Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 53 (2009) 2378–2389
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

An efficient method of estimating the true value of a population characteristic from its discrepant estimates

P.A.V.B. Swamy[a,*], Jatinder S. Mehta[b], I-Lok Chang[c], T.S. Zimmerman[a]

[a] *Bureau of Labor Statistics, Postal Square Building, 2 Massachusetts Ave, N.E., Washington, DC 20212, USA*
[b] *Department of Mathematics, Temple University, Philadelphia, PA 19122, USA*
[c] *Department of Mathematics and Statistics, American University, Washington, DC 20016, USA*

Available online 3 December 2007

Abstract

A fruitful method of pooling data from disparate sources, such as a set of sample surveys, is developed. This method proceeds by finding the first two moments of two conditional distributions derived from a joint distribution of two sample estimators of employment for each of several geographical areas. The nature of the two estimators is such that one of them can yield a better estimate of national employment than the other. The regression of the former estimator on the latter estimator with stochastic intercept and slope is used to generate an improved estimator that is equal to bias- and error-corrected estimator for each area with probability 1. This analysis is extended to cases where more than two estimates of employment are available for each area.

1. Introduction

When a domain-specific sample of large enough size is not available, a reliable estimate of a population characteristic of unknown value for the domain can only be found by pooling data from different sources. The present paper describes a novel method of solving this small area estimation problem by taking into account both sampling and non-sampling errors. The specific problem we solve in this paper is the problem of estimating the true value of employment from its estimates made from different data sets for each of several small geographical areas in the United States (U.S.). The main sources of these data sets are the Current Population Survey (CPS), American Community Survey (ACS), Current Employment Statistics (CES) survey, and Quarterly Census of Employment and Wages (QCEW) program.

Mis-specifications in models can result in mis-estimated coefficients and error variances and covariances (see, e.g., Goldberger (1961)). Using two or more estimators of employment for each of several small areas, the method developed in this paper avoids serious specification errors.

The inapplicability of the generalized least squares (GLS) method to pool a forecast of employment from an autoregressive integrated moving average (ARIMA) model with survey estimates of employment is shown in

* Corresponding author. Tel.: +1 202 691 7424; fax: +1 202 691 5999.
E-mail addresses: paravastu_s@bls.gov (P.A.V.B. Swamy), mehta1007@comcast.net (J.S. Mehta), ilchang@american.edu (I-L. Chang), zimmerman_tamara@bls.gov (T.S. Zimmerman).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.11.012

Section 2. The use of its sample estimate in place of an unknown optimal value of the weight used in a weighted average of two survey estimators may not result in a composite estimator that is better than either survey estimator in terms of mean square error, as shown in Rao (2003, pp. 57–59). This is the second result stated in Section 2, which also contains the following five results: (i) it finds the first two moments of two conditional distributions derived from a joint distribution of the CPS and ACS estimators of employment for several areas; (ii) it shows the advantages of using the direct regression of the CPS estimator on the ACS estimator with stochastic coefficients instead of the reverse regression; (iii) it uses the direct regression to derive an improved ACS estimator that is equal to bias- and error-corrected CPS estimator with probability (w.p.) 1; (iv) one other estimator of employment is derived for each area in Section 2. This estimator has the minimum average mean square error within the class of certain assumed non-linear functions of the ACS estimator. (v) At the end of Section 2, the direct regression of the CPS estimator on the ACS estimator is extended to include more than two estimators. Section 3 uses the method of Section 2 to jointly model monthly CPS, annual ACS, monthly CES, monthly QCEW, and other data for sub-state areas. An empirical example is given in Section 4. Section 5 concludes the paper.

