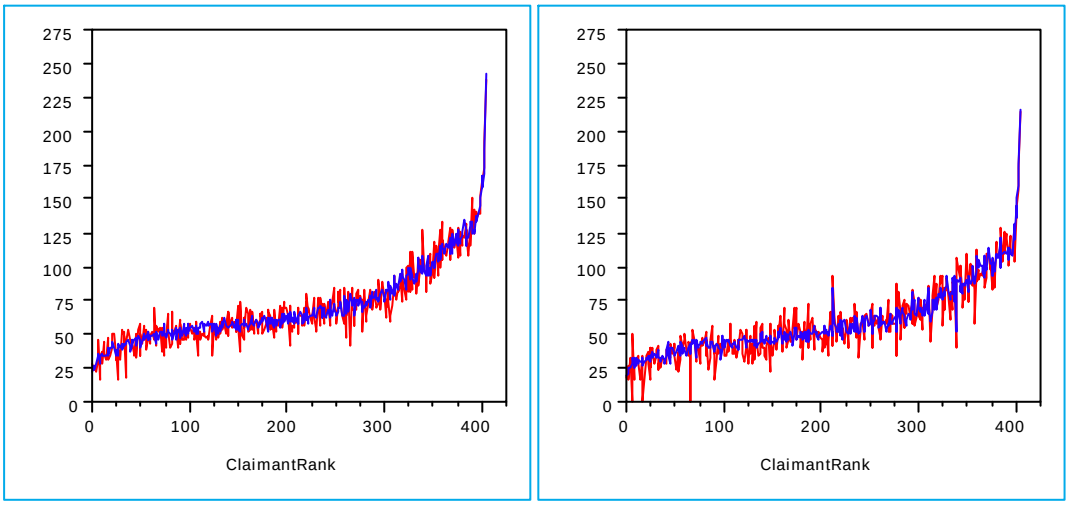
estimates at pre-specified levels of aggregation. This is achieved by iterative scaling. Note that the estimated covariance between the model-based estimators for areas g and h is then

$$(5) \qquad c(\theta_g, \theta_h) = \sum_{i \in g} \sum_{j \in h} \alpha_{ig} N_{ig} \pi_{ig} (1 - \pi_{ig}) \left[\mathbf{x}'_{ig} V(\beta) \mathbf{x}_{jh} \right] \pi_{jh} (1 - \pi_{jh}) N_{jh} \alpha_{jh} .$$

3. Application to Labour Force Survey data

The above procedure was used to produce estimates of unemployment for UA/LADs in Great Britain for the four years 1995-96 to 1998-99. These were based on annual LFS data. The key auxiliary variable was claimant count averaged over the same twelve-month period (March to February) as the LFS data. These counts entered the model as logits of the corresponding population proportions. In addition, the model included indicators for age- sex (2 sex classes by 3 age classes = 6 age-sex classes), geographic region (12 classes) and socio-economic cluster (7 classes). The last classification was based on a grouping of UA/LADs, developed by the ONS, which assigns each authority to one of seven clusters depending on its socio-economic attributes. Local authorities that have similar socio-economic characteristics are in the same cluster. Figure 1 below shows the relationship between the (more variable) "standard" UA/LAD unemployment estimates derived from the LFS and the (less variable) estimates produced by the above methodology for 1995-96 and 1997-98. In both cases the estimates are shown on a square root scale and are plotted in order of increasing value of claimant count in each year. The lower variability of the model-based estimates is clear. These gains are reflected in the much smaller estimated standard errors associated with the model-based estimates. In particular, the average (over the UA/LADs) of the ratio of the estimated standard error of the LFS estimate of unemployment to the estimated standard error of the model-based estimate, based on (5), ranged from 5.5049 in 1995-96 to 5.3851 in 1998-99.

Figure 1 LFS "direct" estimates of unemployment for 1995-96 (left) and 1998-99 (right) superimposed on corresponding model-based estimates.



