Calibration on partly known counts in frequency tables with application to real data. November 2016

Michael Sverchkov and Richard Tiller,

Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC 20212

Abstract

Deville and Särndall (1992, Section 4) considered calibration on the known counts (cell counts or marginal counts) of a frequency table in any number of dimensions (generalized raking procedure). In this paper, we show that a similar procedure can be applied to the case of partly known overlapping counts. As an example we consider calibration of area-month-year unemployment estimates to month-year totals from a time series model of State estimates from the Current Population Survey and area-year totals from the American Community Survey.

Key words: calibration, generalized raking, Current Population Survey, American Community Survey

The opinions expressed in this paper are those of the authors and do not necessarily represent the policies of the Bureau of Labor Statistics

Introduction

Estimates of unemployment in states and local areas in the U.S. are produced as part of the Local Area Unemployment Statistics (LAUS) program of the Bureau of labor Statistics (BLS).

Estimates for the 50 states and the District of Columbia are made from time series models fitted to the direct survey estimates obtained from the Current Population Survey (CPS), so as to reduce variability due to small samples. The model estimates are benchmarked monthly to the highly reliable national CPS estimates (Pfeffermann and Tiller 2006).

With the exception of 5 large metropolitan areas, sub-state estimates are produced by a "handbook method" (HB) (Bureau of Labor Statistics Handbook of Methods, Chapter 4. Measurement of Unemployment in States and Local Areas, *www.bls.gov/opub/hom/)*. The latter is a non-model based method that uses local unemployment insurance claims and payroll employment data along with synthetic estimates of various categories of employment and unemployment not covered by the local data. To ensure consistency with the state-wide model based estimates, the sub-state estimates are ratio benchmarked each month to the model based state-wide totals. This monthly benchmarking is effective in reducing an overall bias in the HB method but it does not use local information.

Direct survey data on total employment and unemployment estimates has become available from the American Community Survey (ACS) on an annual basis for areas with at least a 65,000 population and 5-year estimates for areas of smaller sizes.

While the employment and unemployment definitions in the ACS are similar to the labor force concepts used by the CPS, there are systematic differences in the level estimates from the two surveys at high levels of aggregation, which cannot be explained by sampling error.

The geographic distributions of the ACS labor force estimates, however, appear to be very close to the CPS distributions. To further reduce bias in the HB estimates, this paper proposes adding a second layer of benchmarking constraints which requires the HB estimates to satisfy the annual or multi-year ACS distribution of employment and unemployment across areas within the State, in addition to satisfying the monthly constraints defined by the State model totals.

Calibration on partly known counts in frequency tables.

Let $n_{rc},\ r=1,...,R,$ c=1,...,C be the true (unknown) counts in row r and column c and d_{rc} be the corresponding HB estimates.

Suppose first that all true row totals, $T_1,...,T_R$, and column totals, $T_{R+1},...,T_{R+C}$, are known. Let $\mathbf{T}=(T_1,...,T_{R+C})$. For each (r,c) define the vector $\mathbf{x}_{rc}=(x_{rc1},...,x_{rcR+C})$ where $x_{rci}=1$ if $i=r$ or $i=R+c$ and 0 otherwise, so $\mathbf{x}_{rc}=(0,...,0,1,0,...,0,0,0,...,0,1,0,...,0)$ ($r-1$, $R-r$, $c-1$, $C-c$). Then $\sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}\mathbf{x}_{rc}=\mathbf{T}$.

For example, if the State consists of 15 areas and the data are available for years 2008 – 2013, the State totals are known for each month × year and the area totals are known for each year, then for a unit in area 1, year 2008 (area × year corresponds to column), month 12, year 2008 (month × year corresponds to row):

$$\mathbf{x}_{12,1}=(0,0,...,1,...\ 0,0,...\ 0,\ \ 1\ ,\ 0\ ,\ \ 0\ ,\ \ 0\ ,\ 0\ ,\ \ 0)$$
$$\underbrace{1,2,...12}_{2008},...\underbrace{1,2,...12}_{2013},\underbrace{2008,2009,...2013}_{\text{Area 1}},...\underbrace{2008,2009,...2013}_{\text{Area 15}}$$

Then the estimates $d_{rc},\ r=1,...,R,$ c=1,...,C can be corrected by the known totals similar to Deville and Särndall (1992): find a set of "weights" (new counts), $w_{rc},\ r=1,...,R,$ c=1,...,C , that are as close as possible to $d_{rc},\ r=1,...,R,$ c=1,...,C and satisfy the constraints $\sum_{r=1}^{R}\sum_{c=1}^{C} w_{rc}\mathbf{x}_{rc}=\mathbf{T}$. "As close as possible" is usually defined by some distance function, $D(\mathbf{d},\mathbf{w})$, between $d_{rc},\ r=1,...,R,$ c=1,...,C and $w_{rc},\ r=1,...,R,$ c=1,...,C, for example, if $D(\mathbf{d},\mathbf{w})=\sum_{r=1}^{R}\sum_{c=1}^{C}(w_{rc}-d_{rc})^2/d_{rc}$ then minimization of $D(\mathbf{d},\mathbf{w})$ is explicit, (simple matrix algebra, see Deville and Särndall 1992, Section 1).

Now, consider our case where all true row totals (CPS estimates of year-monthly totals), $T_1,...,T_R$ are known, column totals (ACS estimates of area-year totals), $T_{R+1},...,T_{R+K}$,

$K < C-4$ are known, and five year column totals (ACS estimates of area-year totals), $T_{R+K+1}(5),...,T_{R+C-4}(5)$, are known, where $T_t(5) = T_t + T_{t+1} + T_{t+2} + T_{t+3} + T_{t+4}$.

Define $\mathbf{T}^* = (T_1,...,T_{R+K}, T_{R+K+1}(5),...,T_{R+C-4}(5))$, and $\mathbf{x}^*_{rc} = (x_{rc1},...,x_{rcR+K}, x_{rcR+K+1}(5),...,x_{rcR+C-4}(5))$, $x_{rci}(5) = x_{rci} + x_{rci+1} + x_{rci+2} + x_{rci+3} + x_{rci+4}$.

Then $\sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}\mathbf{x}^*_{rc} = \mathbf{T}^*$, and therefore again the estimates d_{rc}, $r = 1,...,R$, c=1,...,C can be corrected by the known totals: find a set of "weights" (new counts), w_{rc}, $r = 1,...,R$, c=1,...,C , that are as close as possible to d_{rc}, $r = 1,...,R$, $c = 1,...,C$ and satisfying the constraints $\sum_{r=1}^{R}\sum_{c=1}^{C} w_{rc}\mathbf{x}^*_{rc} = \mathbf{T}^*$. Here again, if $D(d,w) = \sum_{r=1}^{R}\sum_{c=1}^{C}(w_{rc} - d_{rc})^2 / d_{rc}$ then minimization of $D(d,w)$ is explicit (see Deville and Särndall 1992, Section 1).

We applied the suggested calibration procedure to the real data (explained in introduction) for a number of states and found that distribution across the counties can be essentially different for some counties after additional calibration on ACS estimates of the totals.

References.

Deville, J. C. and Särndal, C.-E. (1992), Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, pp. 376-382.

Pfeffermann, D., and Tiller, R. (2006) "State-space modelling with correlated measurement errors with application to small area estimation under benchmark constraints". *Journal of the American Statistical Association*, **101**, (476), 1387-1397.