2. A method of simultaneously improving one and correcting another of two estimates of an unknown true value and its extensions

Let Y (or Y^*) be the finite population value of total "place-of-residence" (or "place-of-work") employment for a geographical area in a period. Such a value cannot be known exactly from its sample estimates because these estimates are subject to some uncertainty. It cannot also be known exactly from a census (100% sample) of the total population because censuses are often subject to omissions, duplications, and reporting and recording errors. Failing a method of determining Y exactly, an alternative is to reconcile the available estimates of Y computed from different data sets. The present paper takes this alternative approach and tries to remove their biases and model errors from different estimates of employment for states and sub-state labor market areas (LMAs). LMAs are defined in terms of full counties in all states except New England, where Minor Civil Divisions (MCDs) are used to define LMAs. Let Y_{ivtm} denote the "true" value, Y, for LMA v of state i in month m of year t. Let the number of LMAs in state i be p_i and let the number of states in the nation be n. Then the averages and aggregates that are of our interest are: $Y_{ivt} = (1/12)\sum_{m=1}^{12} Y_{ivtm}$, $Y_{itm} = \sum_{v=1}^{p_i} Y_{ivtm}$, $Y_{it} = (1/12)\sum_{v=1}^{p_i}\sum_{m=1}^{12} Y_{ivtm}$, and $Y_t = \sum_{i=1}^{n} Y_{it}$. Estimation of Y_{it} and Y_t is considered in this section and that of Y_{itm}, Y_{ivt}, and Y_{ivtm} is considered in Section 3 below.

2.1. Three different estimators

Let $\hat{Y}_{it}^{\mathrm{ACS}}$ and $\hat{Y}_{it}^{\mathrm{CPS}}$ denote the ACS and CPS-composite estimators of "place-of-residence" employment, Y_{it}, respectively. We assume that these estimators are subject to both sampling and non-sampling errors and write $\hat{Y}_{it}^{\mathrm{CPS}} = Y_{it} + \varepsilon_{it}^{\mathrm{CPS}}$ with $\varepsilon_{it}^{\mathrm{CPS}} = e_{it}^{\mathrm{CPS}} + u_{it}^{\mathrm{CPS}}$ and $\hat{Y}_{it}^{\mathrm{ACS}} = Y_{it} + \varepsilon_{it}^{\mathrm{ACS}}$ with $\varepsilon_{it}^{\mathrm{ACS}} = e_{it}^{\mathrm{ACS}} + u_{it}^{\mathrm{ACS}}$, where e_{it}^{CPS} and e_{it}^{ACS} denote the sampling errors and u_{it}^{CPS} and u_{it}^{ACS} denote the non-sampling errors of $\hat{Y}_{it}^{\mathrm{CPS}}$ and $\hat{Y}_{it}^{\mathrm{ACS}}$, respectively. The principal aim of this paper is to develop a method for computing standard errors and confidence limits that remain valid when these non-sampling errors are present. To deal with cross-sectional variations in $\varepsilon_{it}^{\mathrm{CPS}}$ and $\varepsilon_{it}^{\mathrm{ACS}}$ first, we fix t and let i vary throughout this paper. Their temporal variations will be dealt with in future work.

Following Rao (2003, p. 76), we use E_m and V_m to denote the expectation and variance with respect to a probability distribution assigned to the Y_{it}'s that are assumed to be a random sample from a "superpopulation", respectively. These operators, E_m and V_m, should be distinguished from E_p and V_p that denote the expectation and variance with respect to the probability (or randomization) distribution induced by a sampling design conditional on Y_{it}, respectively (see Wolter (1985, p. 381)).

QCEW data on "place-of-work" non-farm employment, denoted by Y_{it}^*, may also be used to estimate Y_{it} that is not equal to Y_{it}^*. The problem with these data is that for the current month, say m, they are not available until six to nine months after m. Consequently, the shortest horizon of a forecast of Y_{ivtm}^* computed from the available QCEW data is 6 to 9 months. Let $\hat{Y}_{it}^{\mathrm{ARIMA}}$ denote a forecast of Y_{it}^* made from an ARIMA model fitted to QCEW annual time-series data, $\{y_{i,t-h}^*\}$, on $Y_{i,t-h}^*$ with $h > 0$ and let $\varepsilon_{it}^{\mathrm{ARIMA}^*} = \hat{Y}_{it}^{\mathrm{ARIMA}} - Y_{it} + (Y_{it} - Y_{it}^*)$, with $E_m(\varepsilon_{it}^{\mathrm{ARIMA}^*} | y_{i,t-h}^* \text{ with } h > 0) = 0$.