4. Further work

One problem with the "synthetic" type of model-based estimator used above is that it does not allow for variation between the small areas not explained by variation in the covariates (including claimant count). Typically this results in estimated standard errors that are too small. This seems to be the case with the LFS example above where the average five-fold decrease in standard errors associated with the model-based approach is a little too good to be true. This problem can be tackled by extending the model to include a random effect for each small area, and is the basis of the third (multilevel modelling) approach to this problem that has been investigated by the ONS. Essentially, it consists of replacing the simple logistic model for the $_{ig}$ used so far by a model of the form $\text{logit}(\pi_{ig}) = \mathbf{x}'_{ig}\beta + u_g$, where the area specific values $\{u_g\}$ are assumed to be independent realisations of a random variable with zero mean and variance σ_u^2.

There are a variety of ways this more complicated model can be fitted to the LFS data and estimates of the "random effects" u_g and their variance σ_u^2 obtained. Work so far has focussed on using the estimates produced by the multilevel modelling package MLWin. This has shown that the UA/LAD estimates (4) based on "EBLUP-type" component estimates $\pi_{ig} = \text{antilogit}\ (\mathbf{x}'_{ig}\beta + u_g)$ tend to be closer to the LFS estimates in UA/LADs with larger samples and closer to the "fixed effect" estimates in UA/LADs with smaller samples. A practical problem with this estimator is estimation of its mean squared error. A compromise being considered by ONS at present is to use the synthetic estimates defined by (4), but to replace the variance estimator defined by (5) by the "larger" random effects version

$$(6) \qquad v(\theta_g) = \sum_{i \in g} \sum_{h \in g} \alpha_{ig} N_{ig} \pi_{ig} (1 - \pi_{ig}) \left[\sigma_u^2 + \mathbf{x}'_{ig} v(\beta) \mathbf{x}_{hg}\right] \pi_{hg} (1 - \pi_{hg}) N_{hg} \alpha_{hg}$$

where σ_u^2 is the estimated "between area" variance based on fitting the random effects model.

5. Conclusions

This work has shown that it is possible to improve the direct estimates of unemployment from the LFS by introducing additional information. The approach considered in this paper generates estimates that make considerable improvements in accuracy. However, a number of issues remain - particularly the decision on whether to include a random area effect in the model, and the estimation of the mean squared error of this "EBLUP-type" estimator. Another issue is that of estimating rates rather than population proportions. Calculating rates using model-based estimates of unemployment divided by direct LFS estimates of economic activity will not necessarily achieve the best possible estimates of rates. An alternative is to estimate unemployment, employment and inactivity simultaneously. Covariates for inactivity exist in the form of benefits data on pensions, sickness and disability. In principle, the above approach could therefore be extended to modelling inactivity, with the proportion of employed people then obtained as a residual. These approaches are presently being investigated.

References

Heady, P., Clarke, P., Brown, G., D'Amore, A. & Mitchell, B. (2000). Small area estimates derived from surveys: ONS central research and development programme. *Statistics In Transition* **4**, 635-48

Ghosh, M. & Rao, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science* **9**, 55-93.

Purcell, N. I. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review* **48**, 3-18.

Fay, R.E. & Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269-77.

Goldstein, H. (1995). *Multilevel statistical models*, 2nd edition. London: Arnold.

Resume

Cet article décrit la recherche qui a été menée a l'ONS (institut statistique britannique) afin de produire des estimateurs de qualité du chômage (en nombre et en pourcentage) à un niveau regional et local. Ces estimateurs tirent parti de la corrélation élevée qui existe entre le nombre de chômeurs enregistrés et les estimations de chômage au sens du BIT, provenant de l'enquête Emploi. Les résultats de l'analyse effectuée avec les données de l'enquête Emploi, sur une période de 4 ans (de 1995-96 a 1998-99), indiquent que la méthodologie proposée fonctionne bien.